PLOS One

RESEARCH ARTICLE

Beyond accuracy: Quantifying the reliability of multiple instance learning for whole slide image classification

Hassan Keshvarikhojasteh[1]*, **Marc Aubreville**[2], **Christof A. Bertram**[3], **Josien P. W. Pluim**[1], **Mitko Veta**[1]

1 Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, **2** Flensburg Artificial Intelligence Research Group (FLAIR), Flensburg University of Applied Sciences, Flensburg, Germany, **3** University of Veterinary Medicine Vienna, Vienna, Austria

* h.keshvarikhojasteh@tue.nl

Check for updates

OPEN ACCESS

Citation: Keshvarikhojasteh H, Aubreville M, Bertram CA, Pluim JP, Veta M (2025) Beyond accuracy: Quantifying the reliability of multiple instance learning for whole slide image classification. PLoS One 20(12): e0337261. https://doi.org/10.1371/journal.pone.0337261

Editor: Elnaz Pashaei, Indiana University School of Medicine, UNITED STATES OF AMERICA

Received: May 23, 2025

Accepted: November 5, 2025

Published: December 5, 2025

Copyright: © 2025 Keshvarikhojasteh et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The data underlying the results presented in the study are publicly available. The datasets

Abstract

Machine learning models have become integral to many fields, but their reliability, defined as producing dependable, trustworthy, and domain-consistent predictions, remains a critical concern. Multiple Instance Learning (MIL) models designed for Whole Slide Image (WSI) classification in computational pathology are rarely evaluated in terms of reliability, leaving a key gap in understanding their suitability for high-stakes applications like clinical decision-making. In this paper, we address this gap by introducing three quantitative metrics for reliability assessment and applying them to several widely used MIL architectures across three region-wise annotated pathology datasets. Our findings indicate that the mean pooling instance (MEAN-POOL-INS) model demonstrates superior reliability compared to other networks, despite its simple architectural design and computational efficiency. These findings underscore the need of reliability evaluation alongside predictive performance in MIL models and establish MEAN-POOL-INS as a strong, trustworthy baseline for future research.

Introduction

Machine learning (ML) has become a cornerstone of modern computational pathology by enabling automated analysis of large and complex datasets generated in clinical practice [1,2]. Deep learning, in particular, has shown remarkable potential for tasks such as tumor detection, tissue subtyping, and prognosis prediction [3,4].

Multiple Instance Learning (MIL) is a widely used method for weakly-supervised classification tasks in computational pathology, primarily due to the difficulty and labor-intensive nature of obtaining pixel-wise annotations from histopathology data [5, 6]. Compared to simpler approaches such as patch-level classifiers with propagated slide labels or heuristic aggregation strategies, MIL offers a principled means of reducing label noise and focusing on diagnostically relevant regions. In the MIL framework, groups of instances, or “bags,” are created, with the goal of predicting

used include TCGA via the GDC Data Portal (https://portal.gdc.cancer.gov/), CAMELYON16 via the Grand Challenge site (https://camelyon16.grand-challenge.org/Data/), and CATCH via The Cancer Imaging Archive (https://www.cancerimagingarchive.net/collection/catch/). The code used to reproduce the analysis is available at https://github.com/tueimage/MIL-Reliability.

Funding: This research was supported by the IMI BigPicture project (IMI945358), part of the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 945358. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have no competing interests to declare. This does not alter our adherence to PLOS One policies on sharing data and materials.

the label for each bag. The key assumption in MIL is that positive bags contain at least one key instance, while negative bags do not. Recent advancements in deep learning have spurred the development of novel MIL models tailored to computational pathology, enhancing performance through key instance detection [7], feature space refinement [8], and overfitting mitigation [9,10].

Reliability refers to the ability of a system or model to perform as expected under standard and well-defined conditions. The reliability of MIL models is essential for gaining the trust of physicians, ensuring successful deployment in clinical settings, and adhering to established healthcare standards. Reliable models focus on biologically pertinent features, grounded in established scientific knowledge, when making predictions. In the context of computational pathology, an unreliable MIL model might erroneously focus on artifacts or irrelevant features, leading to false positives or negatives in critical tasks such as tumor detection or subtyping. For example, a model could achieve high slide-level accuracy by correctly predicting tumor presence in most cases, yet consistently rely on irrelevant regions such as staining or scanner-induced artifacts, making it accurate but not reliable. Conversely, a model could consistently focus on biologically meaningful regions (reliable) but exhibit slightly lower overall accuracy due to ambiguous or noisy labels. Such errors could result in misdiagnoses, delays in treatment, or unnecessary interventions, underscoring the need for robust reliability evaluations. Achieving reliability in MIL models for computational pathology is particularly challenging due to the inherent variability in histopathology data, including differences in staining, scanning resolutions, and patient-specific biological heterogeneity [11]. Furthermore, the weakly-supervised nature of MIL exacerbates these challenges by relying on coarse slide-level labels without precise spatial annotations. Addressing these limitations requires not only model improvements but also careful evaluation frameworks that ensure predictions align with domain knowledge and clinical relevance. By fostering reliable MIL models, computational pathology can advance toward broader acceptance and integration into clinical workflows.

However, machine learning models are rarely evaluated in terms of reliability; instead, performance metrics such as accuracy and F1 score are typically reported [12]. Yet, reliable models are essential for real-world deployment and tend to exhibit better generalization. Therefore, gaining insights into models' reliability is critical. For instance, Pendyala et al. [13] assessed the reliability of various models using a mental health dataset and discovered that these models often focused on less relevant features, despite achieving high performance, thereby highlighting the limitations of traditional metrics. Furthermore, Araujo et al. [14] propose an alternative validation method for models and support the assertion that performance metrics are not indicative of the reliability of machine learning networks. In the domain of medical imaging, the trustworthiness of saliency maps has been quantitatively assessed on two public radiology datasets using the area under the precision-recall curve (AUPRC) and the structural similarity index measure (SSIM) as evaluation metrics [15]. The study revealed that none of the saliency maps satisfied all the tested criteria, highlighting limitations in their reliability.

Interpretability and reliability in the context of machine learning (ML) are related but distinct concepts that address different aspects of model performance and trustworthiness. Interpretability means how easily a person can understand why a model made a certain decision or how it comes to its results [16]. We aim to understand and explain the way models work in interpretability, providing transparency. To address questions concerning the decision-making processes and the underlying mechanics, numerous interpretability methods have been developed [17]. Interpretability and explainability methods have proven useful in identifying both data and model deficiencies [18]. For MIL models, attention scores assigned to each patch are frequently used to generate heatmaps, which are key tools for assessing interpretability. However, interpretability in MIL models has traditionally been evaluated qualitatively by presenting specific slides and their corresponding heatmaps. The Camelyon16 dataset [19], which includes pixel-wise ground truth annotations, is the most common resource for such evaluations. While qualitative assessments provide valuable insights, they are not feasible across large test sets due to computational and time constraints. Moreover, analyzing heatmaps requires specialized histopathology knowledge, which may not be readily available to machine learning researchers. Finally, a more robust and comprehensive assessment should involve evaluation across multiple datasets.

