

CAMEL: Contrastive-powered Attention Multiple-instance learning for Enhanced mesothelioma classification

Andrea Zenotto Emanuele Carelli Luca Pellicciotti Sabrina Basile
s327473@studenti.polito.it s333957@studenti.polito.it s331419@studenti.polito.it s329161@studenti.polito.it

Politecnico di Torino

CONTENTS

I	Introduction	1
II	Background	1
III	Materials and methods	2
III-A	Dataset and Preprocessing	2
III-B	Self-Supervised Feature Extraction	3
III-C	Multiple Instance Learning Framework	3
III-D	Comparison with Fine-tuned CNN	4
III-E	Explainable AI (XAI) Methods	4
III-E1	Out-of-Distribution Detection	4
III-E2	Attention-Based Saliency Maps	4
IV	Results and discussion	5
IV-A	Evaluation on the CAMEL Inference Dataset	5
IV-B	Quantitative Results	5
IV-C	Qualitative Discussion	5
IV-D	Explainable AI (XAI) Results: OOD Detection and Attention-Based Interpretability	5
IV-D1	Out-of-Distribution (OOD) Detection	5
IV-D2	Attention Maps for WSI-Level Explainability	6
V	Conclusions and future works	6
	References	7

LIST OF FIGURES

1	Overview of the framework pipeline (patches extraction and architecture)	2
2	Example of the Training Dataset patches	3
3	Attention map for an epithelioid WSI (M-65). Warmer colors indicate patches with higher attention scores, corresponding to regions considered more relevant for the classification. The blue background indicates an absence of data in this region, resulting from patches being discarded during pre-processing.	6

LIST OF TABLES


I	Selected WSI for Inference Dataset	2
II	Inference results	5
III	Comparison between our proposal and a standard CNN	5
IV	OOD detection performance	6

CAMEL: Contrastive-powered Attention Multiple-instance learning for Enhanced mesothelioma classification

Abstract—Malignant pleural mesothelioma (MPM) is a rare and aggressive cancer with high histological heterogeneity, making subtype classification, challenging even for expert pathologists. Accurate classification is essential for guiding treatment strategies and patient management.

In this work, we present a weakly supervised pipeline for MPM subtype classification using whole-slide images (WSIs), combining self-supervised representation learning and attention-based multiple instance learning (MIL). To address the domain gap between natural images and histopathology, we pretrain a feature extractor using SimCLR-based contrastive learning directly on histopathology patches, enabling the extraction of rich and meaningful embeddings without manual annotations. These features are then processed through a multi-head attention MIL module that identifies diagnostically relevant regions and aggregates patch-level information to produce slide-level predictions.

Our approach builds upon the CLAM framework and is tailored to the specific challenges of MPM. Results show promising classification performance, especially in correctly identifying biphasic cases, with potential for improvement through extended training and hyperparameter optimization. Comparative experiments with CNN baselines highlight the advantages of combining self-supervised learning and attention-based MIL in computational pathology tasks. This study underscores the feasibility of building robust AI-assisted diagnostic tools for rare cancers under limited supervision, with also an analysis of clinical reliability and interpretability through explainable AI (XAI) techniques and Out-of-Distribution (OOD) detection.

 GitHub repository

I. INTRODUCTION

Malignant pleural mesothelioma (MPM) is a rare but aggressive cancer arising from the mesothelial cells of the pleura and is strongly associated with asbestos exposure. Histologically, MPM is classified into three major subtypes: **epithelioid, sarcomatoid, and biphasic** [1]. Accurate subtype classification plays a critical role in clinical practice, guiding treatment decisions, surgical eligibility, and patient enrolment in clinical trials. However, MPM diagnosis remains challenging due to histological heterogeneity and the coexistence of multiple tissue components within the same specimen, which increases inter-observer variability even among expert pathologists.

The increasing adoption of digital pathology and Whole Slide Imaging (WSI) has opened new opportunities for computational pathology, enabling the development of artificial intelligence (AI)-assisted diagnostic tools. Deep learning models, especially those based on convolutional neural networks (CNNs), have demonstrated remarkable performance in histopathology image analysis. Nevertheless, their application to MPM subtype classification remains underexplored, primarily due to two challenges: ⁽ⁱ⁾the weakly supervised nature of available annotations, which are typically provided at the

slide level rather than at the cellular or patch level, and ⁽ⁱⁱ⁾the scarcity of large annotated datasets required to train robust models.

