Localization stability of weakly supervised deep neural networks

Rumjana Romanova, Joaquin van Schoren, Veronika Cheplygina
Department of Biomedical Engineering, Eindhoven University of Technology

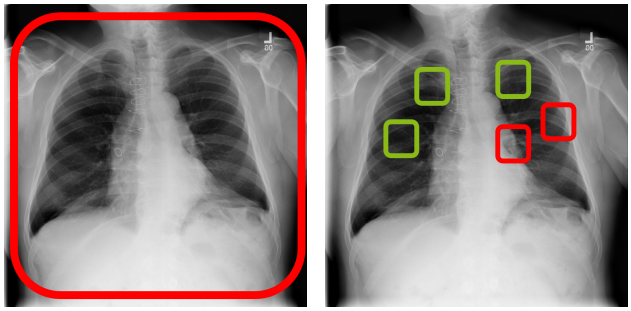
***Abstract*—It is of interest to the medical field to localize abnormalities in large volumes of medical images. Medical images usually comes with two types of annotations – annotations that indicate the presence of an abnormality, and annotations that indicate the exact location of abnormalities. Annotations with exact locations in medical imaging are scarcely available, limiting the use of established fully supervised algorithms for localizing abnormalities. Thus, medical imaging often relies on multiple instance learning (MIL). MIL algorithms rely only on annotations about the presence of an abnormality, which are more widely available, and can perform classification about the presence or location of an abnormality. This type of inference, however, may lead to unstable localization.**

This paper studies stability of localization of abnormalities in MIL algorithms in deep learning architectures. We propose two alternative measures of stability, based on the correlation between predictions; and based on localization agreement between two predictions. Using the proposed scores, we investigate the stability on two large public medical datasets with a common deep learning architectures for MIL. We explore the relationship between stability and performance across various MIL aggregation functions.

I. Introduction

Researchers in the medical field apply machine learning on medical images for computer-aided diagnosis and computer-aided detection [1]–[3]. Computer-aided detection specializes in localizing abnormalities in medical images. However, the detection of abnormalities is challenging due to the *limited availability and quality of data annotations*. Two types of annotations are differentiated. The first type annotates the absence or presence of an abnormality in the image as a whole, and the other type indicates the specific abnormality location within the image. An example is shown in Figure 1. Labels in medical imaging often only indicate the presence of an abnormality, as annotations about exact locations are costly, time-consuming and require the expertise of people in the domain. Hence, to overcome the absence of location annotations, classifiers in such settings are frequently trained in a weakly supervised manner [4], [5], such as multiple instance learning (MIL).

Originally introduced by Dietterich et al. [6], multiple instance learning is an extension of supervised learning. Similarly to fully supervised algorithms, multiple instance learning relies on labeled observations. However, in MIL, observations are defined as a set of bags, where every bag consists of multiple instances. A bag is often an image, and instances are segments of the associated image. Annotation indicating the presence of an abnormality in an entire image is referred to as bag-level annotation, and annotation specifying the exact location of an abnormality is called instance-level annotation. Figure 1a demonstrates a bag-level annotation, and Figure 1b shows instance-level annotation.



(a) (b)

Fig. 1: Types of annotation (a) Annotation indicating presence of an abnormality *(in red)* in the image as a whole (b)Annotation indicating exact location of healthy tissues *(in green)* or abnormal tissues *(in red)* on specific locations.

MIL classifiers are often trained only on available bag labels. However, we are ultimately interested in both, *bag and instance* labels, and MIL classifiers can be designed to predict instances or bags. This type of weak supervision facilitated the application and research of machine learning algorithms in fields where labeled data is rather limited. The limited labeled data in medical field lends itself well to the use of MIL methods. Furthermore, MIL algorithms perform well on various imaging modalities. Successful accounts include computed and optical tomography, radiography, ultrasound, and MRI scans [7], [8] on various anatomic structures such as the brain, eyes, breasts, lungs, abdomen, and cells [7], [9].

Instance classifiers in MIL predict on instance level, but are trained with bag labels. So earlier research examines the relation between bag and instance performance in MIL classifiers [10], [11]. Cheplygina et al. [12] focus on another aspect of the bag-instance relationship - particularly the *stability* of the predictions. Stability reflects the capability of algorithms to *consistently* detect abnormalities. Stability is measured by comparing the predictions of multiple models trained on subsets of the same dataset. In this way, stability can be seen as an unsupervised measure to check for instance performance, without having instance labels. The caveat is that stability does not directly infer good localization, but rather an important prerequisite of predicting instances accurately.

Examining notable MIL algorithms excluding deep learning architectures, Cheplygina et al. [12] show that there is a trade-off between instance stability and bag performance. Currently, deep learning architectures are employed more often to solve MIL problems [13]–[15]. We identify the need to expand the research about stability of deep learning architectures. In addition, the existing stability score is not be an optimal measurement of stability, especially for some classifiers, thus we also need to formulate alternative ways of measuring stability. This project aims to further investigate the following aspects of stability:

- How can we improve the measurement of stability?
- How do current deep learning models perform in terms of stability?
- What is the stability of models with respect to bag performance?
- How do assumptions of MIL algorithms influence the stability?

We provide alternative measures of the stability which can be used for reporting models' stability performance, in addition to the traditional evaluation metrics. Consequently, we examine the stability of a common deep learning architecture for MIL. Results from the experiments demonstrate that stability increases with the bag performance, although the stability is still somewhat low even for datasets with perfect evaluation performance. Finally, we find that stability of MIL algorithms can vary according to the aggregation used from instance to bag predictions.

This paper is organized as follows. Section II describes the previous work in MIL and stability, Section III propose alternatives for a stability score. Section IV introduces the experimental setup, and Section V presents the results.

II. Related Work

This section introduces fundamental aspects of the project, among which are the concepts of MIL, stability score, and common deep architectures for MIL.

A. Multiple Instance Learning

Multiple instance learning [6] is a variation of supervised learning. Each observation is seen as a bag B_i of N_i instances, such that:

$$B_i = \{x_{ij} | j = 1, ..., N_i\} \subset \mathbb{R}^d$$

and x_{ij} is an instance, a d-dimensional feature vector. Ultimately every instance has an associated label $y_{ij} \in \{0, 1\}$, but y_{ij} is *not known* during training. Instead, each bag is annotated with bag label $Y_i \in \{0, 1\}$.

MIL classifiers are trained on bag labels, but they can be designed to predict bag labels, or the instance labels. Depending on their predictions we differentiate between bag classifiers and instance classifiers [16]. A bag classifier g is described as:

$$\hat{Y}_i = g(x_{ij} | j = 1, ..., N_i)$$

while an instance classifier f is:

$$\hat{y}_{ij} = f(x_{ij})$$

Instance classifiers are attained by either assuming an explicit relation between bags and its instances, or aggregating their instance labels within the bag. In the first case, various relations can be assumed, but commonly a bag label is negative when all instances within are negative [17]:

$$\hat{Y}_i = \begin{Bmatrix} 0, & \text{iff } \forall j : y_{ij} = 0 \\ 1, & \text{otherwise} \end{Bmatrix}$$

However, another frequent approach is aggregating all instance labels to derive the bag label [17], such that:

$$\hat{Y}_i = \theta(\hat{y}_{ij} | j = 1, ..., N_i)$$

where θ is the aggregation function for all instances in a bag.

B. MIL Pooling operators

MIL pooling operators refer to the aggregation function which derives a bag label from all its instance predictions. Common choices for MIL operators are `Max`, `Mean` or `Log-Sum-Exp (LSE)`. The `Max` operator implies that the bag has the same label as its most discriminate instance:

$$\hat{Y}_i = \max_j(\hat{y}_{ij}) \tag{1}$$

`Mean` operator considers all the instances equally discriminate, and assigns a bag label which represents all of the encompassing instances.

$$\hat{Y}_i = \frac{1}{N_i} \sum_j^{N_i} (\hat{y}_{ij}) \tag{2}$$

`LSE` is an approximation of the `Max` function [18]:

$$\hat{Y}_i = \log \left[\frac{1}{N} \sum_j \exp(\hat{y}_{ij}) \right] \tag{3}$$

Furthermore, a more flexible version of `Log-Sum-Exp` with hyper-parameter r exists:

$$\hat{Y}_i = \frac{1}{r} \log \left[\frac{1}{N} \sum_j \exp(r\hat{y}_{ij}) \right] \tag{4}$$

The benefit of this function is that according to the value of r, the function can resemble the `Max` function (with large r), or estimate the `Mean` function (with a smaller value of r).
Another way to infer the bag label from the instance labels is the `Noisy-OR (NOR)` aggregation [19]:

$$\hat{Y}_i = 1 - \prod_j (1 - \hat{y}_{ij}) \tag{5}$$

In `NOR`, each instance is an independent binomial variable with a probability $\hat{y}_{ij}$ of being positive. The probability of a bag being negative is equal to the simultaneous probability that all instances are negative. On the other hand, the probability of a bag not being negative is equal to the probability of being positive [20].

C. Stability Score

Consider a bag B_i and two similarly trained instance classifiers, f' and f'', obtained by the same algorithm. We differentiate between two kinds of instance classifiers - yielding probabilities, or binary predictions for every instance.

Let W_i' and W_i'' be the set of instance predictions from f' and f'' on B_i.

$$W_i' = \{f'(x_{ij})\} \text{ and } W_i'' = \{f''(x_{ij})\}$$

Depending on the type of classifier, W_i' and W_i'' are sets of either probabilities, or binary instance predictions of the same bag. If f' and f'' predict instance probabilities, then each instance is ultimately expected to have the same, or closely correlated prediction value in W_i' and W_i''. In case of binary predictions, let $W_i^{'+} \subseteq W_i'$ and $W_i^{''+} \subseteq W_i''$ denote the positive predictions from W_i' and W_i'', respectively. $W_i^{'+}$ and $W_i^{''+}$ are expected to agree on the containing instances if the algorithm yields reliable results. In order to measure the degree of agreement between two predictions a *stability score* is introduced.

Considering that bags are images and instances in a bag are sub-regions of an image, the stability problem in this case is comparable to similarity problems in other fields. In object detection and segmentation, a common agreement metric [21]–[24] is the Jaccard index (also known as intersection over union (IoU)) [25]. The Jaccard index measures the area between predicted bounding boxes and ground truth segmentation:

$$Jaccard(P, Q) = \frac{|P \cap Q|}{|P \cup Q|} \tag{6}$$

The Jaccard index is the stability score originally defined in [12]. Their proposal of stability score measures the agreement of *positive* instances between two predictions on the same image, as

$$S_J(W_i', W_i'') = \frac{n_{11}}{n_{01} + n_{10} + n_{11}} \tag{7}$$

where

$$n_{01} = |\{x_{ij} | (\hat{y}_{ij}' = 0 \wedge \hat{y}_{ij}'' = 1)\}|$$

$$n_{10} = |\{x_{ij} | (\hat{y}_{ij}' = 1 \wedge \hat{y}_{ij}'' = 0)\}|$$

$$n_{11} = |\{x_{ij} | (\hat{y}_{ij}' = 1 \wedge \hat{y}_{ij}'' = 1)\}|$$

So, the score uses the binary labels of the predictions. Binary metrics may suit binary predicting instance classifiers, but measuring agreement from probabilities requires first a conversion to binary predictions. Such a conversion leads to information loss, but also makes the stability score dependent on the threshold for the conversion.

In the context of feature selection, Kuncheva investigates a stability score to measure the agreement between two sets [26]. Acknowledging the Jaccard index as a viable option, Kuncheva demonstrates another major drawback: the method assigns a high stability value even when stability is highly likely by chance. To counteract artificial inflation of the stability score, a metric is proposed where high stability values are attained only after surpassing the stability by chance. The proposed score has the general form:

$$S_C(P, Q) = \frac{\text{Observed}(k) - \mathbb{E}(k)}{\max(k) - \mathbb{E}(k)}, \tag{8}$$

where k is the amount of overlap between the two subsets, P and Q. The novelty of the metric resides in the correction term $\mathbb{E}(k)$. We observe a similar correction $\mathbb{E}(k)$ in several other measures, such as Adjusted Rand Index [27] and Cohen's kappa statistic [28]. Finally, the stability score is required to be **monotonically increasing**, **unsupervised**, have **limits**, and preferably correct for chance [12] [26].

D. Deep learning architectures for MIL

Architectures for MIL tasks encompass different types of layers. The most commonly employed are convolutional layers, sometimes followed by pooling layers [29]–[32]. In contrast, [33] employs only fully connected layers integrated with residual connections.

The existence of large organized databases such as ImageNet [34] enabled various deep and very deep neural net architectures, which capture detailed features and greatly boost performance. In medical images, transfer learning from different tasks and/or domain is also used to improve performance of the classification tasks [35]–[38]. Architectures with pre-trained neural networks are used in segmentation and localization tasks in weakly supervised settings [30], [31]. The influence of transfer learning is evident where successful localization of objects is achieved without bounding boxes during training time [31]. Hence, a representative architecture of the common MIL models has convolutional or fully connected layers, and possibly transfers knowledge from another domain or task.

E. Evaluation Metrics

Parallel to the stability, it is of utmost importance to establish the general performance of the model used in terms of classification and localization capabilities.