In contrast to interpretability, reliability is concerned with ensuring the model consistently produces correct and trustworthy predictions in a wide range of circumstances, independent of transparency of underlying process by which predictions are made. Models focusing on related regions of interest (ROIs) are more reliable, as they process the most relevant parts of the data. However, more interpretable models are not necessarily more reliable. For instance, a model might offer more interpretability through class-specific information or excitatory-inhibitory heatmaps [18], yet fail to concentrate on relevant ROIs. Additionally, while interpretable models may provide explanations for predictions across entire classes, this does not necessarily align with reliability. For example, in binary tumor versus normal classification, since normal samples lack clear ROIs, predictions for these cases are not considered in reliability assessments.

In this paper, we address the above-mentioned challenges by quantitatively evaluating the reliability of MIL models for both binary tumor versus normal classification and multi-class tissue subtyping problems using three metrics. Additionally, we incorporate a recently proposed additive interpretability method [18] to investigate whether improved interpretability translates into increased reliability, providing new insights into the complex relationship between interpretability and reliability. Unlike previous studies that rely mainly on single-dataset evaluations, we extend our experiments beyond the Camelyon16 dataset by including two additional datasets: CATCH [20] and TCGA BRCA. By introducing quantitative reliability metrics, benchmarking multiple MIL architectures across three datasets, and examining the impact of additive interpretability, this study provides a comprehensive, multi-dataset evaluation of MIL model reliability in computational pathology. Our study aims to advance the development of more reliable MIL models for computational pathology.
The main contributions of this paper are as follows:

- Introduction of three quantitative metrics to evaluate the reliability of MIL models in computational pathology.
- Systematic benchmarking of several widely-used MIL architectures across three region-wise annotated pathology datasets: Camelyon16, CATCH, and TCGA BRCA.
- Investigation of the impact of a recently proposed additive interpretability method on MIL model reliability, providing insights into the relationship between interpretability and reliability.
- Demonstration that comprehensive reliability evaluation complements traditional accuracy metrics and is essential for trustworthy deployment in high-stakes clinical settings.

Materials and methods

In this section, we describe the problem formulation, the evaluation metrics for predictive performance and reliability, the datasets used in our experiments, and the details of our implementation.

Problem formulation

For our analysis, we refer to the input Whole Slide Image (WSI) as I and predict its slide-level label $\hat{Y}$. Following established methodologies, we first tessellate the slide into N small patches and then extract features from these patches using a pre-trained model. In the final step, the slide label is predicted using an aggregation module, which is typically attention-based, providing attention scores for each patch. For clarification, the proposed framework is illustrated in Fig 1.

Metrics

Reliability in this study is defined as the consistent focus of MIL models on diagnostically relevant ROIs within WSIs, a prerequisite for trustworthy and clinically useful predictions. To quantify this alignment, we selected three complementary metrics that together capture different aspects of spatial concordance between predicted patch scores and ground truth annotations:

- **Mutual Information (*MI*):** Measures the statistical dependence between predicted scores and ROIs, quantifying how much information the model's output shares with the true ROI distribution [21].
- **Spearman's Correlation** (*Spearman's*)**:** Captures the monotonic relationship between predicted patch importance and true ROI presence, allowing for non-linear associations and reflecting relative ranking fidelity [22].

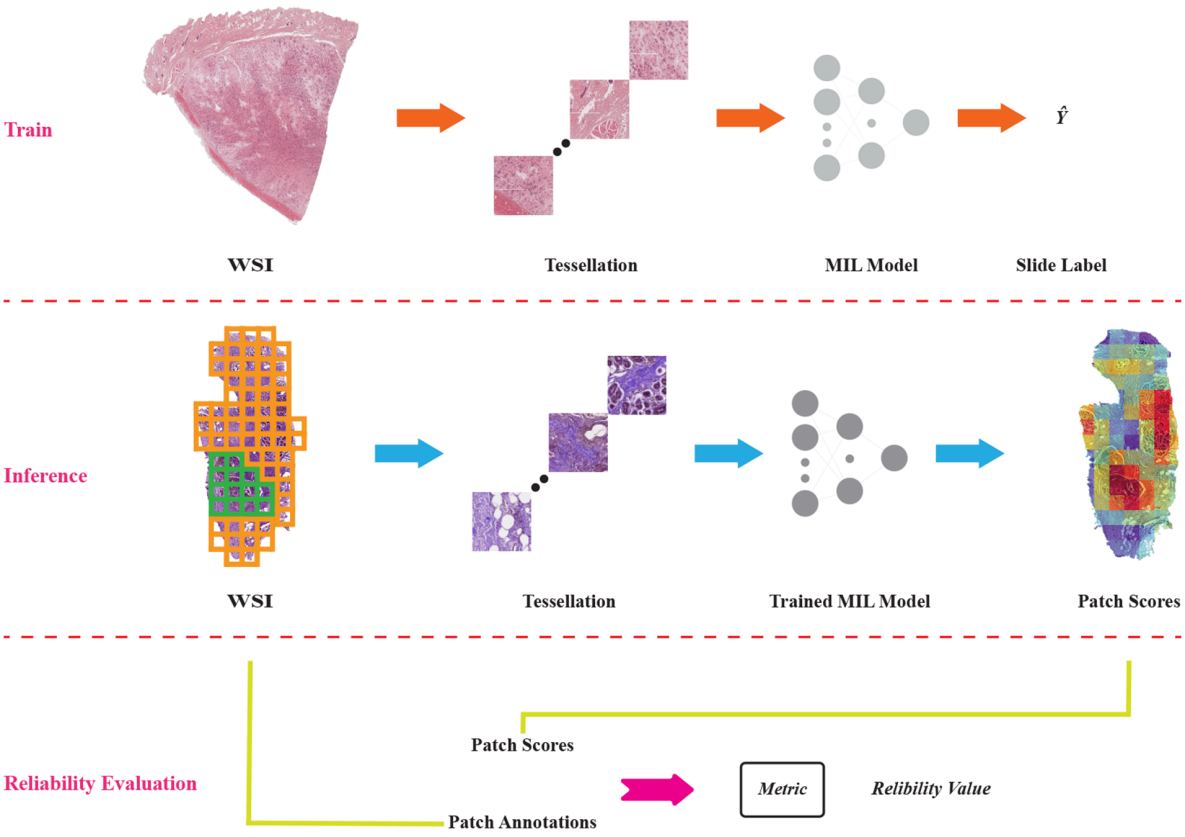


Fig 1. The overall framework for evaluating the reliability of MIL models follows a three-step process. First, a MIL model is trained on a weakly-supervised task for predicting slide-level labels. Next, the trained model is applied to predict scores for individual image patches. Finally, the reliability value is computed based on the predicted patch scores and their corresponding annotations, where tumor patches are highlighted in green and normal patches in orange in the annotation visualization.

https://doi.org/10.1371/journal.pone.0337261.g001

- **Area Under the Precision-Recall Curve (*AUPRC*):** Evaluates the model's ability to localize ROIs under conditions of class imbalance by measuring precision and recall trade-offs [23].

Dataset

Three public WSI datasets were used in the experiments. The Breast Invasive Carcinoma dataset (**TCGA BRCA**) from The Cancer Genome Atlas (TCGA) project includes 1,038 H&E-stained slides representing two types of breast cancer: Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC), with slide-level labels. The **CAMELYON16** dataset contains 399 WSIs with breast cancer lymph node metastasis annotations, divided into 270 slides for training and validation, and 129 for testing. Finally, the **CATCH** dataset provides 350 meticulously annotated slides from seven different canine cutaneous tumors.