Multiple Instance Learning (MIL) [2] has emerged as a promising solution for weakly supervised [3] WSI classification. In this paradigm, each slide is modeled as a bag of instances (patches), with the slide-level label used to guide the learning process. Attention-based MIL approaches, such as the Clustering-constrained Attention Multiple instance learning (CLAM) framework, have shown strong potential in identifying diagnostically relevant regions despite the lack of pixel-level annotations. However, most existing studies rely on backbones pretrained on natural images (e.g., ImageNet), which may fail to capture the fine-grained morphological patterns specific to histopathology.

In this work, we propose a robust pipeline for MPM subtype classification from WSIs under weak supervision, leveraging a MIL framework inspired by the CLAM [4] paradigm. To address the limitation of domain-shift between natural and histopathological images, we employ a SimCLR-based [5] self-supervised contrastive learning strategy to pretrain the feature extractor directly on histopathology patches. This backbone generates rich and semantically meaningful patch-level embeddings without requiring extensive manual annotations. The encoded features are then aggregated using a multi-head attention (MHA) [6] MIL module, which computes attention scores to focus selectively on diagnostically significant regions. The final slide-level representation is used to classify WSIs into the three major MPM subtypes.

Our contributions are fourfold:

- 1) Development of a semi-supervised pipeline (Figure 1) that integrates self-supervised feature learning and attention-based MIL for mesothelioma subtype classification.
- 2) Demonstration of the effectiveness of SimCLR-based pretraining for extracting histopathology-specific representations.
- 3) A comparative analysis with pretrained and fine-tuned CNN baselines is presented to evaluate the robustness of the proposed approach.
- 4) Enhancement of clinical reliability and interpretability through explainable AI (XAI) techniques and Out-of-Distribution (OOD) detection.

II. BACKGROUND

The rapid advancement of deep learning has profoundly influenced digital pathology, particularly in the automatic

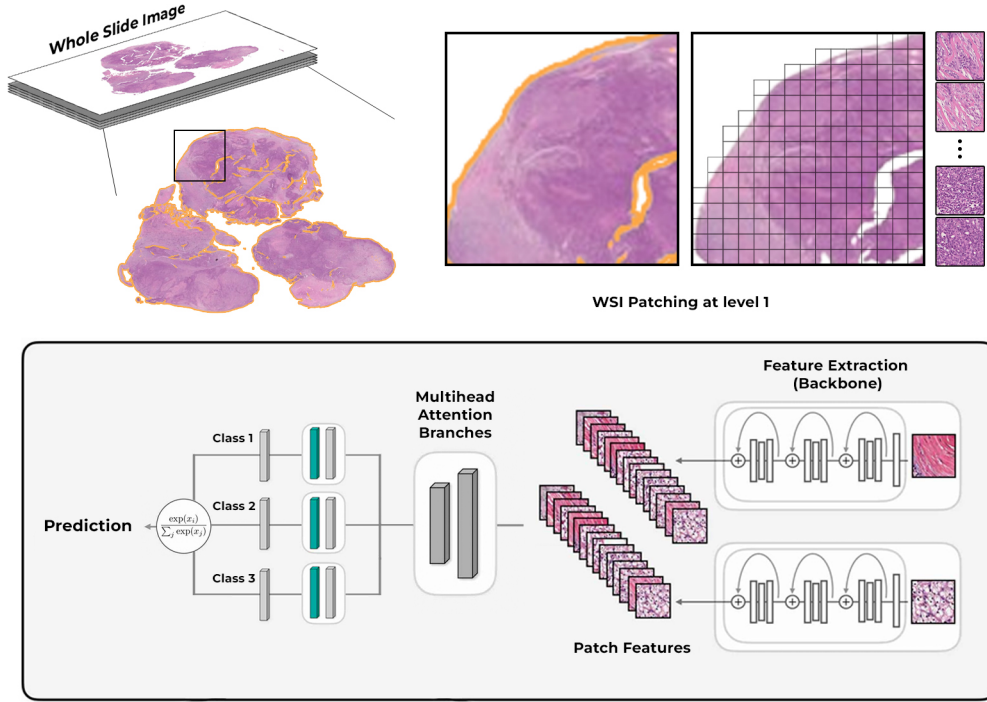


Fig. 1. Overview of the framework pipeline (patches extraction and architecture)