1) Localization performance: Localization performance demonstrates a model's skill on instance level. Instance level evaluation, however, can be achieved only in the presence of images with instance labels. These images are denoted as $x_i|bbox_i$. By comparing the provided bounding boxes with the instance predictions, the DICE coefficient [39] (Equation 9) and the accuracy from the Jaccard index (Equation 6) can be computed.

$$DICE(P, Q) = \frac{2|P \cap Q|}{|P| + |Q|} \tag{9}$$

Depending on its design, an instance classifier may predict region of abnormality with coordinates of the location. Then the predictions are binary and can be directly used to compute the DICE score (Equation 12) and the Jaccard index (Equation 13). On the other hand, an instance classifier can predict probability for each instance, which means that all instances first should get a binary representation. By setting a threshold

of the raw probabilities, instances are divided into 'positive' and 'negative' for the class-related problem:

$$\hat{y}_{ij} = \left\{ \begin{matrix} 1, \text{iff } f(x_{ij}) \geq T(binary) \\ 0, \text{otherwise} \end{matrix} \right\} \tag{10}$$

Furthermore, when instance predictions are probabilities, the anomalous regions can be discrete. Regardless, the positive predictions in a bag are treated as a single anomalous region. In addition, an instance has the 'positive' label as long as it is within the boundaries of the annotation:

$$y_{ij} = \left\{ \begin{matrix} 1, \text{iff } x_{ij} \in \text{BoundingBox}_i \\ 0, \text{otherwise} \end{matrix} \right\} \tag{11}$$

Figure 2a and 2b illustrate how probability predictions are converted to binary predictions on an exemplary bag. 2c and 2d depict the instance labeling from available segmentation, and 2e illustrates the computation of true positive (TP), false positive (FP) and false negative (FN) instances.

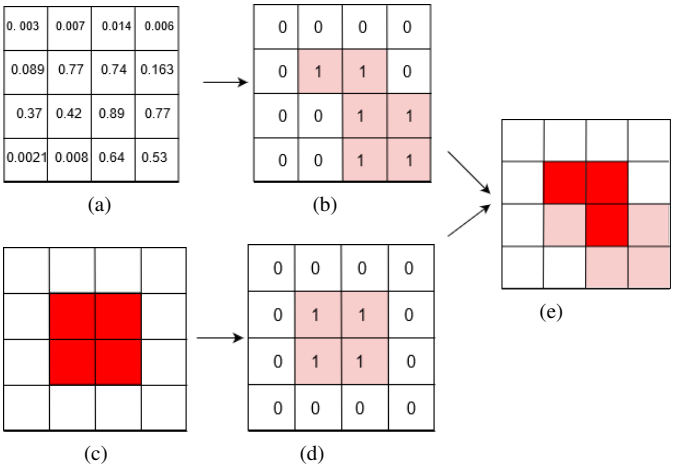


Fig. 2: Computation of TP and FP and FN from probability predictions (a) Instance predictions in a bag (b) Binary predictions after thresholding of 0.5 (c) Ground truth segmentation in red, negative labels in white. (d) Conversion from segmentation to binary labels (e) TP is the overlap of predictions and labels in bright red; instances in light red are FP or FN instances; white has both negative prediction and label

$$TP = \sum_j \mathbb{1}\left\{ y_{ij} = \hat{y}_{ij} = 1 \right\}$$

$$FP = \sum_j \mathbb{1}\left\{ y_{ij} = 0 \wedge \hat{y}_{ij} = 1 \right\}$$

$$FN = \sum_j \mathbb{1}\left\{ y_{ij} = 1 \wedge \hat{y}_{ij} = 0 \right\}$$

The DICE score is:

$$DICE(x_i|bbox_i) = \frac{2TP}{2TP + FP + FN} \tag{12}$$

The Jaccard index is:

$$Jacc(x_i|bbox_i) = \frac{TP}{TP + FP + FN} \tag{13}$$

$Jacc(x_i|bbox_i)$ is a continuous score per image with value between 0 and 1. Interpreting a prediction as accurate with respect to the label is not immediately apparent. Hence, a threshold `T(Jacc)` is used for computing the accuracy of the localization performance:

$$Accuracy(x_i|bbox_i) = \left\{ \begin{matrix} 1, \text{if } Jacc(x_i|bbox_i) \geq T(Jacc) \\ 0, \text{otherwise} \end{matrix} \right\} \tag{14}$$

2) Classification performance: Unlike localization, classification performance is measured for all bags due to their label availability. Finally, classification performance is represented with Area under the ROC curve (AUC) [40] between the bag labels and bag predictions AUC(Y, $\hat{Y}$). AUC value is computed using the false positive rate and the false negative rate, which makes it suitable to both balanced and imbalanced datasets.

III. Proposed Methods

Although a stability score is already proposed by Cheplygina et al. [12], the score may be weak. Agreement between sets of predictions may not be captured accurately for instance classifiers predicting probabilities. For instance classifiers with binary output, high stability may be measured as a result of random agreement between the sets. That is why we propose several alternative measures, which we analyze and choose for further experiments. Finally, the following sections examine stability scores for binary and probability predictions, respectively.

A. Stability scores for binary predictions

We propose several alternative scores for assessing the stability for binary predictions. An overview of the proposed methods, with their weaknesses and strengths is presented in Table Ia. The notation below follows the constituted notation,

$$n_{00} = |\{x_{ij}|W'_{ij} = 0 \wedge W''_{ij} = 0\}|$$

and

$$N = n_{00} + n_{01} + n_{10} + n_{11}$$

1) Positive Overlap (PO): The overlap coefficient, also known as the Szymkiewicz–Simpson coefficient [41], is suggested as a measurement sensitive to the size of the sets compared. Contrary to the Jaccard index, which stays invariant despite of the set sizes, the overlap coefficient is influenced by the set sizes. The set sizes are the same when comparing predictions on one and the same bag. However, considering only the positive predicted instances of the two classifiers, the assumption of equivalence of the two set sizes no longer holds. Then, the two sets have varying sizes and the overlap coefficient may be a better fit. The stability score (S_{PO}) proposed is:

$$S_{PO}(W'_i, W''_i) = \frac{n_{11}}{\min(n_{10}, n_{01}) + n_{11}} \tag{15}$$

Limits: $S_{PO} \in [0,1]$. $S_{PO} = 0$ is attained when $n_{11} = 0$, and $S_{PO} = 1$ when $n_{11} = N$. Although the overlap coefficient takes the set sizes into account, the maximum value of the score can be achieved even if there is a large discrepancy between the sets, as long as one of them is a complete subset of the other. Thus, we conclude that this score is actually not a good measure of stability.

2) Adjusted Positive Overlap (APO): Inspired by the general score in Equation 8, we adjust Equation 15 and derive stability score S_{APO}:

$$S_{APO}(W_i', W_i^{''}) = \frac{n_{11} - \mathbb{E}(n_{11})}{\min(n_{1*}, n_{*1}) - \mathbb{E}(n_{11})}, \quad (16)$$

where $n_{1*} = |\{x_{ij}|W_{ij}' = 1\}|, n_{*1} = |\{x_{ij}|W_{ij}^{''} = 1\}|$. Considering $n_{11} \sim B(N,p)$ is a binomially distributed random variable with probability p, and N is the total number of instances in the bag, the expected mean value is:

$$\mathbb{E}(n_{11}) = pN$$

where p is the probability of positive labels from both W_i' and from $W_i^{''}$, denoted with $p^{'}$ and $p^{''}$:

$$\mathbb{E}(k) = p^{'}p^{''}N$$

In addition, we know that: $p' = \frac{n_{1*}}{N}$ and $p'' = \frac{n_{*1}}{N}$.

Limits: $S_{APO} \in [1-N, 1]$. $S_{APO} = 1 - N$ for full disagreement when $n_{01} = N - 1, n_{10} = 1$. With such a limit, the score diverges, which makes it less suitable for measuring stability. Moreover, similarly to S_{PO}, S_{APO} is 1 as long as one of the sets is a complete subset of the other one. Thus, this measure is discarded as well as it is not a suitable measure of stability.

3) Heuristic Adjustment of Positive Jaccard (HAPJ): First, let us assume an extreme scenario. A bag is predicted with distinct positive instances by two classifiers, such that there is no overlap between the two sets of positive predictions. However, if the total positive predictions from both classifiers $n_{*1} + n_{1*}$ are more than the total instances in a bag N, as $n_{*1}+n_{1*} > N$, then it must be that at least $N-(n_{*1}+n_{1*})$ are positive predictions, which are overlapping. So the correction with such a heuristic is:

$$S_{HAPJ}(W_i', W_i^{''}) = \\ = \frac{n_{11} - \max(n_{*1} + n_{1*} - N, 0)}{n_{10} + n_{01} + n_{11} - \max(n_{*1} + n_{1*} - N, 0)} \quad (17)$$

Limits: $S_{HAPJ} \in [0,1]$. S_{HAPJ} yields a maximum value 1 as long as $n_{11} \leq n_{00}$, and $n_{10} = n_{01} = 0$. Minimum S_{HAPJ} can be attained for $n_{11} = 0$, and $n_{10} > 0 \vee n_{01} > 0$. Additionally, in cases where all the predictions from both classifiers agree: for $n_{00}/n_{11} = N$, then S_{HAPJ} = `NaN`. It is important to note that the metric is suitable for extreme cases when the positive instance predictions are abundant. Else, if the positive predictions from both sets are less or equal than N, the metric is equal to positive Jaccard, which was proposed initially by Cheplygina et al. [12].

4) Adjusted Positive Jaccard (APJ): In a similar manner to Equation (8), the positive Jaccard index can be corrected for chance:

$$S_{APJ}(W_i', W_i^{''}) = \frac{n_{11} - \mathbb{E}(n_{11})}{n_{10} + n_{01} + n_{11} - \mathbb{E}(n_{11})} \quad (18)$$

where $n_{11} \sim B(N,p)$, n_{11} is a binomially distributed random variable with probability p, and N is the total number of instances in the bag. The expected mean value is

$$\mathbb{E}(n_{11}) = \frac{n_{1*}n_{*1}}{N}$$

.

Limits: $S_{APJ} \in [-0.333, 1]$. Exact proof of the limits can be found in Appendix A. The minimum is attained at $n_{01} = n_{10} = 1$ and $n_{11} = n_{00} = 0$, while the maximum value is achieved for $n_{01} = n_{10} = 0$ and $n_{11} = n_{00} = 1$. It is important to note that the score is undefined for the case where all the predictions from both classifiers agree: for $n_{00}/n_{11} = N$ then S_{APJ} = `NaN`. However, for another extreme case where one of the classifiers predicts only one class, while the other classifier yields only predictions from the contrary class (n_{01}/n_{10}= N), then the score S_{APJ} is 0. Although derived in another way, we acknowledge that adjusted positive Jaccard resembles Cohen's Kappa statistics [42] on positive instances.

5) Adjusted Jaccard (AJ): With the correction for chance, the overlap of positive and negative instances can be a relevant stability measure. This can be very suitable to imbalanced bags with numerous negative instances. The correction decreases the superficial inflation expected due to the negative overlap.

$$S_{AJ}(W_i', W_i^{''}) = \\ = \frac{n_{11} + n_{00} - \mathbb{E}(n_{11}) - \mathbb{E}(n_{00})}{n_{00} + n_{01} + n_{10} + n_{11} - \mathbb{E}(n_{11}) - \mathbb{E}(n_{00})} \quad (19)$$

with $n_{11}, n_{00} \sim B(N,p)$ and $\mathbb{E}(n_{11})$ and $\mathbb{E}(n_{00})$ as:

$$\mathbb{E}(n_{11}) = \frac{n_{1*}n_{*1}}{N}$$

$$\mathbb{E}(n_{00}) = \frac{n_{0*}n_{*0}}{N}$$

Limits: $S_{AJ} \in [-1, 1]$. The minimum is attained at $n_{01} = n_{10} = 1$ and $n_{11} = n_{00} = 0$. The maximum value can be achieved for $n_{01} = n_{10} = 0$ and $n_{11}, n_{00} > 0$ (proof of the limits can be found in the appendix A). It is important to note that the score is undefined for the case where all the predictions from both classifiers agree: for $n_{00}/n_{11} = N$ then S_{AJ} =`NaN`. However, for another extreme case where one of the classifiers predicts only one class, while the other classifier yields only predictions from the contrary class, then the score S_{AJ} is 0. Finally, we recognize that adjusted Jaccard is identical to Cohen's Kappa statistics.

All of the aforementioned scores comply with the properties defined by Cheplygina et al. [12] and Kuncheva [26]. Firstly, $S_{APJ}, S_{AHPJ}, S_{AJ}$ are monotonically increasing, where S_{AJP} and S_{AHPJ} grow with positive prediction agreement, while for S_{AJ} the monotonic increase is with respect to positive and negative prediction agreement. Secondly, the scores have limits, although each of them has a distinct lower limit. In addition, all scores have a correction for chance. Lastly, stability can be measured in an unsupervised manner as no ground truth labels are needed.

B. Stability score for probability predictions

The previous measurements are highly appropriate for algorithms predicting bounding boxes, but classifiers yielding probability predictions cannot use them directly. Instead, probabilities are aggregated to binary labels based on an arbitrary threshold (e.g. 0.5). The major drawback of a threshold is that everything, above (or below) it, is treated equally, and essential raw prediction data is lost, preventing accurate evaluation between two raw sets. Thus, we propose the following methods for stability measures for probability predictions. An overview of the stability scores is in Table Ib.