Unlike CAMELYON16 and CATCH, the TCGA BRCA dataset lacks ground truth region-level annotations. To address this, tumor regions in the test slides were manually annotated with the assistance of Dr. Christof A. Bertram, a pathologist [24].

Implementation details

Throughout the experiments, we employ the Adam optimizer with a learning rate of 1×10^{-4} and set the maximum number of epochs to 50. The weight decay is tuned based on the validation loss. All experiments are repeated with five different random seeds, and we report the mean *AUC* and *F1* scores as classification metrics, along with the mean *MI*, r_s, and *AUPRC* values as reliability metrics. The *AUC* measures the model's overall discriminative ability in classifying positive and negative cases across all thresholds, while the *AUPRC* emphasizes the model's reliability in identifying true positive regions under conditions of class imbalance. Using *AUPRC* as a reliability metric thus complements the classification-based *AUC*, providing a more comprehensive evaluation of MIL model performance. Additionally, we evaluate *model size* and *Floating-Point Operations Per Second (FLOPS)* as computational metrics. For dataset preprocessing, we follow the procedures outlined in [8], which involve extracting features from non-overlapping 256×256 patches at $20\times$ magnification using a ResNet50 pretrained on ImageNet. These extracted features are fed into the model with an input dimensionality of $input_dim = 1024$, which is subsequently reduced to $D = 512$.

We evaluate several state-of-the-art models, including ABMIL [7], CLAM [8], DTFD [9], ACMIL [10], MADMIL [25], mean pooling (MEAN-POOL), max pooling (MAX-POOL), max pooling instance (MAX-POOL-INS), and mean pooling instance (MEAN-POOL-INS). For the MEAN-POOL and MAX-POOL models, we calculate the slide representation after feature compression using either the mean or max operation, and then predict the slide label. It should be noted that reliability metrics are not reported for the MEAN-POOL model, as its mean aggregation does not yield patch-level scores or attention weights. In contrast, for the MAX-POOL-INS and MEAN-POOL-INS models, we first predict patch probabilities, which are then aggregated to generate slide-level probabilities. Patch probabilities are also used to compute reliability metrics for MEAN-POOL-INS. In the case of MAX-POOL-INS, we use both the selected patch index, similar to MAX-POOL, and the patch probabilities.

For multi-head models such as MADMIL and ACMIL, average attention scores or patch probabilities are employed to calculate reliability metrics. We use a shorthand notation to indicate the number of heads, e.g., MADMIL/4 refers to the MADMIL model with four heads. To ensure a thorough analysis, we also investigate additive versions of these models, in accordance with the approach proposed by [18] , which reformulates the final predictor as an additive function of individual instances. For additive models, either attention scores or patch probabilities are used to assess reliability.

To compute the reliability metrics, it is essential to obtain both patch labels and the corresponding attention scores or probabilities. Following tessellation, patch labels are assigned based on the region-wise ground truth annotations. Specifically, for each patch, if it falls within a annotated region , it is assigned a label of 1, indicating the presence of the

target feature. Conversely, if the patch does not intersect with the annotated region, it is assigned a label of 0, denoting the absence of the target feature.

Results

We present the predictive performance and reliability results for CAMELYON16, CATCH, and TCGA BRCA datasets. This section highlights how well each model identifies regions of interest and compares their overall effectiveness.

CAMELYON16. The results for this dataset are presented in Tables 1 and 2. In Table 1, the ACMIL/2 model demonstrates moderate classification performance but high reliability. Conversely, DTFD achieves strong classification performance with lower reliability. Interestingly, the MAX-POOL model, while computationally efficient and yielding reasonable classification results, exhibits the lowest reliability. This finding reinforces our argument that model selection should not be based solely on classification metrics. To illustrate this point, we present a tumor slide that was correctly classified by MAX-POOL, along with its heatmap in Fig 2. The heatmap reveals that the model concentrates on normal patches of lymph node tissue rather than metastatic tumor patches, indicating the unreliability of its predictions.

The results in Table 2 show that additive models do not exhibit significant improvements in classification performance, but they do show slight enhancements in reliability metrics. Notably, the simple MEAN-POOL-INS model achieves the highest reliability values, although having the lowest classification performance. Fig 3 illustrates a slide and its corresponding heatmap, highlighting the model's focus on tumor areas.

Table 1. Average reliability, classification, and computational metrics (± standard deviation) over five repetitions for the CAMELYON16 dataset.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL	0.12 ± 0.00	0.29 ± 0.00	0.59 ± 0.01	0.85 ± 0.01	0.84 ± 0.01	94 M	789 K
CLAM	0.12 ± 0.01	0.29 ± 0.02	0.59 ± 0.02	0.86 ± 0.01	**0.85 ± 0.01**	94 M	791 K
DTFD	0.13 ± 0.01	0.29 ± 0.01	0.61 ± 0.01	**0.89 ± 0.01**	0.85 ± 0.02	126 M	1053 K
ACMIL/2	**0.16 ± 0.02**	**0.34 ± 0.02**	0.59 ± 0.05	0.82 ± 0.03	0.81 ± 0.01	94 M	791 K
MADMIL/2	0.13 ± 0.00	0.31 ± 0.00	**0.61 ± 0.00**	0.86 ± 0.01	0.85 ± 0.01	79 M	658 K
MAX-POOL	0.00 ± 0.00	-0.00 ± 0.00	0.23 ± 0.01	0.80 ± 0.01	0.79 ± 0.01	**63 M**	**526 K**
MEAN-POOL[a]	–	–	–	0.58 ± 0.02	0.59 ± 0.02	63 M	526 K

FLOPs are calculated using 120 instances per bag.

[a]Reliability metrics are not computed for this model, as it does not provide patch scores.

https://doi.org/10.1371/journal.pone.0337261.t001

Table 2. Average reliability, classification, and computational metrics (± standard deviation) over five repetitions for the CAMELYON16 dataset using additive models.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL-ADD-ATT	0.13 ± 0.00	0.32 ± 0.01	0.62 ± 0.01	0.86 ± 0.01	0.83 ± 0.02	95 M	789 K
ABMIL-ADD-PATCH	0.04 ± 0.00	0.24 ± 0.01	0.48 ± 0.01				
CLAM-ADD-ATT	0.12 ± 0.01	0.29 ± 0.01	0.60 ± 0.02	0.87 ± 0.01	0.83 ± 0.02	95 M	791 K
CLAM-ADD-PATCH	0.03 ± 0.00	0.24 ± 0.05	0.48 ± 0.02				
DTFD-PATCH[a]	0.04 ± 0.00	0.29 ± 0.02	0.53 ± 0.01	**0.89 ± 0.01**	**0.85 ± 0.02**	126 M	1053 K
ACMIL/2-ADD-ATT	0.17 ± 0.01	0.35 ± 0.01	0.62 ± 0.02	0.79 ± 0.02	0.81 ± 0.01	95 M	791 K
ACMIL/2-ADD-PATCH	0.03 ± 0.00	-0.03 ± 0.01	0.29 ± 0.01				
MAX-POOL-INS	0.00 ± 0.00	0.05 ± 0.03	0.20 ± 0.02	0.77 ± 0.14	0.76 ± 0.13	63 M	526 K
MAX-POOL-INS-PATCH	0.15 ± 0.06	0.29 ± 0.15	0.55 ± 0.20				
MEAN-POOL-INS	0.31 ± 0.04	0.53 ± 0.07	0.78 ± 0.08	0.58 ± 0.02	0.59 ± 0.02	**63 M**	**526 K**