classification of WSIs. Due to the extremely high resolution of WSIs and the scarcity of region-level annotations, weakly supervised methods, especially MIL, have become the standard approach to tackle slide-level classification. Among these, the CLAM framework by Lu et al. [4] represents a milestone, introducing an attention-based mechanism that simultaneously identifies diagnostically relevant regions and performs slide-level classification without requiring pixel-wise labels. CLAM and its variants have been successfully applied to several cancer types, including breast, lung, and renal carcinoma, demonstrating the effectiveness of attention-based aggregation in handling large-scale histopathological data.

Building upon this, alternative MIL architectures such as DSMIL [7] and TransMIL [8] have explored different attention mechanisms and transformer-based models to better capture inter-instance relationships, further improving performance in weakly supervised settings. However, most of these methods still rely on backbones pretrained on natural images, which may not optimally capture the domain-specific morphological patterns of histopathology.

To mitigate this limitation, self-supervised and semi-supervised learning strategies have gained increasing attention. Contrastive learning methods such as SimCLR [5] and MoCo [9] have been successfully adapted to histopathology, enabling the learning of robust patch-level representations from large unlabeled datasets. Ciga et al. [10], for instance, demonstrated that self-supervised pretraining significantly boosts downstream MIL-based classification, particularly in scenarios with limited labeled data.

These advances collectively highlight the importance of combining MIL frameworks with self-supervised feature extraction to achieve accurate slide-level classification under

weak supervision, especially for rare diseases where annotated data is scarce.

III. MATERIALS AND METHODS

A. Dataset and Preprocessing

The original dataset consisted of over 100 WSIs in NDPI format, highly class-imbalanced: only 5 sarcomatoid, about 20 biphasic, and more than 70 epithelioid cases. To mitigate this imbalance and ensure a fair evaluation, we constructed a **balanced dataset** by selecting 5 WSIs per class. The selection was based on the number of valid tissue patches extracted from each slide, as described below.

TABLE I
SELECTED WSI FOR INFERENCE DATASET

MPM Subtype		Selected WSI			
Epithelioid	M-59	M-13	M-70	M-68	M-85
Sarcomatoid	M-30	M-73	M-108	M-90	M-92
Biphasic	M-65	M-101	M-86	M-114	M-87

WSIs are typically stored at multiple resolution levels. We performed a preliminary analysis to evaluate the number of tissue patches generated at different zoom levels. Our goal was to select a level that produced a sufficiently large and informative patch set, while keeping the dataset size computationally manageable. Based on this analysis, **level 1** was selected as the optimal trade-off between resolution and dataset size.

Each selected slide was tiled into non-overlapping **224 × 224 pixel patches** (Figure 2) at level 1 using the OpenSlide library. To exclude background or non-informative regions, we applied a **saturation-based filtering**: the saturation channel of

each patch was computed in the HSV color space, and patches with a mean saturation below a threshold were discarded.

The threshold value was determined by grid search, analyzing the number of valid patches for saturation values ranging from 15 to 40. A threshold of 30 was chosen, as it preserved tissue-rich areas while reducing the inclusion of white background.

For each diagnostic group, we ranked WSIs according to the number of valid tissue patches and selected the **top 5 slides per class** (Table I). This ensured that the final dataset was both balanced and representative of tissue variability within each category.

From the original WSIs, we generated **three subsets**:

- **15 WSIs** for training
- **8 WSIs** for evaluation (unfortunately, none sarcomatoid due to scarcity of original WSIs in this class)
- A third dataset composed of **OOD patches** from the previous datasets was created for a subsequent explainability analysis.

All valid patches from the selected WSIs were extracted and organized into class-specific folders. This preprocessing step produced a dataset of several thousand patches, which was subsequently used for feature extraction and slide-level classification.