1) Pearson's product moment correlation coefficient: Pearson correlation coefficient [43] measures correlation with respect to the linearity between pairs of data points. The coefficient relies, however, on the assumption that the data points have normal distribution. However, instances in a bag may not have a normal distribution, making Pearson correlation coefficient rather inaccurate.

2) Spearman rank correlation coefficient: Unlike Pearson, Spearman rank correlation coefficient [44] is a nonparametric test, making no assumptions about the distribution of the underlying data. The coefficient does not use raw data, but ranks the observations to evaluate the correlation coefficient. Furthermore, this test is capable of detecting monotonic relationships between the observations without assuming their normal distribution. The formula of Spearman rank (see in Table I) does not provide intuitive explanation of the meaning of the score, but Spearman rank measures deviations between the ranks. From the formula it becomes apparent that the score is sensitive to small number of large deviations between ranks; and it is not harsh on high number of small deviations between ranks. For the medical imaging context, small deviations in the prediction ranks are tolerated between classifiers, as no two classifiers will be exactly the same. However, large deviations between ranks would potentially mean conflicting localization of abnormality, further increasing the suitability of Spearman coefficient. Thus Spearman coefficient is regarded as a good stability score.

3) Kendall's tau: Similar to Spearman rank correlation, Kendall's tau [45] is a non-parametric correlation test. It measures correlation in terms of concordant pairs.
A pair (a_r, b_r) is defined as concordant with respect to another pair (a_s, b_s) when: $(a_r - a_s)(b_r - b_s) > 0$. Analogously, (a_r, b_r) is discordant if $(a_r - a_s)(b_r - b_s) < 0$. The Kendall's tau score can be seen as the ratio between concordant and discordant pairs. Measuring concordant and discordant pairs in the context of stability can be quite intuitive. The score measures whether a classifier ranks instances as another classifier with the same relative probability to the other instances in the bag. As long as one classifier tends to give repeatedly lower or higher probabilities to all the instances within the bag, the tau correlation is not going to be influenced. Furthermore, having large value differences within the pair is not influencing the correlation.

C. Kappa Weakness

As earlier noted, the adjusted positive Jaccard and the adjusted Jaccard are closely related to Cohen's kappa statistic. And that is why it is important to explain a peculiar behavior of Cohen's kappa, stated in paradox cases [46].
Feinstein and Cicchetti [46] describe the phenomena in terms of symmetry and balance. Symmetry is a property of two classifiers, where both classifiers predict *the same ratio* to positive, and to negative labels. Balance is the property where both classifiers predict each class with *equal ratio*. So Feinstein and Cicchetti [46] show that in settings of symmetrical classifiers ($n_{0*} = n_{*0}$, and consequently $n_{1*} = n_{*1}$), the kappa statistic considerably penalizes imbalanced classifiers ($n_{0*} >> n_{1*}$ or $n_{0*} >> n_{1*}$) for the same number of agreement instances. Thus, a paradox arises. The second paradox is observed in imbalanced, but asymmetrical settings. This occurs when each classifier predicts predominantly one class, but the predominant class of the two classifiers is not the same. The imbalance in classes leads to a minimal correction of the expected value and as a results, the kappa value is higher than in settings of symmetrical imbalance.

Table II shows the influence of balance to the adjusted Jaccard and the adjusted positive Jaccard for the same positive agreement. Table IIa introduces the case of symmetrical balanced classifiers, while Table IIb shows symmetrical imbalance classifiers. Both stability scores are demonstrating lower scores in imbalanced settings. Table III, on the other hand, shows the influence of symmetry on S_{AJ} and S_{APJ} for fixed agreement. Again the behavior correspond to the phenomena described - symmetrical imbalanced receive lower score than asymmetrical imbalanced classifiers.

In conclusion, the adjusted Jaccard and the positive Jaccard exhibit the similar weakening behavior as kappa in relation with symmetry and balance of the confusion matrix of the predictions. The paradoxical situation described earlier makes an interpretation of kappa value problematic [46]–[49]. Additionally, while the conflicting cases are caused by the correction introduced in the kappa statistic, Cicchetti and Feinstein [48] and Hoehler [49] clearly state the significance of the correction. Furthermore, to mitigate the counter-intuitive behaviour, Byrt et al. [47] suggest an alternative of the Cohen's kappa, correcting for symmetry and balance. However, symmetry-free and balance-free agreements are rather theoretical, changing the original situation measured [49]. Finally, considering more information additional to the kappa value helps understanding the presence of balance and symmetry

TABLE I: Overview of stability indices for (a) binary predictions (b) probability predictions

(a)

Stability Score	Formula	Strengths	Weaknesses
Overlap	$\frac{\text{Observed overlap}^+}{\lvert\text{Smaller positive set}\rvert}$	• considers the set sizes	• does not measure the discrepancy between the sets - not considering total number of n01 and n10 • it rather measures if a set is a complete subset of another one;
Adjusted Overlap	$\frac{\text{Observed overlap}^+ - \mathbb{E}(\text{overlap}^+)}{\lvert\text{Smaller positive set }\rvert - \mathbb{E}(\text{overlap}^+)}$	• correction for chance	• measures if a set is a complete subset of another one, similarly to its counterpart • exhibits divergent behaviour
Positive Jaccard	$\frac{\text{Observed overlap}^+}{\text{Max overlap}^+}$	• interested in total number of agreeing and disagreeing instances	• not considering the distribution of the agreement and disagreement
Heuristic Adjustment of Positive Jaccard	$\frac{\text{Observed overlap}^+ - \mathbb{E}(\text{overlap}^+)}{\text{Max overlap}^+ - \mathbb{E}(\text{overlap}^+)}$	• it adds a correction for positive Jaccard • it is a lower bound correction	• it is for extreme cases where the positive instance predictions are abundant; • if positive instance predictions are scarce, then no correction is applied and the score is the same as positive Jaccard
Adjusted Positive Jaccard	$\frac{\text{Observed overlap}^+ - \mathbb{E}(\text{overlap}^+)}{\text{Max overlap}^+ - \mathbb{E}(\text{overlap}^+)}$	• exhibit similar behaviour like Jaccard but slightly lower indices • corrects for overlap by chance	-
Adjusted Jaccard	$\frac{\text{Observed overlap} - \mathbb{E}(\text{overlap})}{\text{Max overlap} - \mathbb{E}(\text{overlap})}$	• considers positive and negative instances but corrects for chance	-

(b)

Pearson's Product Moment Correlation Coefficient	$\frac{\sum_j^N (W'_{ij} - \bar{W}'_i)(W''_{ij} - \bar{W}''_i)}{\sqrt{\sum_j^N (W'_{ij} - \bar{W}'_i)^2}\sqrt{\sum_j^N (W''_{ij} - \bar{W}''_i)^2}}$	• tests linearity	• assumes normality of the data distribution
Spearman Rank Correlation Coefficient	$1 - \frac{6\sum_j^N (rank(W'_{ij}) - rank(W''_{ij}))^2}{N(N^2-1)}$	• tests monotonicity between observations • non parametric coefficient • small difference in ranks are tolerable • correlation plummets at few but large deviations in prediction ranks	• not a good test if two classifiers detect same region as anomalous, but one of the classifiers consistently assigns higher/lower predictions than another classifier
Kendall's Tau coefficient	$\frac{\text{\# concordant pairs} - \text{\# discordant pairs}}{\binom{N}{2}}$	• examines the number of concordant and discordant pairs relative to the pair	• not influenced by the magnitude of the differences between prediction values

Classifier 1		Classifier 2		
		0	**1**	Total
	0	40	10	50
	1	10	40	50
	Total	50	50	100

(a)

Classifier 1		Classifier 2		
		0	**1**	Total
	0	0	30	30
	1	30	40	70
	Total	30	70	100

(b)

TABLE II: Balanced classifiers with $S_{AJ} = 0.6$ and $S_{APJ} = 0.43$ vs imbalanced classifiers with $S_{AJ} = -0.43$ and $S_{APJ} = -0.18$

Classifier 1		Classifier 2		
		0	**1**	Total
	0	20	20	40
	1	20	40	60
	Total	40	60	100

(a)

Classifier 1		Classifier 2		
		0	**1**	Total
	0	20	35	55
	1	5	40	55
	Total	25	75	100

(b)

TABLE III: Symmetrical imbalanced classifiers with $S_{AJ} = 0.17$ and $S_{APJ} = 0.09$ vs Asymmetrical imbalanced with $S_{AJ} = 0.24$ and $S_{APJ} = 0.14$

[47], [48]. The additional scores, applicable for the adjusted Jaccard and the adjusted positive Jaccard, are total agreement ratio (TAR), and positive agreement ratio (PAR) and negative agreement ratio (NAR) [48]:

$$TAR = \frac{n_{11}}{N} \tag{20}$$

$$PAR = \frac{2n_{11}}{2n_{11} + n01 + n_{10}} \tag{21}$$

$$NAR = \frac{2n_{00}}{2n_{00} + n01 + n_{10}} \tag{22}$$

Additionally, the difference between the positive and the negative agreement ratio as well as $n_{0*} - n_{1*}$ are computed to establish the presence of balance and symmetry and how they influence the result.

D. Learning from segmented images

To help neural networks recognize the region of interest, learning from segmented images can be achieved in a supervised manner [32]. This is achieved by employing separate pooling methods for segmented and non-segmented images. The pooling operators in Equation 1 to 5 are used for deriving bag predictions of non-segmented images. For segmented images, the pooling method is changed such that the core of the operator is preserved, but the pooling is executed in a supervised manner. Li et al. use a modified `NOR` (Equation 5) [32]:

$$\hat{Y}_i = \prod_{j \in N_i} \hat{y}_{ij} \prod_{j \in N_i \setminus S_i} \hat{y}_{ij}, \tag{23}$$

where S_i is the set of instances within a segmentation, and $S_i = \{x_{ij} | y_{ij} = 1\}$.
Later in our experiments, we examine the difference in the stability results between different pooling operators, and it is important to synchronize the pooling of segmented images with the global pooling operator. `Max` pooling (Equation 1) for images with available instance labels is:

$$\hat{Y}_i = \max_{j \in S_i} (\hat{y}_{ij})$$

`Mean` pooling (Equation 2) for segmented images is modified to:

$$\hat{Y}_i = \frac{1}{N_i} \left(\sum_{j \in S_i} \hat{y}_{ij} + \sum_{j \in N_i \setminus S_i} (1 - \hat{y}_{ij}) \right)$$

`LSE` pooling (Equation 4) for images with available local labels is:

$$\hat{Y}_i = \frac{1}{r} \log \left[\frac{1}{N} \left(\sum_{j \in S_i} \exp\left(r \hat{y}_{ij}\right) + \sum_{j \in N_i \setminus S_i} \exp\left(r(1 - \hat{y}_{ij})\right) \right) \right]$$

IV. Experimental Setup

In this section we introduce to the architecture, datasets and exact details for the conduction of the following experiments.

A. Model

A good representative architecture in MIL should be chosen as the baseline. Another requirement when choosing a baseline architecture is its capability for predicting instance labels, since we are exploring the stability of instance predictions. However, we acknowledge that performance of a single architecture cannot represent all deep learning architectures for MIL.

Li et al. [32] describes such a common architecture. The model uses a pre-trained ResNet [50] for feature extraction, followed by a max pooling layer which preserves the features locality and results in down-sampled features of instances. Subsequently, two convolutional layers are used. Figure 3 shows the building blocks of the architecture. The novelty resides in training from images with or without annotated segmentation via weak supervision. Including the limited number of segmented images demonstrated a clear advantage in terms of performance [32]. Finally, the model yields a prediction on instance level, which is consequently aggregated to a bag label. In addition, the model's performance is tested on a publicly available dataset [51] which eases the replication of the model.

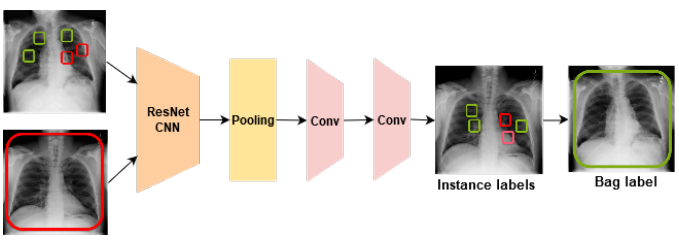


Fig. 3: Summarized view on the architecture proposed in [32].

The original model solves multi-label classification on 14 different lung diseases for every image. Our re-implementation of the model based on the paper, however, focuses on binary classification. Finally, unlike the original architecture, we do not use L2 regularization as the model did not suffer from overfitting.

B. Datasets

The following datasets are used in the experiments as they are public and some provide images with annotated instances, which can be used in the baseline model:

NIH Chest X-Ray Dataset [51], referred to as X-Ray dataset*:* comprised of X-Rays with 14 common thorax disease categories (cardiomegaly, pneumonia, etc.), includes more than 100,000 multi-label scans from 30,000 patients. While the majority of the data is annotated on an image-level, little under 1000 images are annotated with exact coordinates of a lung disease. This is the dataset used in [32].