[a]We use the provided patch probabilities based on GRAD-CAM.

https://doi.org/10.1371/journal.pone.0337261.t002

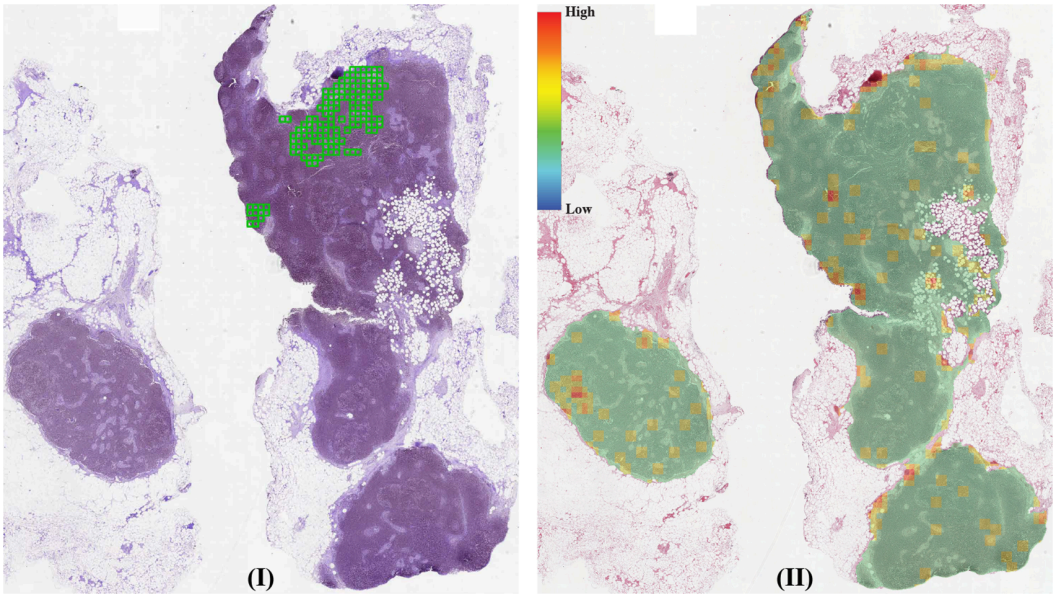


Fig 2. (I) The test-30 slide with ground truth annotations (green) overlaid on the tissue section. (II) Corresponding heatmap generated by MAX-POOL, showing predicted patch scores distribution from low (blue) to high (red). The annotation and heatmap are spatially aligned for comparison.

https://doi.org/10.1371/journal.pone.0337261.g002

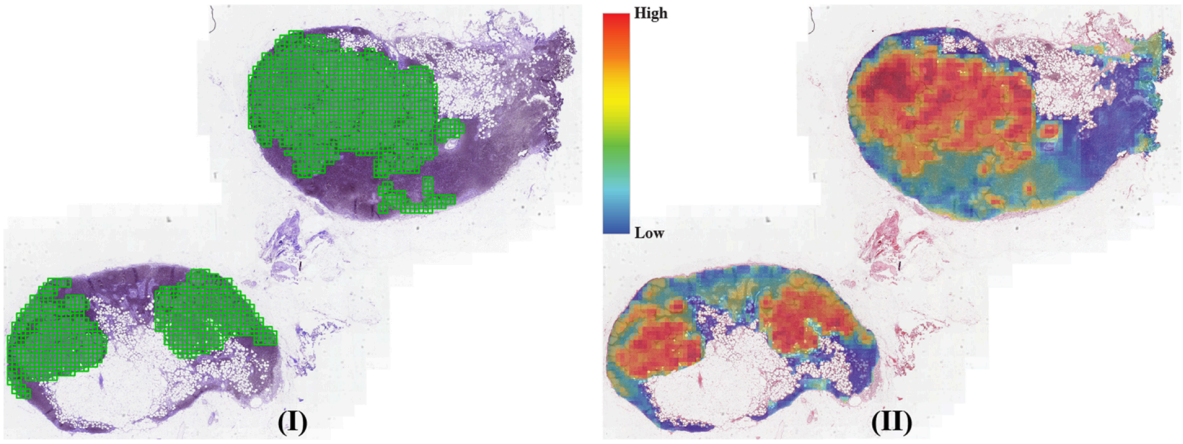


Fig 3. (I) The test-40 slide with ground truth annotations (green) overlaid on the tissue section. (II) Corresponding heatmap generated by MEAN-POOL-INS, showing predicted patch scores distribution from low (blue) to high (red). The annotation and heatmap are spatially aligned for comparison.

https://doi.org/10.1371/journal.pone.0337261.g003

CATCH. Tables 3 and 4 present the results for CATCH dataset. Notably, the MAX-POOL model achieves the highest classification metrics but records the lowest reliability scores. Selecting the MAX-POOL model without considering its reliability could lead to unpredictable outcomes. As shown in Fig 4, the model primarily focuses on non-tumorous regions, further highlighting its lack of reliability.

Table 3. Average reliability, classification, and computational metrics (± standard deviation) over five repetitions for CATCH.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL	0.30 ± 0.02	0.53 ± 0.01	0.91 ± 0.01	0.99 ± 0.00	0.90 ± 0.02	94 M	791 K
CLAM	**0.32 ± 0.01**	0.54 ± 0.00	0.91 ± 0.00	0.99 ± 0.00	0.88 ± 0.01	94 M	798 K
DTFD	0.20 ± 0.06	**0.56 ± 0.01**	0.91 ± 0.01	1 ± 0.00	0.93 ± 0.01	126 M	1842 K
ACMIL/4	0.28 ± 0.01	0.55 ± 0.01	**0.92 ± 0.01**	1 ± 0.00	0.93 ± 0.01	95 M	807 K
MADMIL/2	0.30 ± 0.01	0.55 ± 0.00	0.92 ± 0.00	0.99 ± 0.00	0.90 ± 0.02	79 M	660 K
MAX-POOL	0.00 ± 0.00	-0.05 ± 0.00	0.65 ± 0.00	**1 ± 0.00**	**0.94 ± 0.02**	**63 M**	**528 K**
MEAN-POOL	-	-	-	0.98 ± 0.00	0.81 ± 0.02	63 M	528 K