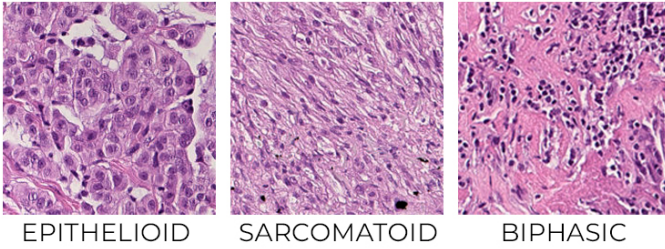


Fig. 2. Example of the Training Dataset patches

B. Self-Supervised Feature Extraction

To efficiently handle the large number of patches extracted from WSIs and accelerate training and inference of the downstream multiple instance learning model (mhaMIL), we employed a self-supervised learning (SSL) approach based on the **SimCLR framework**.

SimCLR learns feature representations by maximizing the agreement between **two augmented views of the same patch** (positive pairs) while minimizing the similarity with views of different patches (negative pairs), without requiring manual annotations. Specifically, given a batch of patches, each patch is augmented twice to generate two correlated views z_i and z_j . The model maximizes the cosine similarity between these views while reducing similarity with all other representations in the batch.

The **Normalized Temperature-scaled Cross-Entropy (NT-Xent)** loss for a positive pair (i, j) is defined as:

$$\ell_{i,j} = -\log \frac{\exp\left(\frac{\text{sim}(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\frac{\text{sim}(z_i, z_k)}{\tau}\right)}$$

where τ is a temperature parameter, N the batch size, and $\mathbf{1}_{[k \neq i]}$ excludes self-similarity.

We fine-tuned all layers of a **ResNet50 backbone** pretrained on ImageNet. A projection head, composed of a global average pooling followed by two fully connected layers with 2048 ReLU units and 128 linear units, mapped the encoder outputs to the embedding space used for contrastive learning.

During training, each patch was augmented twice with random flips, crops, brightness and contrast jittering, and Gaussian noise to promote invariance. The model was trained for 40 epochs with a batch size of 128 using the AdamW optimizer (initial learning rate 2×10^{-4} , weight decay 1×10^{-5}), leveraging multi-GPU parallelism via TensorFlow's MirroredStrategy.

After training, the projection head was discarded, and the 2048-dimensional encoder output was used as the patch feature representation. These embeddings were precomputed for all patches, providing a compact representation that significantly reduces memory usage and computational costs during mhaMIL training and inference.

C. Multiple Instance Learning Framework

In this work, each WSI is modeled as a *bag* of instances, where instances correspond to patch-level feature embeddings extracted by a frozen backbone network. Since only slide-level labels are available, we adopt a **MIL** approach to aggregate instance information into a global representation.

We implement a **MHA** mechanism to assign different importance weights to individual patches. Given instance embeddings $\{h_1, h_2, \dots, h_n\}$, they are projected into query (Q), key (K), and value (V) spaces, and attention scores are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k is the dimensionality of the key vectors. Using **four attention heads**, the model learns to capture complementary morphological patterns, enhancing the discriminative power of the aggregated representation.

The outputs from all heads are concatenated and normalized, and a **weighted pooling** guided by the attention scores aggregates patch-level information into a single slide-level embedding. This embedding is then passed through a dense projection layer and finally through a fully connected classifier to predict the slide label.

The dataset is constructed by iterating through the patch directory structure, where patches are organized by WSI and class label. Each patch is assigned a one-hot encoded label corresponding to one of the three diagnostic categories: *epithelioid*, *sarcomatoid*, and *biphasic*. Formally, for each class $c \in \{\text{epithelioid}, \text{sarcomatoid}, \text{biphasic}\}$, the mapping is defined as:

- epithelioid $\rightarrow [1, 0, 0]$
- sarcomatoid $\rightarrow [0, 1, 0]$
- biphasic $\rightarrow [0, 0, 1]$

A data pipeline is built using **TensorFlow Datasets (TFDS)**, applying standard data augmentation and batching to improve generalization.

Training is performed end-to-end on the extracted patch features using a cross-folding strategy to enhance model generalization. Specifically, at each epoch, the dataset is randomly partitioned into training and validation subsets, ensuring diverse data splits throughout the training process.

The backbone network, either a ResNet50 pretrained on ImageNet or a self-supervised pretrained model, is frozen to prevent updates during MIL training, thus reducing computational cost and overfitting. The downstream multiple instance learning model employs a multi-head attention mechanism to aggregate patch-level features into a slide-level representation for classification.