In binary classification settings, the X-Ray dataset population is no longer representative for the chosen classes because it becomes overwhelmingly negative. Instead, we consider all images of patients who have at least a single positive image for the specified class category. In this way, the class imbalance is preserved without having an overpowering number of negative labels. This dataset includes few segmentation images for some classes. The main prerequisite for choice of class prediction is the availability of segmentation images, which are used in training and evaluation phase. Hence, the first classification is on cardiomegaly detection. In addition, we have arbitrarily chosen another class - effusion detection to test and generalize the model performance. This class is not used in the stability experiments.

MURA dataset [52]: consists of a large number of X-Ray imagesTbones. Some of the present classes are elbow, finger, forearm, and hand. In total there are 7 classes. This dataset contains only image-level annotations. From the MURA dataset we have arbitrarily chosen to conduct the experiments on the shoulder class.

Pascal VOC, from 2005 [53]: a relatively small dataset of about 1000 photos of cars, bicycles, motorcycles and people. We have decided to predict cars to yield a binary problem. Some images have segmentations, which are used for evaluating the instance performance.

Table IV shows the exact statistics of bags and instances in each dataset and class.

C. Preprocessing

Firstly, the accepted input size is 512 $\times$512. Images from the X-Ray dataset originally had a size of 1024 $\times$1024, so they are decreased to size of 512 $\times$512 pixels. Images from MURA and Pascal VOC, on the other hand, have varying size. Images with at least one axis larger than 512 pixels are first proportionately scaled down. Their larger side is set to 512 pixels, and zero-padding is used for the smaller side. For images smaller than 512 pixels on both axes we have used zero-padding for enlarging the image. Furthermore, all the images are normalized between $[-1, 1]$ values.

Dataset	Class	Bags	Instances
X-Ray	Cardiomegaly	2722+, 14742-	146 bags $\times$ 256 inst.
X-Ray	Effusion	13056+, 37693-	149 bags $\times$ 256 inst.
MURA	Shoulder	4446+, 4211-	-
Pascal VOC	Cars	335+, 826-	-

TABLE IV: Dataset population for each class. The effusion class is only used for generalization of the baseline results.

D. Train-test splitting

With far less observations for the images without segmentation it is decided that it is more beneficial to keep 80% of the data for training and 20% for testing. Similarly, the images with segmentation are limited so they are divided into 80%/20% training to test split, as described to be most beneficial [32].

The publishers of the MURA dataset publicly provide training and validation set. The validation set is used as testing set, and the original training set we divide into training and validation with 80%/20% ratio. Furthermore, patients with multiple scans in MURA and X-Ray are only in the training or in the testing dataset.

E. Method for measuring stability

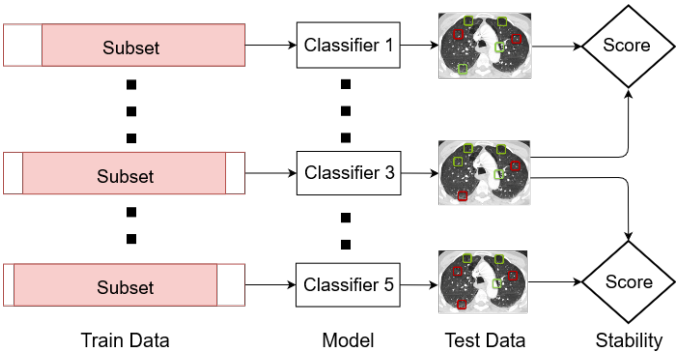


Fig. 4: 5 classifiers are trained with very similar dataset. The predictions of each classifier is compared with the other 4 to compute the stability score

To measure stability, we train 5 very identical classifiers. To preserve similarity of the training set, each classifier uses 95% subset of the whole training data to train. The test dataset is fixed. The stability scores are pairwise measurements, so a score is a result of comparing the predictions of 2 classifiers on the same image. Ultimately, comparing all classifiers' predictions for each image derives the stability score of the each image across all classifiers.

F. Choice of Stability Score

After initial analysis we have eliminated some of the possibilities for measuring stability. So for binary predictions stability can be measured with the adjusted positive Jaccard and the adjusted Jaccard, and probability predictions can be measured with Kendall's tau and Spearman correlation coefficient. We perform correlation analysis, similar to the experiments performed by Taha and Hanbury [54], between the stability scores. We used the pairwise Pearson's correlation and the Spearman rank correlation coefficients to examine the stability indices. Results yield correlation coefficients of 0.98 and 0.99, between adjusted positive Jaccard and adjusted Jaccard, and between Kendall's tau and Spearman, respectively. The extreme pairwise correlation coefficients allow us to conclude that the choice of a score within each group is rather superficial - adjusted positive Jaccard is interchangeable with respect to adjusted Jaccard, and Kendall's tau is no more informative than Spearman rank correlation. Thus no information loss occurs if one of the score within each group is ignored. Eventually, we have chosen to use adjusted positive Jaccard for measuring stability and Spearman correlation coefficient for the consecutive experiments.

G. Stability variability with threshold

The current classifier yields probability instance predictions, and that is why, a stability score for binary and probability predictions can be used. However, using a stability score for binary label requires first to convert the probability predictions to binary. The choice of threshold for this conversion influences the stability score, but the degree of influence it not known. Instead of choosing an arbitrary threshold for converting to binary predictions, first we examine the stability sensitivity to threshold choice. Performing experiments where the threshold for binary conversion is changed from 0.1 to 0.9 with a step size of 0.1 show that there is a general tendency of the stability to decrease with increase of the binary threshold. However, the observed stability decrease is quite steady, and no rapid jumps are observed. So the results confirm that a threshold of 0.5 is a reasonable choice, which is used further in the experiments.

V. Experiments and Results

In this section the stability is examined on several datasets. Furthermore, we investigate the stability in relation to the performance on bag and instance level, and the effect of the common pooling operators.

A. Baseline

Based on the evaluation metrics in Section II-E, Table V shows the results obtained from the baseline model on a 5-fold cross validation. We use `T(binary)=0.5` in Equation 10 for converting to binary predictions, and `T(Jacc)=0.1` in Equation 14. In addition, the table includes results of Li et al. [32] on the architecture with the X-Ray dataset with the same values of `T(binary)` and `T(Jacc)`. Because we use the architecture for a single class prediction, the model is not exactly the same as the described by Li et al. That is why the results of the two experiments are not directly comparable. However, their results are a reference to the model we have trained.

Class	AUC	DICE *	Accuracy *
Cardiomegaly	0.90 ± 0.02	0.72 ± 0.07	0.97 ± 0.05
Cardiomegaly Reference. [32]	0.80 ± 0.01	-	0.98 ± 0.02
Effusion	0.92 ± 0.01	0.33 ± 0.02	0.74 ± 0.08
Effusion Reference [32]	0.87 ± 0.01	-	0.87 ± 0.03
Shoulder	0.84 ± 0.01	-	-
Cars	0.99 ± 0.01	0.40 ± 0.03	0.82 ± 0.03

* Instance performance is measured only on images with available instance labels.

TABLE V: Performance evaluation from 5-fold cross validation. Bag performance: AUC and instance performance: DICE coefficient and accuracy from the Jaccard index. Accuracy is a result from Jaccard threshold of 0.1.

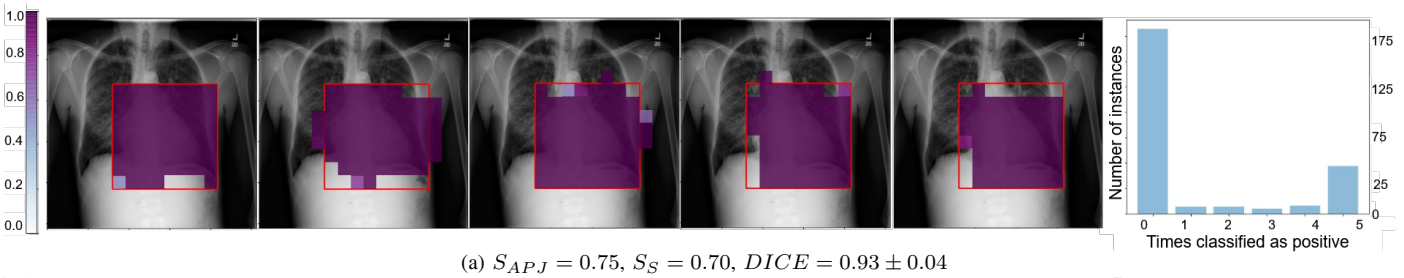
B. Illustrative example

Figure 5 shows several images and their predicted localization. Furthermore, the histogram for each image depicts the frequency of each instance to be predicted as positive. Ultimately, perfectly stable classifiers are expected to predict the same instance as positive or negative for the models trained. As a result, the histogram for an image should have the instances of each image predicted either 0 or 5 times as positive. An example of a such histogram is in Figure 5a, which shows one of the most stable bags from the X-Ray dataset.

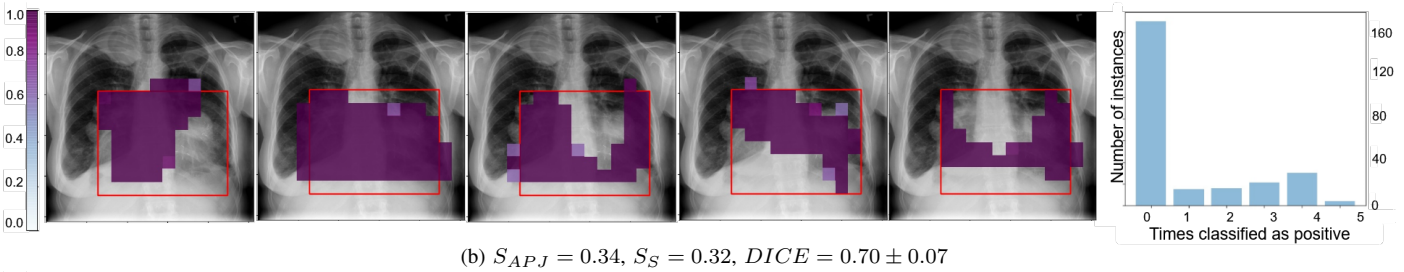
However, any instance predicted between 1 and 4 times as positive highlights the instability of the models. Such behavour can be seen in Figure 5b, where about 30% of instances are classified incosistently as positive. This examples also highlights that good localization ($DICE = 0.70 \pm 0.07$) does not lead to stable predictions ($S_{APJ} = 0.34, S_S = 0.32$).

Figure 5d shows one of the most stable bags in the MURA dataset ($S_{APJ} = 0.69$, $S_S = 0.73$). In contrast, the average bag in the MURA dataset (in Figure 5e) shows less stable behaviour, where some predictions show larger areas than the other models as abnormal.

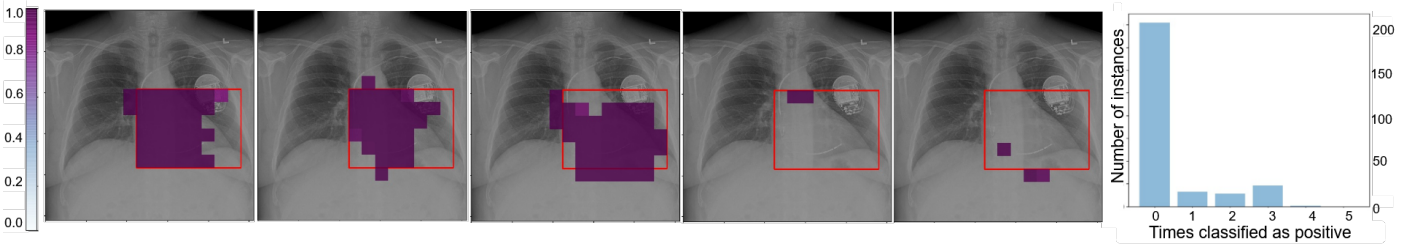
When we discuss the performance of the Pascal VOC, it is important to note that the dataset has several image sources, each with its own image dimensions. As a result there are some discrepancy in the image quality. Some of the sources provide images as the example in Figure 5f that need just a bit of padding. While other images require a lot of additional padding to, resultingas the image in Figure 5g. Figure 5f is among the most stable bags in the Pascal VOC and is well performing on instance-level ($DICE = 0.85$). Figure 5g, on the other hand, shows an average performing bag from Pascal VOC. The classifiers in this case exaggerate the locations of the car. Its S_{APJ} score is mediocre, but its S_S shows relatively high



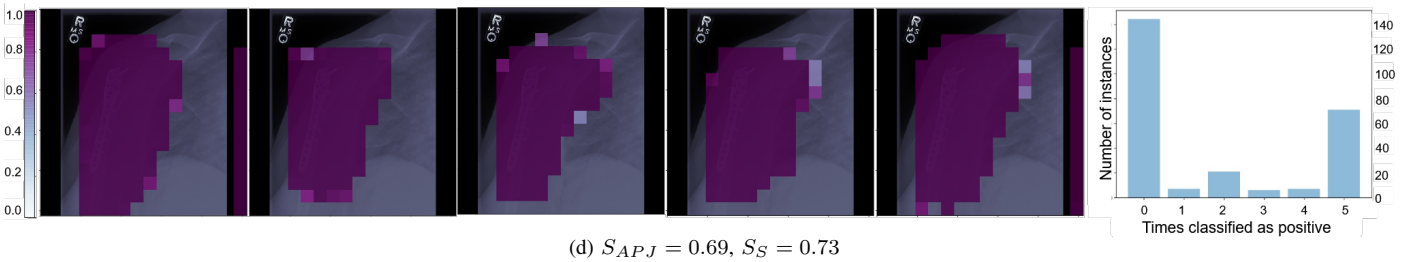
(a) $S_{APJ} = 0.75$, $S_S = 0.70$, $DICE = 0.93 \pm 0.04$