https://doi.org/10.1371/journal.pone.0337261.t003

Table 4. Average reliability, classification, and computational metrics (± standard deviation) over five repetitions for the CATCH using additive models.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL-ADD-ATT	0.29 ± 0.02	0.53 ± 0.00	0.90 ± 0.00	0.98 ± 0.00	0.90 ± 0.02	95 M	791 K
ABMIL-ADD-PATCH	0.06 ± 0.01	0.35 ± 0.02	0.83 ± 0.01				
CLAM-ADD-ATT	0.31 ± 0.02	0.53 ± 0.01	0.90 ± 0.01	0.98 ± 0.00	0.89 ± 0.02	95 M	799 K
CLAM-ADD-PATCH	0.07 ± 0.00	0.35 ± 0.01	0.83 ± 0.01				
DTFD-PATCH	0.09 ± 0.00	0.42 ± 0.01	0.87 ± 0.00	**1 ± 0.00**	**0.93 ± 0.01**	126 M	1842 K
ACMIL/4-ADD-ATT	0.23 ± 0.08	**0.55 ± 0.01**	**0.91 ± 0.01**	0.99 ± 0.00	0.92 ± 0.01	95 M	806 K
ACMIL/4-ADD-PATCH	0.17 ± 0.01	0.44 ± 0.02	0.88 ± 0.01				
MAX-POOL-INS	0.00 ± 0.00	0.00 ± 0.01	0.66 ± 0.00	0.99 ± 0.01	0.89 ± 0.01	63 M	528 K
MAX-POOL-INS-PATCH	0.29 ± 0.01	0.47 ± 0.06	0.88 ± 0.03				
MEAN-POOL-INS	**0.32 ± 0.00**	0.45 ± 0.01	0.85 ± 0.00	0.98 ± 0.00	0.81 ± 0.02	**63 M**	**528 K**

https://doi.org/10.1371/journal.pone.0337261.t004

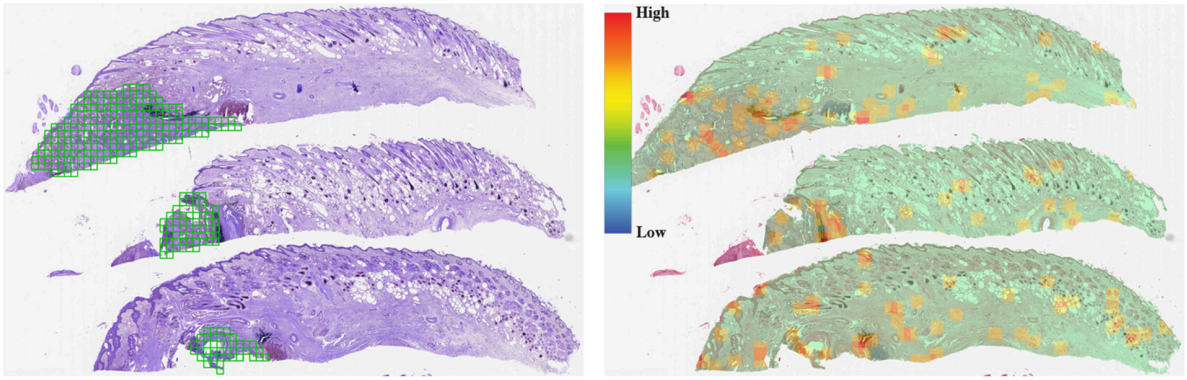


Fig 4. (I) A slide from CATCH with ground truth annotations (green) overlaid on the tissue section. (II) Corresponding heatmap generated by MAX-POOL, showing predicted patch scores distribution from low (blue) to high (red). The annotation and heatmap are spatially aligned for comparison.

https://doi.org/10.1371/journal.pone.0337261.g004

Additionally, different models excel in various reliability metrics. Among them, the ACMIL/4 models (both plain and additive) strike a good balance between classification and reliability performance, although they are computationally demanding. In Fig 5, the heatmap produced by the plain ACMIL/4 model demonstrates that the model accurately focuses on regions of interest as annotated by pathologists.

Finally, the MEAN-POOL-INS model, while being the most computationally efficient, exhibits the lowest F1 score but achieves moderate reliability metrics.

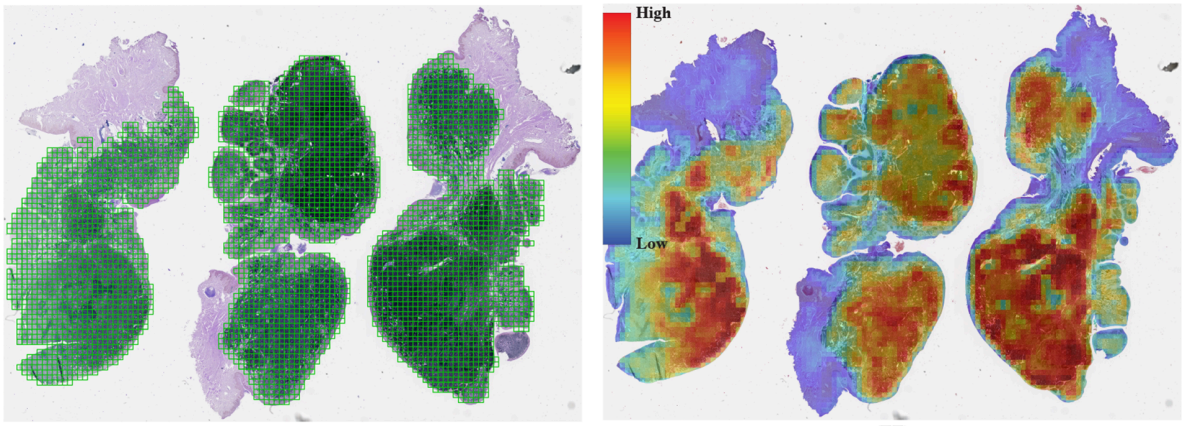


Fig 5. (I) A slide from CATCH with ground truth annotations (green) overlaid on the tissue section. (II) Corresponding heatmap generated by ACMIL/4, showing predicted patch scores distribution from low (blue) to high (red). The annotation and heatmap are spatially aligned for comparison.

https://doi.org/10.1371/journal.pone.0337261.g005

TCGA BRCA. The MADMIL/3 models exhibit the highest reliability performance, a high AUC classification metric, and low computational costs for this dataset, as shown in Table 5.

Fig 6 visually supports these numerical results. On the other hand, the DTFD model achieves the highest F1 score but demonstrates lower reliability, AUC, and computational efficiency. Additionally, we present a slide's heatmap generated by the MEAN-POOL-INS model in Fig 7, which aligns with the model's moderate reliability, high classification AUC, and low computational cost reported in Table 6.

Discussion

Based on the overall mean of the average results across the three datasets, as presented in Tables 7, 8, and Fig 8, several key conclusions can be drawn. Repeated-measures ANOVA revealed significant differences across models for all evaluated metrics. For each model the three datasets were treated as repeated measures. Specifically, significant effects of model were observed on MI ($F(5, 20) = 178.47$, $p<0.001$), Spearman correlation ($F(5, 20) = 12.34$, $p = 0.0001$), AUPRC ($F(5, 20) = 9.12$, $p = 0.0005$), AUC ($F(6, 24) = 15.21$, $p<0.001$), and F1 score ($F(6, 24) = 8.45$, $p = 0.0012$), indicating that model performances differed significantly across datasets. The learning curves for all methods on the TCGA BRCA dataset are provided in the Supporting information. These curves illustrate the convergence dynamics and stability across folds, supporting the reported performance metrics.