Model training is conducted over 20 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} . The categorical cross-entropy loss is optimized, and several metrics, including accuracy, precision, recall, area under the ROC curve (AUC), and F1-score, are monitored to evaluate performance. A model checkpointing mechanism saves the best-performing model based on the training loss to ensure optimal model selection.

At each epoch, features are re-extracted from augmented patches to incorporate variability, and the model is trained for a single epoch on the current training split while being validated on the corresponding validation split. This cross-folding approach improves robustness and helps mitigate overfitting by exposing the model to different training-validation distributions during the course of training.

D. Comparison with Fine-tuned CNN

To provide a baseline for evaluating the effectiveness of the MIL framework, we trained a fully supervised convolutional neural network on patch-level annotations. The approach leverages a **fine-tuned ResNet50** pretrained on ImageNet, adapting it to the mesothelioma classification task.

The network architecture consists of two main components:

- 1) **Feature Extraction** – A ResNet50 backbone, initialized with ImageNet weights, is used to extract high-level morphological features from 224×224 patches. The backbone operates in frozen mode during training to preserve pre-learned low-level representations.
- 2) **Classification Head** – The extracted 2048-dimensional feature vector is passed through a lightweight fully connected classifier composed of a 128-unit ReLU layer followed by a softmax layer with three output neurons.

The model is trained using the **AdamW optimizer** (learning rate $lr = 1 \times 10^{-4}$, weight decay 1×10^{-5}) and **categorical cross-entropy loss** for 10 epochs, with mini-batches of 128 patches.

E. Explainable AI (XAI) Methods

To improve model interpretability and reliability, we integrated two key explainability techniques: OOD detection via thresholding mechanisms and visualization of attention saliency maps.

1) *Out-of-Distribution Detection*: To enable robust OOD detection, we implemented a threshold-based rejection mechanism applied to the model’s output probabilities. Specifically, class-specific thresholds were defined to distinguish confident, in-distribution (ID) predictions from uncertain or OOD cases. During inference, if the predicted class probability falls below the corresponding threshold, the sample is rejected as OOD rather than assigned a potentially incorrect label.

These thresholds were not set arbitrarily but optimized via a fine-tuning procedure over a grid of candidate thresholds for each class, aiming to maximize a weighted metric called the *Out-of-distribution Rejection Score* (ORS):

$$ORS_{\alpha} = \alpha \cdot TRR + (1 - \alpha) \cdot TAR$$

where:

- **True Accept Rate (TAR)** measures the proportion of correctly accepted ID samples (sensitivity):

$$TAR = \frac{\text{Correctly accepted ID samples}}{\text{Total ID samples}}$$

- **True Reject Rate (TRR)** measures the proportion of correctly rejected OOD samples (specificity):

$$TRR = \frac{\text{Correctly rejected OOD samples}}{\text{Total OOD samples}}$$

The parameter $\alpha \in [0, 1]$ balances the trade-off between clinical safety (rejecting OOD samples) and coverage (correctly accepting ID samples). Higher values of α emphasize safety by penalizing false acceptance of OOD samples more heavily.

By optimizing thresholds according to the ORS, the system identifies cases where no class prediction reaches sufficient confidence; such inputs are rejected rather than arbitrarily classified, reducing diagnostic risk and improving reliability in clinical deployment. This thresholding strategy, grounded in explainable and tunable criteria, enhances model robustness to unseen or ambiguous inputs without requiring retraining or dedicated OOD architectures.

2) *Attention-Based Saliency Maps*: To complement the slide-level classification performed by our Multiple Instance Learning (MIL) model, we extract and visualize attention maps that spatially reflect the relevance of each patch within the whole-slide image (WSI). This approach integrates seamlessly without requiring pixel-level annotations.

Patch coordinates are extracted directly from the input filenames, which follow the format `patch_{x}_{y}.png`, encoding the spatial position of each patch within the slide. A parsing function maps each patch to its (x, y) location on the slide grid.

During inference, the MIL model produces self-attention matrices over patch-level embeddings. We compute a scalar attention score for each patch by averaging the diagonal elements (self-attention weights) across all attention heads, effectively quantifying the importance of each patch for the final prediction.

These scores are projected back onto the spatial grid of the original slide. Missing patches are handled gracefully by

marking their positions as NaN, preserving the spatial integrity of the map.