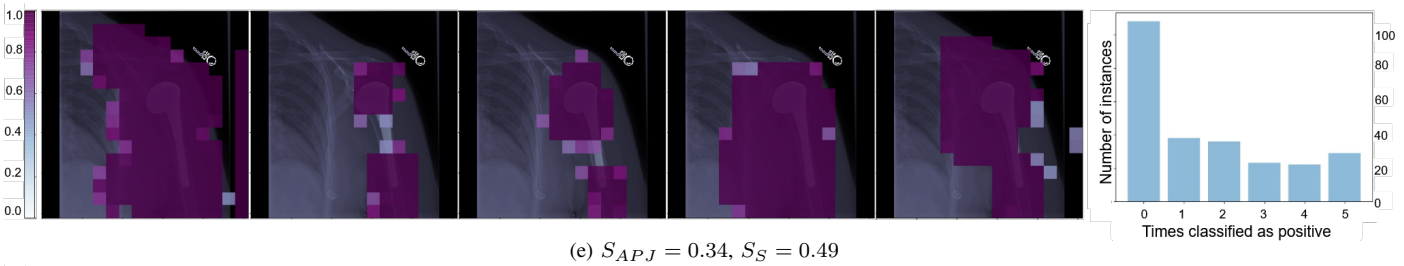
(b) $S_{APJ} = 0.34$, $S_S = 0.32$, $DICE = 0.70 \pm 0.07$



(c) $S_{APJ} = 0.17$, $S_S = 0.12$, $DICE = 0.49 \pm 0.35$

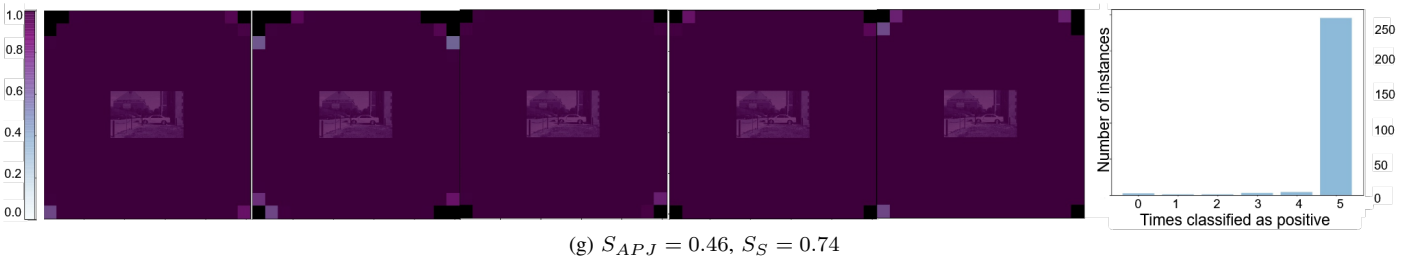


(d) $S_{APJ} = 0.69$, $S_S = 0.73$



(e) $S_{APJ} = 0.34$, $S_S = 0.49$

(f) $S_{APJ} = 0.79$, $S_S = 0.76$, $DICE = 0.85 \pm 0.02$



(g) $S_{APJ} = 0.46$, $S_S = 0.74$

Fig. 5: (a) stable bag from X-Ray, (b) bag with unstable but well localizing predictions from X-Ray, (d) stable bag from MURA, (e) average bag from MURA, (f) bag with good localization from Pascal, (g) bag from Pascal

stability (0.46 and 0.74, respectively). Furthermore, this bag fairly represents images with lower dimensions in the Pascal VOC. The predictions on these images tend to encompass the whole image and often part of the padding, as well. The two bag examples from Pascal VOC represent the difference in performance of the models on various image quality.

C. Stability and Instance Performance

The stability scores are designed as an unsupervised measure to check for consistency of instance localization. To evaluate the scores, we compare the stability scores and the instance performance for each bag. These experiments are conducted only on limited bags with available instance-level annotation. There are 29 and 41 bags with instance-level annotations in the test sets of the X-Ray and in the Pascal VOC, respectively.

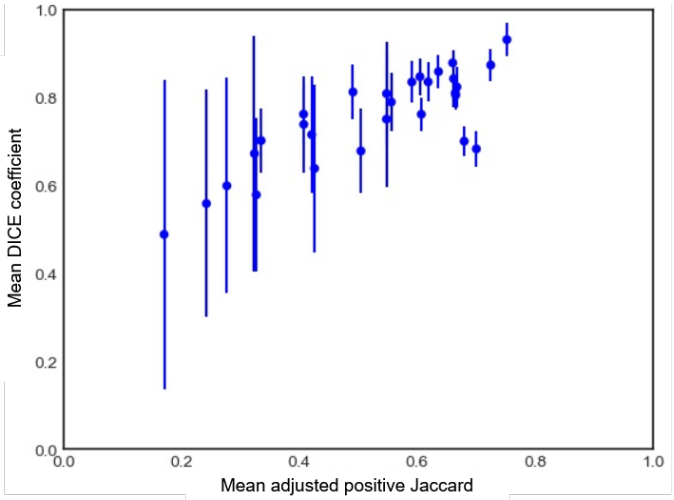
Firstly, we compute for each bag the DICE coefficient (Equation 12) between the ground truth instance labels and the predictions. Next, the stability score per bag is computed. It is important to note that the stability is a pairwise measurement - so each bag has a stability score for each combination of two classifiers. That means that per bag there are 10 stability scores for all combinations of the 5 classifiers examined. On the other hand, there are 5 DICE coefficients - 1 for each classifier and the ground truth annotation. The mean results are derived by averaging the available measurements (10 stability scores, and 5 instance DICE scores).

The results from the X-Ray dataset are shown in Figure 6 with stability in terms of adjusted positive Jaccard (Figure 6a), and in terms of Spearman correlation (Figure 6b). The initial deduction is the linearity between both of the stability scores and the DICE coefficient. On closer inspection, we differentiate two clusters of bags. However, the cluster separation is more pronounced and concise with adjusted positive Jaccard. Nonetheless, one of the clusters identified is in the central upper part of the figures. Bags situated there have significant mean DICE coefficient, and relatively higher stability score. What is more important is that these bags exhibit lower standard deviation of DICE than the rest, showing that all 5 classifiers perform relatively well on these bags.

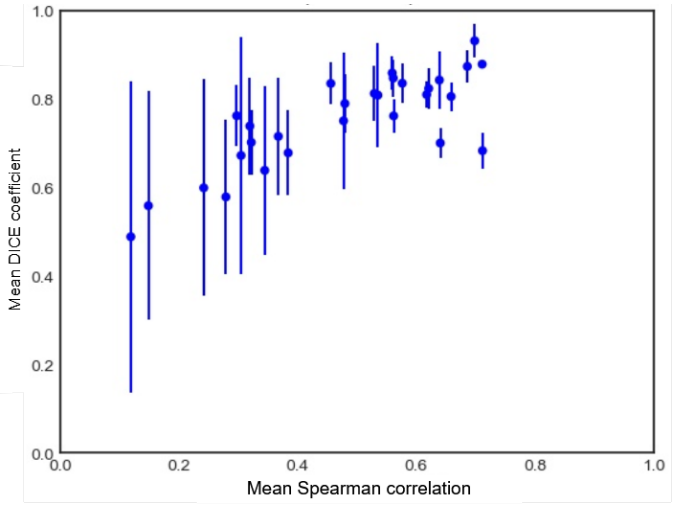
The other cluster is situated on the left side of the figures. The bags populating it look more heterogeneous with respect to mean DICE. Bags in this cluster have mean DICE from low to reasonably high, but majority of them show large standard deviation of DICE coefficient. The large standard deviation is evidence that some of the classifiers are localizing well, while others are performing worse on the bag.

Figure 5c is a prime example of a bag with high standard deviation. From the figure it is apparent that the contributors to the DICE score are only 2 or 3 classifiers. The stability scores are low ($S_{APJ} = 0.17, S_S = 0.12$) and it is good to see that discrepancy between classifiers is reflected in lower stability score.

The results in Figure 7 show bags from the Pascal VOC dataset and their average stability score and the corresponding DICE coefficient. Similar to the X-Ray dataset, Figure 7a suggest some linearity between localization performance and the stability scores. We observe that bags with higher DICE



(a)



(b)

Fig. 6: Stability score of each bag against its instance performance in the X-Ray dataset.

(and small standard deviation) tend to have a higher stability score S_{APJ}. The plot of S_S (Figure 7b) shows that all the bags are stable, independent of their localization performance. In both figures we can differentiate two clusters. One of them has higher DICE scores and slightly higher standard deviation, and the other one has lower DICE scores and almost 0 standard deviation. The second cluster receives nearly the same stability scores as bags in the first cluster despite of their difference in DICE. Further analysis shows that the second cluster with low DICE consists of images with lower quality. These images receive the consistent localization, but the predictions encompass also the wrong place - an example is Figure 5g.

Finally, Table VI shows the average results of the stability scores against the instance performance. From the aggregated results it becomes clear that the X-Ray dataset is relatively

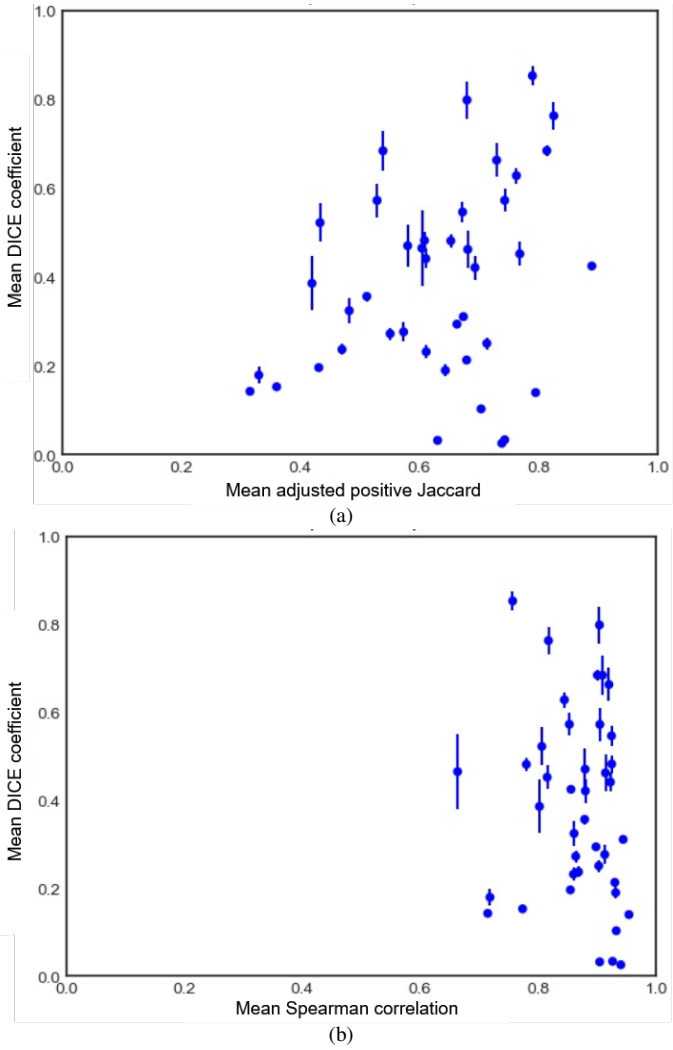


Fig. 7: Stability score of each bag against its instance performance in the Pascal VOC dataset.

high on the localization, but the predictions are not very stable. On the other hand, localization performance of Pascal VOC is poorer than on the X-Ray, but the localizations are more far more stable.

Dataset	S_{APJ}	S_S	DICE coefficient
X-Ray	0.53 ± 0.16	0.48 ± 0.17	0.75 ± 0.06
Pascal VOC	0.63 ± 0.14	0.87 ± 0.07	0.38 ± 0.21

TABLE VI: Mean stability scores and instance performance on images with available instance-level annotation. *All results here are derived only on images with available segmentation labels. That is why results of S_{ADJ} and S_S are not the same as in Table VII.*

D. Stability and bag performance

It is interesting to investigate the relation between the stability and the performance on bag level - shown in Figure 8. The plot shows some linearity between the stability and the bag performance, but the correlation cannot conclude any causality between the two. In addition, all datasets have decent to very good performance on bag level (bag AUC above 0.8). However, their stability scores are not high. Models from the X-Ray and MURA have bag AUC of about 0.8, but their stability scores are in the low to mediocre range. The models from the Pascal VOC dataset perfectly perform on bag level with AUC of 1, but the stability score show medium to substantial stability. So, the stability scores are not as high as the bag AUC, which implies the need for more stable algorithms. Finally, doing evaluation only on bag level may be misleading. It is important to incorporate stability in the evaluation metrics and use its score to support the assessment of the quality of the localization.