Table 5. Average reliability, classification, and computational metrics (± standard deviation) over five repetitions for TCGA BRCA.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL	0.19 ± 0.04	0.44 ± 0.04	0.82 ± 0.02	0.95 ± 0.01	0.85 ± 0.02	94 M	789 K
CLAM	0.18 ± 0.01	0.48 ± 0.01	0.84 ± 0.01	0.95 ± 0.01	0.84 ± 0.01	94 M	791 K
DTFD	0.15 ± 0.03	0.47 ± 0.01	0.84 ± 0.01	0.96 ± 0.01	**0.87 ± 0.02**	126 M	1053 K
ACMIL/3	0.22 ± 0.01	0.49 ± 0.02	0.85 ± 0.01	0.96 ± 0.01	0.85 ± 0.02	95 M	792 K
MADMIL/3	**0.22 ± 0.02**	**0.49 ± 0.03**	**0.85 ± 0.02**	0.97 ± 0.01	0.85 ± 0.04	74 M	615 K
MAX-POOL	0.00 ± 0.00	-0.06 ± 0.00	0.61 ± 0.00	0.93 ± 0.00	0.80 ± 0.02	63 M	526 K
MEAN-POOL	-	-	-	**0.98 ± 0.01**	0.86 ± 0.02	**63 M**	**526 K**

https://doi.org/10.1371/journal.pone.0337261.t005

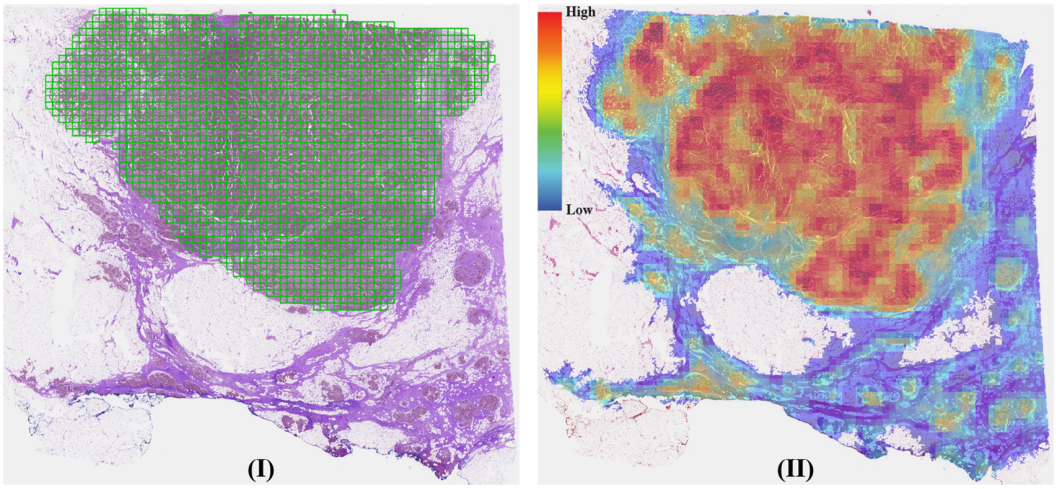


Fig 6. (I) A slide from TCGA BRCA with ground truth annotations (green) overlaid on the tissue section. (II) Corresponding heatmap generated by MADMIL/3, showing predicted patch scores distribution from low (blue) to high (red). The annotation and heatmap are spatially aligned for comparison.

https://doi.org/10.1371/journal.pone.0337261.g006

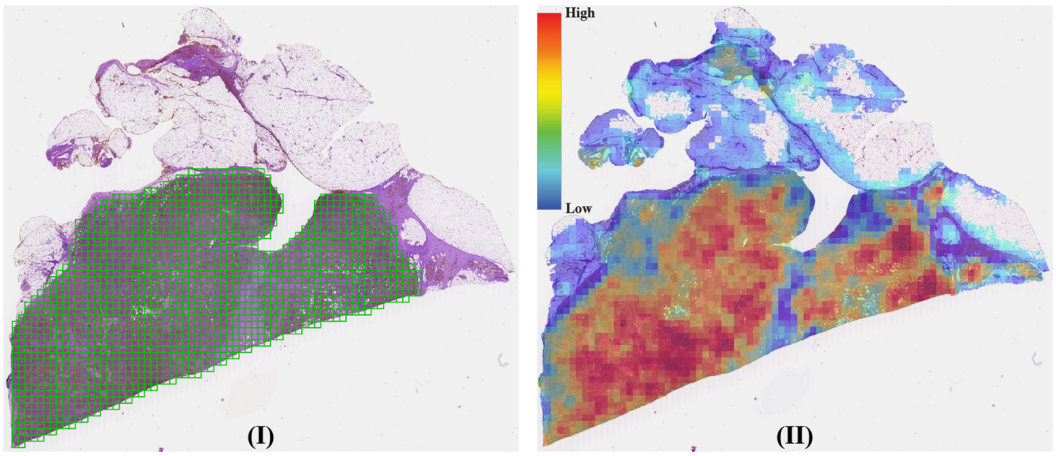


Fig 7. (I) A slide from TCGA BRCA with ground truth annotations (green) overlaid on the tissue section. (II) Corresponding heatmap generated by MEAN-POOL-INS, showing predicted patch scores distribution from low (blue) to high (red). The annotation and heatmap are spatially aligned for comparison.

https://doi.org/10.1371/journal.pone.0337261.g007

The model comparisons highlight important trade-offs. First, while the simple MAX-POOL network achieves high classification metrics and low computational cost, it demonstrates poor reliability values. This inconsistency may compromise clinical trust, suggesting that despite its efficiency, MAX-POOL is unsuitable for deployment in real-world applications where reliability is critical. In contrast, the MEAN-POOL-INS model shows high reliability, comparatively good classification performance, and low computational cost, making it a more favorable choice for scenarios where consistent predictions and resource constraints are both important.

Table 6. Average reliability, classification, and computational metrics (± standard deviation) over five repetitions for TCGA BRCA using additive models.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL-ADD-ATT	0.19 ± 0.03	0.41 ± 0.10	0.81 ± 0.04	0.91 ± 0.04	0.80 ± 0.04	95 M	789 K
ABMIL-ADD-PATCH	0.07 ± 0.03	0.33 ± 0.03	0.80 ± 0.01				
CLAM-ADD-ATT	0.14 ± 0.04	**0.48 ± 0.02**	0.84 ± 0.02	0.92 ± 0.01	0.79 ± 0.06	95 M	791 K
CLAM-ADD-PATCH	0.06 ± 0.02	0.29 ± 0.04	0.79 ± 0.00				
DTFD-PATCH	0.07 ± 0.01	0.29 ± 0.02	0.79 ± 0.01	0.96 ± 0.01	**0.87 ± 0.02**	126 M	1053 K
ACMIL/3-ADD-ATT	0.12 ± 0.01	0.23 ± 0.30	0.73 ± 0.12	0.94 ± 0.01	0.84 ± 0.02	95 M	792 K
ACMIL/3-ADD-PATCH	0.06 ± 0.01	0.11 ± 0.08	0.70 ± 0.04				
MAX-POOL-INS	0.00 ± 0.00	0.01 ± 0.01	0.61 ± 0.00	0.89 ± 0.04	0.73 ± 0.06	63 M	526 K
MAX-POOL-INS-PATCH	0.24 ± 0.04	0.45 ± 0.04	**0.84 ± 0.02**				
MEAN-POOL-INS	**0.27 ± 0.00**	0.30 ± 0.01	0.79 ± 0.01	**0.98 ± 0.01**	0.85 ± 0.02	**63 M**	**526 K**