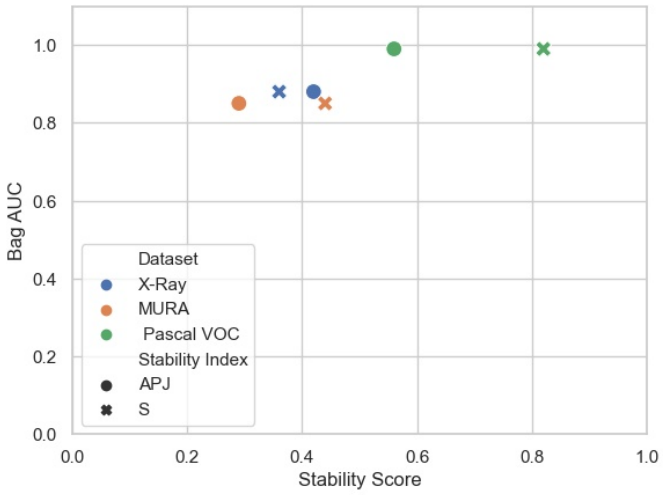


Fig. 8: Average stability and average bag AUC

E. Pooling Operators Stability

Initial experiments are performed on the `NOR` (Equation 5) pooling operator. However, since the aggregating function derives the bag prediction, the choice of pooling operator directly influences the training and the learning. In this section we investigate how models with different pooling operators perform with respect to the stability. Table VII shows the results of the stability scores. For completeness, we include the performance on bag level. Furthermore, we include the instance performance based on the limited available bags with instance-level annotation. These bags are under 1% in X-Ray and under 25% in Pascal VOC. So the instance performance provides a glimpse about the models' localization, but it is not representative.

For the experiments, we test `NOR` (Equation 5), `Mean` (Equation 2), `Max` (Equation 1) and flexible `LSE` operator (Equation 4). We use two values of the hyperparameter with

Dataset	Pooling Operator	S_{APJ} *	S_S *	bag AUC **	DICE **	split bag AUC +	split DICE +
X-Ray	NOR	0.42 ± 0.20	0.36 ± 0.20	0.88 ± 0.02	**0.75 ± 0.06**	0.87	0.72
X-Ray	Mean	0.31 ± 0.17	0.52 ± 0.22	0.88 ± 0.03	0.56 ± 0.06	0.88	0.52
X-Ray	LSE, r=0.1	**0.49 ± 0.16**	**0.68 ± 0.21**	**0.91 ± 0.01**	0.59 ± 0.10	0.89	0.60
X-Ray	LSE, r=1	0.37 ± 0.18	0.58 ± 0.23	0.90 ± 0.02	0.58 ± 0.10	0.90	0.56
X-Ray	Max	0.05 ± 0.07	0.15 ± 0.10	0.87 ± 0.02	0.11 ± 0.04	0.87	0.12
MURA	NOR	0.29 ± 0.19	0.44 ± 0.17	**0.85 ±0.01**	-	0.85	-
MURA	Mean	0.12 ± 0.09	0.41 ± 0.14	0.85 ±0.01	-	0.84	-
MURA	LSE, r=0.1	0.13 ± 0.10	0.40 ± 0.15	0.85 ± 0.01	-	0.84	-
MURA	LSE, r=1	**0.29 ± 0.14**	**0.48 ± 0.14**	0.84 ± 0.01	-	0.84	-
MURA	Max	0.11 ± 0.15	0.24 ± 0.12	0.82 ± 0.01	-	0.82	-
Pascal VOC	NOR	**0.56 ± 0.14**	0.82 ± 0.09	0.99 ± 0.00	0.40 ± 0.03	0.98	0.39
Pascal VOC	Mean	0.28 ± 0.27	0.82 ± 0.10	0.99 ± 0.00	0.24 ± 0.03	0.99	0.23
Pascal VOC	LSE, r=0.1	0.29± 0.26	0.83 ± 0.10	0.99 ± 0.00	0.24 ± 0.03	0.98	0.03
Pascal VOC	LSE, r=1	0.36 ± 0.29	**0.85 ± 0.09**	0.99 ± 0.00	0.26 ± 0.03	0.98	0.25
Pascal VOC	Max	0.52 ± 0.29	0.64 ± 0.11	**0.99 ± 0.00**	**0.57 ± 0.02**	1.00	0.57

* Score derived from positive bags.
** Score derived from 5-fold CV
+ Score on the specific fold used for training subsets

TABLE VII: Results of different pooling methods. Stability scores and bag AUC are computed on the whole test set, but DICE coefficient is derived only from bags with instance labels.

the flexible LSE operator (Equation 4). With $r = 1$ the pooling operator is an estimation of the Max, and with $r = 0.1$ it is an estimation of the Mean.

Results vary greatly between different operators and between each dataset. The most stable results on the X-Ray are achieved with the LSE, r=0.1 operator, while on the MURA dataset it is LSE, r=1. On the Pascal VOC, the results vary, but it seems that the NOR operator is the most stable. Although not best, the NOR operator has relatively good and stable predictions on S_{ADJ} on all 3 datasets. In contrast, the Max operator yields the lowest S_S stability index in all three datasets. However, the results from the other operators are not as conclusive about its performance.

Furthermore, the Mean and LSE, r=0.1 give relatively similar results, which is expected as LSE, r=0.1 is an approximation of the Mean. In a similar manner, LSE, r=1 is an approximation of Max, but there is a greater difference in the results of these two operators.

Furthermore, results show that bag performance is somewhat invariant to different operators. The instance performance, on the other hand, varies with the pooling operator. This again shows that bag performance on its own may be misleading. In addition, we find no relationship between the instance performance and stability scores. However, the results cannot be conclusive due to the limited images used for evaluating instance performance.

VI. Discussion

A. Results

Experiments studying the relation between the instance performance and the stability scores reveal important aspects of the stability scores. The relation between the stability and localization performance is different in the two datasets. However, we observe that bags, whose instances are consistently well localized (high mean DICE coefficient with small standard deviation) are with the highest S_{ADJ} and are among the highest S_S. This confirms the soundness of the stability scores.

Designing stability measures as a surrogate for unsupervised instance predictions does not mean that stable models are good at localization. Furthermore, linearity between the stability score and localization is found, but the correlation should not be confused for causality. The results from the Pascal VOC are evidence of it. In both graphs of Figure 7, there are bags with relatively high stability score, but low DICE coefficient. These are bags, where the localization is incorrect but consistent. In this regard, the score is rightfully reflecting their prediction consistency.

Finally, while good performance and stability does not imply good localization, stability is a prerequisite for good localization. Thus, low stability in models with good performance can be seen as a first indicator of inaccurate localization.

With the designed scores, we measure the stability on 3 datasets. The results demonstrate that predictions are not very stable. Furthermore, results show linearity between the bag performance and the stability. We observe that the models which perform better on a bag level are also more stable, suggesting that the stability increases together with better performance on the bag level. Despite of the linearity observed, the stability is still not high. Models with perfect bag performance are moderately stable. This highlights the need to measure stability in MIL models, and to tailor more stable models. The relation between stability and bag performance, however, contrasts with earlier findings, which show a trade-off between stability and bag performance [12]. However, differences in the results are expected due to the fundamental differences in the algorithms compared. Cheplygina et al. [12] use several algorithms, some

of which are bag embedded classifiers, while others maximize margins in the hyperplanes to determine the instance labels.

While measuring linearity between bag performance and stability, Table VII shows that the same bag performance can have varying stability according to the pooling operator used. This demonstrates the importance of choice of pooling operator. The lack of agreement between the pooling operators on all datasets suggests that there is no best/worst pooling operator. It depends on the nature of the operator, together with the type of images and localization size.

The `NOR` operator shows good stability with S_{ADJ} on all three datasets. The operator treats the instances within a bag as independent binomial variables, which does not hold. Our hypothesis, however, is that this operator compared to the others is more suited to the underlying data. From the formula of the `NOR` (Equation 5), the bag prediction becomes high probability as long as there are some positive instances. Here we assume using normalized or logarithmic prediction values, else underflow can occur. In contrast, the `Mean` operator averages all the instance predictions, so in order to have a bag label of '1', all the instances should be '1'. The operator forces more instances to be positive. So `Mean` is suitable to bags with a high number of positive instances. Otherwise `Mean` exaggerates the number of positive instances and the model may learn co-occurring noise, which introduces instability. Similar behavior is observed with the flexible `LSE` operator with $r = 0.1$, which estimates the `Mean`.

On the other hand, using `Max`, bags are predicted as '1' as long as a single instance is predicted as positive. This pooling is expected to suit bags with very few positive instances. Otherwise, it is likely that a single distinct instance is predicted as positive from all positive instances. This can lead to high bag performance, but not to stable predictions. In a similar manner, `LSE, r=1` is an approximation of `Max`, but there is a great difference in the results of these two operators. `LSE` with $r = 1$ behaves more similarly to the `Mean` and `LSE` with $r = 0.1$. A plausible reason is that `LSE, r=1` considers all instance predictions to aggregate to bag prediction, while `Max` considers only one instance prediction.

In addition, the `Max` operator has the lowest S_S stability index among all three datasets. We believe that this is a result from the combination of the metric and pooling operator. The `Max` operator selects the most active instances, and the S_S is sensitive to large prediction difference in prediction results (for more see III). Thus, the low stability results of Max highlight large differences in the activated instances.

Finally, positive instances on each bag in the three datasets are neither abundant, nor highly limited, making `Max`, `Mean`, and `LSE` operator less suitable and, thus, less stable.

Throughout all the experiments we measure stability with Spearman rank correlation coefficient and the adjusted positive Jaccard. Due to their different formulas the two scores are not identical. Pearson's correlation test shows varying degree of correlation between the two scores of different classifiers. This is evidence that the two scores are complementary. Spearman correlation is always higher than the adjusted positive Jaccard, as it considers both positive and negative instances in a bag, and it does not have the correction of S_{APJ}. However, S_S score can be much higher than S_{APJ} in some cases - such as some bags in Figure 5g and Figure 7. So one should be aware of the inflation in S_S score.

B. Limitations of the proposed stability measurements

The adjusted positive Jaccard can have an undefined value when all instances in both bags are predicted with the same label. So there are two scenarios: all instances by both classifiers are predicted as 1) positive, or as 2) negative. One can reason that both of these scenarios are trivial and that a more meaningful value can be assigned. A viable argument is that when all predictions of two classifiers are the same, then the stability score should be at its maximum, demonstrating perfect stability. On the other hand, these two scenarios are quite extreme in the context of an abnormality localization. In the case of only negative predictions by both classifiers, one may argue that there is nothing positive to compare and as such the stability score is 0. Analogously, only positive predictions by both classifiers mean 100% agreement on the localization, and as such the stability score should be 1. However, there are counter arguments as well. Assigning a stability score of 1 for only positive predictions does not consider that both classifiers may be yielding the same label on all instances for all bags, thus the model has no predictive power. The controversy of such an assignment might be evidence that replacement with more meaningful value should not be done at all. So this remains the main limitation of the stability score.

However, undefined values cannot be used in any algebraic operations, excluding them from the stability score of the whole population. That is why it is crucial to keep track of their total number when drawing conclusions. In addition, finding the cause of undefined values may provide helpful insights about the models.

The adjusted positive Jaccard corrects for chance, so it is no longer a ratio between agreement of two classifiers, which makes the score less intuitive, and any interpretation of the score should be carefully handled.

The results from the Spearman correlation coefficient are reported as is. Although a more descriptive interpretation of the coefficient values is more conclusive about the degree of correlation, the interpretation among fields varies greatly [55], making any interpretation possibly misleading.

C. Considerations

1) Measuring stability: In our work we have focused on the stability of positive bags. The reason for this is that we focus on the stability of localizing abnormalities and the main prerequisite of measuring one is its presence. Furthermore, the adjusted positive Jaccard is tailored to measure positive instance predictions, so the score is not suited to negative bags. The adjusted Jaccard could be used for measuring stability on all images, but it is designed to measure positive images. Ultimately, measuring bags with only negative instance predictions results with S_{AJ} = `NaN`, which does not provide any insights. So for measuring stability in negative images, a better suited score should be proposed.

As mentioned earlier, the strength of Spearman correlation is evaluated from the absolute value of the coefficient. However, when computing average stability, we do not comply with these instructions. The reason for this is that negative correlation indicates change in different directions between two predictions. This phenomenon in and of itself is evidence of instability. Hence, we consider that the negativity of the correlation coefficient should not be omitted.

2) Quality of datasets: While we have used several datasets to generalize our results, there are some differences between the datasets that could influence the outcome of the experiments. The X-Ray dataset includes a very limited number of images with segmentation (under 1%). These images are included in the training of the network as it has been shown to be beneficial [32]. However, in the MURA dataset, no segmentations are available and thus cannot be used during training. The Pascal VOC dataset contains masks for some images. However, we train entirely weakly supervised because the training set is quite small (under 400 car images) and keeping the ratio of segmented images of around 1% means having 3 segmented images. Such a low number of segmentation images cannot be representative and bias is introduced in the training. So no segmentation images are used. Furthermore, using no segmentation images makes the dataset more similar to the medical datasets, where segmentation images are scarcely provided.

Another difference between the datasets is the angle of the objects on the images. In the X-Ray dataset, the images always consists of a frontal chest scan. In MURA, on the other hand, the scans are done from different angles (see Figure 5d and Figure 5e). Pascal VOC contains even more variability, where images of cars are taken at various viewpoints and scales.