https://doi.org/10.1371/journal.pone.0337261.t006

Table 7. The overall mean of the average reliability, classification, and computation metrics across CAMELYON16, CATCH, and TCGA BRCA.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL	0.20	0.42	0.77	0.93	0.87	94 M	790 K
CLAM	0.21	0.44	0.78	0.93	0.86	94 M	793 K
DTFD	0.16	0.44	0.79	**0.95**	**0.88**	126 M	1316 K
ACMIL	**0.22**	**0.46**	0.78	0.93	0.86	94 M	797 K
MADMIL	0.22	0.45	**0.79**	0.94	0.87	77 M	644 K
MAX-POOL	0.00	-0.04	0.50	0.91	0.84	**63 M**	**527 K**
MEAN-POOL	-	-	-	0.85	0.75	63 M	527 K

https://doi.org/10.1371/journal.pone.0337261.t007

Table 8. The overall mean of the average reliability, classification, and computation metrics for additive models across CAMELYON16, CATCH, and TCGA BRCA datasets.

Model	Reliability			Classification		Computation	
	MI	Spearman's	AUPRC	AUC	F1	FLOPs	Size
ABMIL-ADD-ATT	0.21	0.42	0.78	0.92	0.84	95 M	790 K
ABMIL-ADD-PATCH	0.06	0.30	0.70				
CLAM-ADD-ATT	0.19	**0.43**	0.78	0.92	0.83	95 M	793 K
CLAM-ADD-PATCH	0.05	0.29	0.70				
DTFD-PATCH	0.07	0.33	0.73	**0.95**	**0.88**	126 M	1316 K
ACMIL-ADD-ATT	0.18	0.37	0.75	0.91	0.86	95 M	797 K
ACMIL-ADD-PATCH	0.09	0.17	0.62				
MAX-POOL-INS	0.00	0.02	0.49	0.88	0.79	63 M	527 K
MAX-POOL-INS-PATCH	0.23	0.40	0.76				
MEAN-POOL-INS	**0.30**	0.43	**0.81**	0.85	0.75	**63 M**	**527 K**

https://doi.org/10.1371/journal.pone.0337261.t008

The multihead networks, ACMIL and MADMIL, offer both high reliability and classification performance, but at a higher computational expense, highlighting their potential for research or clinical applications where both accuracy and consistency are essential. Their higher computational cost, particularly for ACMIL, indicates a trade-off that must be considered in resource-limited settings. Meanwhile, the DTFD model delivers the highest classification performance, though this comes at the cost of increased computational demands and reduced reliability. Therefore, when classification performance is a priority, and computational cost and reliability are secondary concerns, DTFD may be a suitable choice.

Regarding the additive models, while they do not exhibit a significant improvement in classification metrics, they do show a subtle increase in reliability scores, albeit with slightly higher computational cost. While they offer enhanced interpretability, their limited benefits in reliability and accuracy indicate that they may be most valuable in research contexts emphasizing model explainability rather than operational performance.

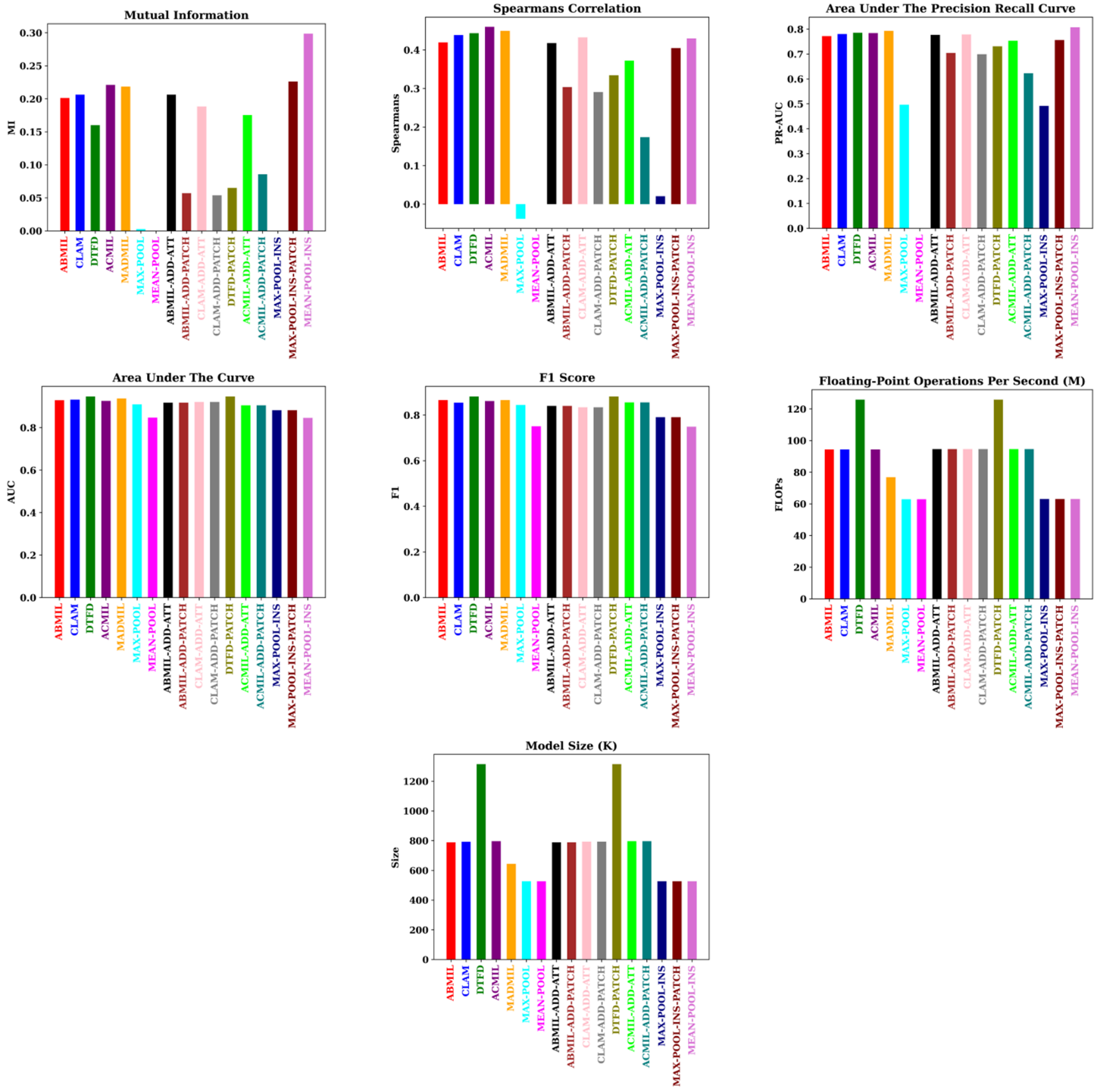


Fig 8. Bar plots comparing the models with different metrics.

https://doi.org/10.1371/journal.pone.0337261.g008

To better understand the contribution of each reliability metric, we examined the model performance by selectively excluding MI, Spearman correlation, or AUPRC from the overall model ranking (Table 9). The results show that excluding

Table 9. Analysis of reliability metrics showing the effect of excluding each metric on model ranking. Rankings are obtained by summing the scores of the selected reliability metrics (with equal weight) across CAMELYON16, CATCH, and TCGA BRCA datasets, and ranking models based on the aggregated score.

Metric Excluded	Model Ranking (Reliability)
None (All metrics)	MEAN-POOL-INS > ACMIL > MADMIL > CLAM > ABMIL-ADD-ATT > CLAM-ADD-ATT
MI excluded	MEAN-POOL-INS = ACMIL = MADMIL > DTFD > CLAM > CLAM-ADD-ATT
Spearman's excluded	MEAN-POOL-INS > MADMIL > ACMIL > CLAM = ABMIL-ADD-ATT = MAX-POOL-INS-PATCH
AUPRC excluded	MEAN-POOL-INS > ACMIL > MADMIL > CLAM > ABMIL-ADD-ATT = MAX-POOL-INS-PATCH

https://doi.org/10.1371/journal.pone.0337261.t009

MI substantially reduces the ability to differentiate models with consistent predictions, as several top models become tied, highlighting MI's critical role in capturing prediction consistency. Excluding AUPRC also affects the rankings, emphasizing its importance in assessing realiability. Similarly, removing Spearman correlation impacts the overall ranking, causing several models to tie at lower positions. These findings indicate that all three metrics contribute meaningfully to assessing model reliability, and a multi-metric framework is necessary to capture different aspects of predictive behavior effectively.

Overall, these findings underscore that classification metrics alone do not capture the full utility of a model. Incorporating reliability and computational cost alongside accuracy provides a more comprehensive assessment, guiding model selection. While standard metrics such as AUC, accuracy, and F1 capture overall classification performance, they can fail in certain scenarios, such as rare positive instances, highly imbalanced classes, or heterogeneous MIL bags. In our experiments, the MAX-POOL network achieves high AUC and F1 scores but demonstrates low reliability, highlighting that strong classification metrics alone do not guarantee trustworthy predictions. Incorporating reliability metrics alongside standard metrics allows identification of models that not only perform well on average but also make consistent and robust predictions, providing a more comprehensive assessment of model behavior and guiding informed model selection. By explicitly considering these factors, researchers can design MIL architectures that balance predictive performance, computational efficiency, and deployability, ultimately advancing the development of models suited for real-world WSI analysis.

Conclusion

In this paper, we proposed a novel approach to comparing MIL models in terms of reliability using three distinct metrics. Notably, we found that the MEAN-POOL-INS model achieves high reliability despite its simple architecture and low computational cost. Additionally, the multihead models, while slightly more computationally expensive, demonstrate higher classification performance and maintain fair reliability.

We hope that future research will incorporate reliability and computational metrics, leading to the development of more dependable and efficient models for WSI classification.

Supporting information

S1 Fig. Learning curve of the models (set 1). Each panel shows the training curve for one model on TCGA BRCA.
(TIF)

S2 Fig. Learning curve of the models (set 2). Each panel shows the training curve for one model on TCGA BRCA.
(TIF)

Author contributions

Conceptualization: Hassan Keshvarikhojasteh, Mitko Veta.

Data curation: Marc Aubreville, Christof A. Bertram.

Formal analysis: Hassan Keshvarikhojasteh.

Methodology: Hassan Keshvarikhojasteh, Mitko Veta.

Supervision: Josien P.W. Pluim, Mitko Veta.

Writing – original draft: Hassan Keshvarikhojasteh.

Writing – review & editing: Hassan Keshvarikhojasteh, Marc Aubreville, Christof A. Bertram, Josien P.W. Pluim, Mitko Veta.

References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8. https://doi.org/10.1038/nature21056 PMID: 28117445
2. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
3. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559–67. https://doi.org/10.1038/s41591-018-0177-5 PMID: 30224757
4. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301–9. https://doi.org/10.1038/s41591-019-0508-1 PMID: 31308507
5. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell. 1997;89(1–2):31–71.
6. Carbonneau M-A, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognition. 2018;77:329–53. https://doi.org/10.1016/j.patcog.2017.10.009
7. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning (ICML 2018); 2018 Jul 10–15; Stockholm, Sweden. PMLR; 2018. p. 2127–36.
8. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555–70. https://doi.org/10.1038/s41551-020-00682-w PMID: 33649564
9. Zhang H, Meng Y, Zhao Y, Qiao Y, Yang X, Coupland SE. Dtfd-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA. 2022. p. 18802–12.
10. Zhang Y, Li H, Sun Y, Zheng S, Zhu C, Yang L, et al. Attention-challenging multiple instance learning for whole slide image classification. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2024: Proceedings of the 18th European Conference on Computer Vision; 2024 Sep 29–Oct 4; Milan, Italy. Cham: Springer Nature; 2024. p. 125–43.
11. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J. 2018;16:34–42.
12. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. https://doi.org/10.1186/s12916-019-1426-2 PMID: 31665002
13. Pendyala V, Kim H. Assessing the reliability of machine learning models applied to the mental health domain using explainable AI. Electronics. 2024;13(6):1025. https://doi.org/10.3390/electronics13061025
14. Araujo Santos VC, Cardoso L, Alves R. The quest for the reliability of machine learning models in binary classification on tabular data. Sci Rep. 2023;13(1):18464. https://doi.org/10.1038/s41598-023-45876-9 PMID: 37891221
15. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol Artif Intell. 2021;3(6):e200267. https://doi.org/10.1148/ryai.2021200267 PMID: 34870212
16. Biran O, Cotton C. Explanation and justification in machine learning: a survey. In: Melbourne, Australia. 2017. p. 8–13.
17. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput Biol Med. 2022;140:105111.
18. Javed SA, Juyal D, Padigela H, Taylor-Weiner A, Yu L, Prakash A, et al. Additive MIL: intrinsically interpretable multiple instance learning for pathology. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022). New Orleans, LA, USA. 2022. p. 20689–702.
19. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318(22):2199–210. https://doi.org/10.1001/jama.2017.14585 PMID: 29234806
20. Wilm F, Fragoso M, Marzahl C, Qiu J, Puget C, Diehl L, et al. Pan-tumor CAnine cuTaneous Cancer Histology (CATCH) dataset. Sci Data. 2022;9(1):588. https://doi.org/10.1038/s41597-022-01692-w PMID: 36167846
21. Ross BC. Mutual information between discrete and continuous data sets. PLoS One. 2014;9(2):e87357. https://doi.org/10.1371/journal.pone.0087357 PMID: 24586270

22. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15(1):72–101.
23. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806
24. Banerjee S, Bertram CA, Ammeling J, Weiss V, Conrad T, Klopfleisch R, et al. Comprehensive dataset of coarse tumor annotations for the cancer genome atlas breast invasive carcinoma. In: Palm C, Breininger K, Deserno TM, Handels H, Maier A, Maier-Hein K, et al., editors. Proceedings of the BVM Workshop; 2025 Mar 2–4; Wiesbaden, Germany. Wiesbaden: Springer Fachmedien; 2025. p. 260–5.
25. Keshvarikhojasteh H, Pluim J, Veta M. Multi-head attention-based deep multiple instance learning. arXiv preprint 2024. https://arxiv.org/abs/2404.05362