In the preprocessing step we use zero-padding to enlarge any image dimension, which is smaller than the desired input. We used zero-padding instead of interpolation as interpolation may distort the image patterns in the process. There are some repercussions of using distorted images - a possible outcome is that the abnormality may not be recognized. What is more, the architecture can be mislead to learn the wrong presentation. In contrast, zero-padding demonstrates no negative effect on performance [56].

Another difference is that MURA and, particularly some images in Pascal VOC, contain lower resolution images. Using padding for smaller size images means that there are fewer instances that should be activated, as the rest of the instances will be on the padding. However, lower proportion of positive instances in some positive bags may be a reason for bad performance, and maybe unstable behavior [8].

We believe that the limited number of segmented images used during training, the similarity between the images and their higher resolution can be the main reasons for better localization on the X-Ray dataset. Thus the lack of segmentation during training, the variability in scale and perspective as well as the low resolution could account for the efforts of the models to exaggerate the positive predictions in the MURA and Pascal VOC datasets.

In addition, Oakden-Rayner raises awareness about the quality of the annotation in the X-Ray dataset, where 10% to 30% lower values of positive labels are reported than actually present on the images [57]. The deviation between labels and visual content can inevitably confuse training the network, influencing both the bag performance and the stability.

D. Future and Outlook

The research in this project can evolve in several directions. Future improvements may involve the architecture, the stability score, or the loss function. However, these improvements have the ultimate goal of creating more stable and well performing on instance level models. Similarly to the `Max` operator, with supervised bounding boxes, where the bag label is equal to the prediction of the most discriminate instance, Zhu proposes taking the k largest abnormal probabilities as positive instances [58]. All the other instances $(N-k)$ are considered as negative. In this way, the pooling operator forces negative instances to have low abnormal probability, and positive instances to have high probabilities. This approach sounds quite reasonable as it may combat with predicting as many instances as positive. This solution has several caveats with it. Firstly, the choice of k requires a prior knowledge about the abnormality. Furthermore, taking fixed k instances as most discriminate assumes that an abnormality has a fixed size. That means that abnormalities scanned at different angles, or variability of the abnormality size are not accounted for. Alternative pooling could be based on the most discriminate region [59].

Carbonneau et al. [8] shows that proportion of positive rate, relation between instances, and noise in the bag are just some properties of datasets which has an influence on the performance. So it may be beneficial designing a pooling operator or a loss function where those characteristics are considered.

A common technique of boosting models' performance is using an ensemble of models [60]. A solution towards more stable classifiers may be using an ensemble of weights for determining the weights of the final model.

A more robust approach to detecting local abnormalities can be achieved with a two-stage learning. In the first stage, the most discriminate regions are detected, which are consequently used in the second stage to boost the localization [59].

VII. Conclusions

In this paper alternative ways to measure stability are presented as an unsupervised measurement of localization. The novelty resides mainly in the concept of measuring how stable predictions are when localizing abnormalities in multiple instance learning. The experiments are evaluated with a common deep learning architecture for MIL and on 3 public datasets, two of which are large medical datasets. The results show that models are not very stable, even in cases with perfect bag performance. This is evidence of the importance of measuring stability, and reporting it in addition to the evaluation metrics. Our proposed stability scores can be applied to all fields and data, as long as algorithms are trained in multiple instance learning settings. Finally, the current research on the topic is immature, so there is more to be explored and developed. Future

work should focus on exploring how to make algorithms more stable and well localizing. Source code is available at `https://github.com/romanovar/evaluation_MIL`.

REFERENCES

[1] M. Firmino, G. Angelo, H. Morais, M. da Câmara Ribeiro-Dantas, and R. Valentim, "Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy," *BioMedical Engineering OnLine*, vol. 15, 01 2016.

[2] S. Katsuragawa and K. Doi, "Computer-aided diagnosis in chest radiography," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 212 – 223, 2007, computer-aided Diagnosis (CAD) and Image-guided Decision Support. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895611107000286

[3] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 198 – 211, 2007, computer-aided Diagnosis (CAD) and Image-guided Decision Support. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895611107000262

[4] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Medical Image Analysis*, vol. 18, no. 3, pp. 591 – 604, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841514000188

[5] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Computerized Medical Imaging and Graphics*, vol. 42, pp. 44 – 50, 2015, breakthrough Technologies In Digital Pathology. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895611114001852

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, pp. 31–71, 1997.

[7] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Reviews in Biomedical Engineering*, vol. PP, 01 2017.

[8] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329 – 353, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317304065

[9] V. Cheplygina, M. de Bruijne, and J. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 5 2019, copyright 2019. Published by Elsevier B.V.

[10] M. Kandemir and F. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, vol. 42, 11 2014.

[11] G. Vanwinckelen, V. Tragante, D. Fierens, and H. Blockeel, "Instance-level accuracy versus bag-level accuracy in multi-instance learning," *Data Mining and Knowledge Discovery*, vol. 30, 05 2015.

[12] V. Cheplygina, L. Sørensen, D. M. Tax, M. de Bruijne, and M. Loog, "Label stability in multiple instance learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Cham, 2015, vol. 9349, pp. 539–546. [Online]. Available: http://link.springer.com/10.1007/978-3-319-24553-9{_}66

[13] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. Metaxas, and X. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1–1, 02 2016.

[14] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Medical Image Analysis*, vol. 43, pp. 157 – 168, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841517301524

[15] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. N. Metaxas, and X. S. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1332–1343, May 2016.

[16] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81 – 105, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370213000581

[17] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. S. Tarragó, and S. Vluymans, "Multiple instance learning," in *Springer International Publishing*, 2016.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[19] O. Maron, "Learning from ambiguity," 1998.

[20] D. Heckerman and J. S. Breese, "Causal independence for probability assessment and inference using bayesian networks," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 26, no. 6, pp. 826–831, Nov 1996.

[21] S. Tang, "Object detection based on convolutional neural network," 2015.

[22] S. Gidaris and N. Komodakis, "Locnet: Improving localization accuracy for object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[23] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 424–432.

[24] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2018, pp. 624–628.

[25] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912. [Online]. Available: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x

[26] L. I. Kuncheva, "A stability index for feature selection," in *Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, ser. AIAP'07. Anaheim, CA, USA: ACTA Press, 2007, pp. 390–395. [Online]. Available: http://dl.acm.org/citation.cfm?id=1295303.1295370

[27] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec 1985. [Online]. Available: https://doi.org/10.1007/BF01908075

[28] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104

[29] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," 2018.

[30] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? – weakly-supervised learning with convolutional neural networks," 2015.

[31] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with Convolutional Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 1713–1721, 2015.

[32] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," *2018 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00865

[33] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[35] R. Guerrero, C. Ledig, and D. Rueckert, "Manifold alignment and transfer learning for classification of alzheimer's disease," in *Machine Learning in Medical Imaging*, G. Wu, D. Zhang, and L. Zhou, Eds. Cham: Springer International Publishing, 2014, pp. 77–84.

[36] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, "Knowledge transfer for melanoma screening with deep learning," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Apr 2017. [Online]. Available: http://dx.doi.org/10.1109/ISBI.2017.7950523

[37] M. Kandemir, "Asymmetric transfer learning with deep gaussian processes," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 730–738. [Online]. Available: http://proceedings.mlr.press/v37/kandemir15.html

[38] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 294–297, 2015.

[39] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, July 1945. [Online]. Available: http://www.jstor.org/pss/1932409

[40] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145 – 1159, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320396001422

[41] D. Szymkiewicz, "Une conlribution statistique à la géographie floristique," *Acta Societatis Botanicorum Poloniae*, vol. 11, no. 3, pp. 249–265, 1934.

[42] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104

[43] K. Pearson and O. M. F. E. Henrici, "Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1896.0007

[44] C. SPEARMAN, "'footrule' for measuring correlation," *British Journal of Psychology, 1904-1920*, vol. 2, no. 1, pp. 89–108, 1906. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1906.tb00174.x

[45] M. Kendall, "Rank correlation methods," *Harvard Book List*, vol. 99, no. 98, 1955.

[46] A. R. Feinstein and D. V. Cicchetti, "High agreement but low Kappa: I. the problems of two paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 543–549, 1990.

[47] T. Byrt, J. Bishop, and J. B. Carlin, "Bias, prevalence and kappa," *Journal of Clinical Epidemiology*, 1993.

[48] D. V. Cicchetti and A. R. Feinstein, "High agreement but low kappa: II. Resolving the paradoxes," *Journal of Clinical Epidemiology*, 1990.

[49] F. K. Hoehler, "Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity," *Journal of Clinical Epidemiology*, 2000.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[51] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.

[52] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," 2017.

[53] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[54] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," in *BMC Medical Imaging*, 2015.

[55] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[56] M. Hashemi, "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation," *Journal of Big Data*, vol. 6, no. 1, p. 98, Nov 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0263-7

[57] L. Oakden-Rayner, "Exploring large scale public medical image datasets," 2019.

[58] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," *Lecture Notes in Computer Science*, p. 603–611, 2017. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-66179-7_69

[59] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. Metaxas, and X. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1–1, 02 2016.

[60] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, Aug 1995, pp. 278–282 vol.1.

APPENDIX A
PROOF OF STABILITY SCORE LIMITS

A. Adjusted Positive Jaccard

$$S_{APJ}(W_i^{'}, W_i^{''}) = \frac{n_{11}n_{00} - n_{01}n_{10}}{n_{00}n_{11} - n_{01}n_{10} + (n_{10} + n_{01})N},$$

Let A denote for short the numerator in the score, and B be the denominator, such $A := n_{11}n_{00} - n_{01}n_{10}$; and $B := n_{00}n_{11} - n_{01}n_{10} + (n_{10} + n_{01})N$. In addition, the following constraints apply:

1) $n_{00}, n_{11}, n_{01}, n_{10} \geq 0$
2) $n_{00} + n_{11} + n_{01} + n_{10} = N$
3) $\mathrm{N} \geq 1$
4) ensuring denominator is always is non-zero: $n_{00}n_{11} - n_{01}n_{10} + (n_{10} + n_{01})N \neq 0$

1) Minimum Value: So the strategy for finding minimum value of the score is outlined with two simultaneous targets. Firstly, we should note that the smallest number for S_{APJ} is achieved if the score is negative, which can be achieved only via the numerator. Secondly, a minimum S_{APJ} is accomplished if the numerator's absolute value is maximized, while the denominator is minimized.

$$\textbf{Sub-goal I:} A < 0 \tag{24}$$

$$\textbf{Sub-goal II:} \min B \wedge \max |A| \tag{25}$$

To achieve $A < 0$ in 24, it is enough that $n_{00} = 0 \vee n_{11} = 0$. In addition, $n_{00} = 0 \vee n_{11} = 0$ contributes to Equation 25. As a result

$$S_{APJ} = \frac{-n_{01}n_{10}}{-n_{01}n_{10} + (n_{01} + n_{10})N} \tag{26}$$

Despite the score dependency on the exact distribution of all instances, which we do not consider in depth at this point, we see that the stability score attains a minimum value at low N, and it converges **around** 0 with large N.

If we consider $|S_{APJ}|$ from (26), the problem can be seen as solving maximization problem. We maximize $n_{01}n_{10}$, while minimizing $-n_{01}n_{10} + (n_{01} + n_{10})N$. Acknowledge that minimizing N also leads to lower denominator in (26), so $n_{00} = n_{11} = 0$, leading to $\min N = (n_{01} + n_{10})$. So next, to find minimum value assume $S_{APJ} \leq -1$, hence:

$$-1 \geq \frac{-n_{01}n_{10}}{-n_{01}n_{10} + (n_{01} + n_{10})(n_{01} + n_{10})} \tag{27}$$

Rewriting the expression we derive to:

$$(n_{01} + n_{10})^2 \leq 2n_{01}n_{10}$$

It quickly becomes apparent that the left side of inequality has quadratic growth in contrast to the linear growth on the right side, and the inequality holds only if $n_{01} = n_{10} = 0$. However, for non-trivial case where $N > 0$, the inequality leads to contradiction. That concludes that it must be that $S_{APJ} > -1$. Another observation is that the numerator's absolute value cannot grow larger that the denominator as the absolute score value decreases with the increase of $(n_{01}+n_{10})$. So keeping $\min(n_{01} + n_{10}) = 2$ and $\max(n_{01}n_{10}) = 1$, then $\arg\max_{n_{01}n_{10}} |S_{APJ}| = \{1, 1\}$. So the minimum value of $S_{APJ} = -0.333$.
Here, it is important to note that from 26 it is apparent that for large N, $lim_{N\to\infty} S_{APJ} = 0$.

2) Maximum Value: The maximum value of the score can be outlined with two simultaneous targets. One is finding a maximal positive value of the numerator, while minimizing the value of the denominator.

$$\textbf{Sub-goal I:} A > 0 \wedge \max A \tag{28}$$

$$\textbf{Sub-goal II:} \min B \tag{29}$$

Equation 28 can be achieved by considering $n_{01}n_{10} = 0$. Considering 29 with respect to these variables, B is minimized if we minimize $n_{01}+n_{10}$ such that $n_{01}+n_{10} = 0$. That results in $n_{01} = n_{10} = 0$ and $S_C = \frac{n_{11}n_{00}}{n_{11}n_{00}} = 1$

3) Monotonicity: Finding the positive and negative intervals of the first derivative of S_{APJ} in respect of n_{11} can show us the values for which the function is monotonically increasing. Firstly we use the quotient rule for differentiation:

$$\frac{\partial}{\partial x}\frac{f(x)}{g(x)} = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}$$

to find the first derivative of S_{APJ}.

Let $u = n_{11}n_{00} - n_{01}n_{10}$ and $v = n_{00}n_{11} - n_{01}n_{10} + (n_{10} + n_{01})(n_{00} + n_{11} + n_{10} + n_{01})$. Hence:

$$\frac{\partial}{\partial n_{11}} S_{APJ} = \frac{v\frac{\partial u}{\partial n_{11}}(u) - u\frac{\partial v}{\partial n_{11}}(v)}{v^2}$$

where

$$v\frac{\partial u}{\partial n_{11}}(u) = v * n_{00}$$

and

$$u\frac{\partial v}{\partial n_{11}}(v) = u * (n_{00} + n_{01} + n_{10})$$

After simplifying the arithmetic steps we derive to:

$$v\frac{\partial u}{\partial n_{11}}(u) - u\frac{\partial v}{\partial n_{11}}(v) = (n_{01} + n_{10})(n_{01} + n_{00})(n_{10} + n_{00})$$

and

$$v^2 = (n_{00}n_{11} - n_{01}n_{10} + (n_{10} + n_{01})(n_{11} + n_{10} + n_{01} + n_{00}))^2$$

Despite the multiple elements involved we observe that the derivative is always positive for $n_{11} > 0$, such $\frac{\partial}{\partial n_{11}} S_{APJ}(n_{11}) > 0$. From this we can conclude that the function is monotonically increasing with respect to the positive overlap observed.

B. Adjusted Jaccard

$$S_{AJ}(W_i', W_i'') = \frac{2n_{11}n_{00} - 2n_{01}n_{10}}{2n_{11}n_{00} - 2n_{01}n_{10} + (n_{01} + n_{10})N} \quad (30)$$

Let A denote for short the numerator in the score, and B be the denominator, such $A := 2n_{11}n_{00} - 2n_{01}n_{10}$; and $B := 2n_{00}n_{11} - 2n_{01}n_{10} + (n_{10} + n_{01})N$. In addition, the following constraints apply:

1) $n_{00}, n_{11}, n_{01}, n_{10} \geq 0$
2) $n_{00} + n_{11} + n_{01} + n_{10} = N$
3) N ≥ 1
4) ensuring denominator is always is non-zero: $2n_{00}n_{11} - 2n_{01}n_{10} + (n_{10} + n_{01})N \neq 0$

1) Minimum Value: In a similar manner to adjusted positive Jaccard, we can show that negative and minimum score is achieved when $n_{11} = n_{00} = 0$ such that:

$$S_{AJ}(W_i', W_i'') = \frac{-2n_{01}n_{10}}{-2n_{01}n_{10} + (n_{01} + n_{10})(n_{01} + n_{10})} \quad (31)$$

Next, assume that $S_{AJ} \leq -1$ leading to:

$$-2n_{01}n_{10} + (n_{01} + n_{10})(n_{01} + n_{10}) \leq 2n_{01}n_{10}$$

and simplifying it to:

$$(n_{01} + n_{10})^2 \leq 4n_{01}n_{10} \quad (32)$$

From 32 it is derived that:

$$n_{01}^2 + n_{10}^2 \leq 2n_{01}n_{10} \quad (33)$$

From the simplified version in 33, we see that $S_{AJ} \leq -1$ only for $n_{01} = n_{01} = 1$. For any larger value of n_{01} or n_{10} the inequality will no longer hold and as such the score score decreases. In addition, smaller value of n_{01} or n_{10} decreases the stability score. So the minimum value of S_{AJ} = -1 is found for $n_{11} = n_{00} = 0$ and $n_{01} = n_{01} = 1$.

2) Maximum Value: The strategy is again similar to what we see in the adjusted positive Jaccard. The lower bound of the score can be outlined with two simultaneous targets. One is finding a maximal positive value of the numerator, while minimizing the value of the denominator.
Maximizing the denominator by setting $n_{01}n_{10} = 0$. Furthermore, the denominator is minimized with respect to these variables if we minimize $n_{01} + n_{10}$. As a result, we set: $n_{01} + n_{10} = 0$.
So for $n_{01} = n_{10} = 0$ and $S_{AJ} = \frac{2n_{11}n_{00}}{2n_{11}n_{00}} = 1$.

3) Monotonicity:
Let $u = 2n_{11}n_{00} - 2n_{01}n_{10}$ and
$v = 2n_{00}n_{11} - 2n_{01}n_{10} + (n_{10} + n_{01})(n_{00} + n_{11} + n_{10} + n_{01})$.

Finding the positive and negative intervals of the first derivative of S_{AJ} in respect of n_{11} and n_{00} show us the values for which the function is monotonically increasing. We show separately the monotonicity with respect of n_{00} and n_{11}. Firstly, we prove monotonicity regarding n_{00}. We apply the quotient rule from Equation A-A3:

$$\frac{\partial}{\partial n_{00}} S_{AJ} = \frac{v\frac{\partial u}{\partial n_{00}}(u) - u\frac{\partial v}{\partial n_{00}}(v)}{v^2}$$

where

$$v\frac{\partial u}{\partial n_{00}}(u) = 2v * n_{11}$$

and

$$u\frac{\partial v}{\partial n_{00}}(v) = u * (2n_{11} + n_{01} + n_{10})$$

After simplifying the arithmetic steps, we derive to:

$$v\frac{\partial u}{\partial n_{00}}(u) - u\frac{\partial v}{\partial n_{00}}(v) =$$

$$2(n_{01}^2 n_{10} + n_{01}^2 n_{11} + n_{01}n_{10}^2 + 2n_{01}n_{10}n_{11} + n_{01}n_{11}^2 + n_{10}^2 n_{11} + n_{10}n_{11}^2)$$

and

$$v^2 = (2n_{00}n_{11} - 2n_{01}n_{10} + (n_{10} + n_{01})(n_{00} + n_{11} + n_{10} + n_{01}))^2$$

Despite the multiple elements involved we observe that the derivative is always positive for $n_{00} > 0$, such $\frac{\partial}{\partial n_{00}} S_{AJ}(n_{00}) > 0$. From this we can conclude that the function is monotonically increasing with respect to the positive overlap observed.

Finding the derivative in a similar manner with respect to n_{11}, we find that:

$$v\frac{\partial u}{\partial n_{11}}(u) = 2v * n_{00}$$

and

$$u\frac{\partial v}{\partial n_{11}}(v) = u * (2n_{00} + n_{01} + n_{10})$$

and

$$v\frac{\partial u}{\partial n_{00}}(u) - u\frac{\partial v}{\partial n_{00}}(v) =$$

$$2(n_{01}n_{00}^2 + n_{10}n_{00}^2 + n_{01}^2 n_{00} + 2n_{01}n_{10}n_{00} + n_{01}^2 n_{10} + n_{01}n_{10}^2 + n_{10}^2 n_{00})$$

Similarly, we observe that the derivative is always positive for $n_{11} > 0$, such $\frac{\partial}{\partial n_{11}} S_{AJ}(n_{11}) > 0$. From this and the monotonicity in respect to n_{00}, we can conclude that the function is monotonically increasing with respect to the positive and negative overlap observed.

APPENDIX B
TRAINING SETTINGS

Exact settings during for the results derived are described in Table VIII.

Dataset	Class	CV fold used for stability	Operator	Settings
Xray	cardiomegaly	fold 2	`NOR`	Epochs: 6, Learning rate: 5×10^{-5} Activity L2 regularization: 10^{-4}
Xray	cardiomegaly	fold 2	`Mean`	Epochs: 5, Learning rate: 6×10^{-5} Activity L2 regularization: 7.5×10^{-4}
Xray	cardiomegaly	fold 2	`LSE, r=0.1`	Epochs: 4, Learning rate: 1×10^{-5} Activity L2 regularization: 10^{-4}
Xray	cardiomegaly	fold 2	`LSE, r=1`	Epochs: 5, Learning rate: 2×10^{-5} Activity L2 regularization: 10^{-4}
Xray	cardiomegaly	fold 2	`Max`	Epochs: 9, Learning rate: 3×10^{-5} Activity L2 regularization: 10^{-4}
MURA	shoulder	fold 2	`NOR`	Epochs: 7, Learning rate: 7×10^{-6} Activity L2 regularization: 10^{-4}
MURA	shoulder	fold 2	`Mean`	Epochs: 10, Learning rate: 1×10^{-5} Activity L2 regularization: 10^{-4}
MURA	shoulder	fold 2	`LSE, r=0.1`	Epochs: 8, Learning rate: 1×10^{-5} Activity L2 regularization: 10^{-4}
MURA	shoulder	fold 2	`LSE, r=1`	Epochs: 7, Learning rate: 1×10^{-5} Activity L2 regularization: 10^{-4}
MURA	shoulder	fold 2	`Max`	Epochs: 16, Learning rate: 1×10^{-5} Activity L2 regularization: 10^{-4}
Pascal VOC	cars	fold 2	`NOR`	Epochs: 20 Learning rate: 1×10^{-6}
Pascal VOC	cars	fold 2	`Mean`	Epochs: 15 Learning rate: 1×10^{-6}
Pascal VOC	cars	fold 2	`LSE, r=0.1`	Epochs: 14 Learning rate: 1×10^{-6}
Pascal VOC	cars	fold 2	`LSE, r=1`	Epochs: 15 Learning rate: 1×10^{-6}
Pascal VOC	cars	fold 2	`Max`	Epochs: 24 Learning rate: 1×10^{-5}
Xray	Effusion	-	`NOR`	Epochs: 4 Learning rate: 5×10^{-5} Activity L2 regularization: 5×10^{-4}

TABLE VIII: Training settings for each dataset

APPENDIX C
SCORE VALUE

The correction for chance in adjusted positive Jaccard makes it counter-intuitive. So in this section we examine the behaviour of the adjusted positive Jaccard. Figure 9 shows the results of several simulations, where at the start $n_{00} = n_{01} = n_{10} = n_{11} = 50$. The total bags are 200 throughout the whole experiment, and it is the ratio of the sub-groups that alters.

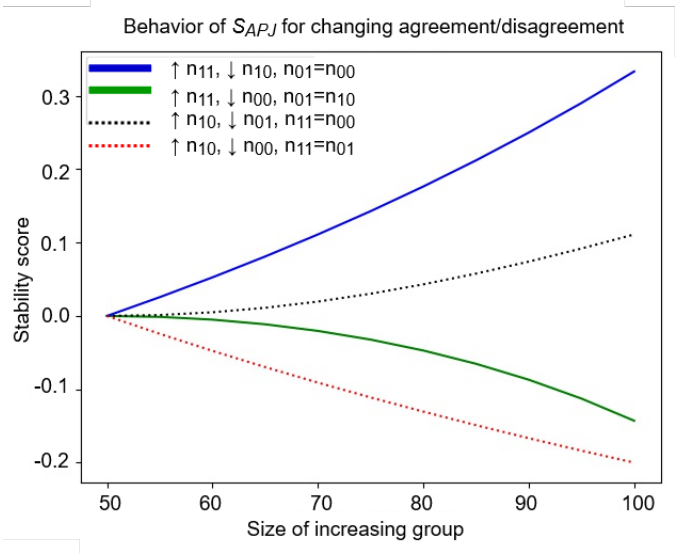


Fig. 9: Adjusted Positive Jaccard with varying sizes of subgroups. *Blue line:* Positive agreement increases, while disagreement decreases. *Green line:* Positive agreement increases, while negative agreement decreases. *Black dotted line:* Disagreement demography changes, agreement is fixed. *Red dotted line:* Disagreement increases, negative agreement decreases.

In this initial settings, the stability score is 0, although there is 50% positive agreement from total positive predictions of each classifier. Consequently, the solid lines follow the stability score with the increase of the positive agreement n_{11}. The increase of n_{11} is proportionate to the decrease of n_{00} on the green line, and proportionate to the decrease of n_{10} on the blue line. While both line depict same rate of positive agreement increase, the behaviour is rather contrasting. What is more, the behaviour on the green line is counter-intuitive since the stability score decreases with the increase of positive agreement between the sets.

The dotted lines, on the other hand, depict simulations where the positive agreement is fixed for the whole experiment. These lines show the change of the stability score when one of the classifier starts predicting more positive instances (n_{10} or n_{01} increases). On the black dotted line, the positive instances of a classifier increases inversely proportional to the positive instances of the other classifier (e.g. n_{10} increasing with respect to n_{01}), changing the positive prediction division between classifiers. On the red line, n_{10} increases inversely to the negative agreement n_{00}, decreasing the total negative agreement. The behavioral difference in these intuitively similar situations shows contrasting result in terms of the stability score.

Finally, comparing the black dotted line and the green line in the last step of the simulations (exact distribution is in Table IX) provides another aspect of the behaviour. A smaller positive agreement (the black dotted line) has higher stability score than much larger ratio of agreement (the green line).

		Classifier 2 Predictions			
		Green line		Black dotted line	
Classifier 1 Predictions		**0**	**1**	**0**	**1**
	0	0	50	50	0
	1	50	*100*	100	*50*

TABLE IX: Classification population in the last step of the simulation