For visualization, the raw attention map is normalized and resized to match a low-resolution WSI thumbnail. The resulting heatmap is overlaid on the slide using a colormap, highlighting regions most influential in the model’s decision. This overlay is displayed alongside the original WSI, enabling qualitative comparison and expert assessment.

This attention-based mechanism enhances the transparency of our MIL approach by associating predictions with interpretable spatial evidence, thus supporting clinical insights and model trustworthiness.

IV. RESULTS AND DISCUSSION

A. Evaluation on the CAMEL Inference Dataset

The performance of the proposed MIL framework was evaluated on the *CAMEL inference dataset*, which consists of eight WSIs: four epithelioid and four biphasic cases. No sarcomatoid cases were included in this test set, and all samples were exclusively used for inference to ensure an unbiased assessment.

B. Quantitative Results

Table II reports the classification results obtained with our attention-based MIL model. The model achieved an overall accuracy of **87.50%**, correctly classifying seven out of eight WSIs. Only one epithelioid case was misclassified as sarcomatoid. The misclassification counts per subtype are also reported.

TABLE II
INFERENCE RESULTS

Inference Results								
WSI	I	II	III	IV	V	VI	VII	VIII
Prediction	E	B	E	B	S	E	B	B
Label	E	B	E	B	E	E	B	B
Accuracy	87.50%							

For comparison (Table III), the ResNet-based WSI-level classifier trained with a standard supervised strategy achieved an accuracy of only **50.00%** on the same dataset, highlighting the superior generalization capability of the MIL approach when dealing with weakly labeled WSIs.

TABLE III
COMPARISON BETWEEN OUR PROPOSAL AND A STANDARD CNN

Model	Accuracy
Our Pipeline	87.50%
Standard CNN	50.00%

C. Qualitative Discussion

The attention-based MIL model demonstrated strong performance in distinguishing between epithelioid and biphasic cases, correctly identifying all biphasic WSIs. The only misclassification occurred in an epithelioid case, predicted as sarcomatoid, likely due to the morphological heterogeneity of the sample or the limited representation of sarcomatoid features in the training set, which may have caused the attention mechanism to focus on misleading regions.

The comparatively lower performance of the supervised ResNet classifier suggests that end-to-end WSI-level training struggles to cope with the high intra-class variability and weak labeling typical of mesothelioma histopathology. In contrast, the MIL paradigm effectively aggregates patch-level information, yielding a more robust representation for final decision-making.

Despite limited resources, including a highly imbalanced and biased dataset collected from a single center, and time constraints that prevented extensive hyperparameter fine-tuning, the results obtained appear promising. The current pipeline should therefore be considered a baseline.

D. Explainable AI (XAI) Results: OOD Detection and Attention-Based Interpretability

To enhance the clinical reliability and interpretability of the proposed CLAM attention-based model, we conducted experiments focusing on two complementary aspects of explainability: **OOD detection** and **attention-based visualization at the WSI level**.

1) Out-of-Distribution (OOD) Detection: To evaluate the robustness of the model under domain shifts and unexpected input variations, we performed dedicated OOD detection experiments following the methodology described in Section III-E. The objective was to quantify the model’s ability to reject unreliable predictions on distorted WSI patches while maintaining high accuracy on in-distribution (ID) samples.

The evaluation was conducted on a combined dataset including both ID and synthetically generated OOD patches. The OOD samples were produced by applying *Gaussian noise* (mean = 0, standard deviation = 30, pixel intensity scale [0–255]) and *Gaussian blur* (kernel size = 11, $\sigma = 5$) to a balanced subset of WSI patches from the three mesothelioma subtypes (epithelioid, sarcomatoid, biphasic). The trained CLAM model was used to extract class probabilities for each patch, and a **threshold-based rejection mechanism** was applied: a prediction was rejected as OOD if the maximum class probability did not exceed a class-specific threshold.

The thresholds were optimized by maximizing the *Out-of-distribution Rejection Score* (ORS), balancing *True Accept Rate* (TAR) and *True Reject Rate* (TRR) with $\alpha = 0.7$ to prioritize clinical safety. A grid search identified the best performing thresholds as:

$$T_e = 0.20, \quad T_s = 0.10, \quad T_b = 0.10 \quad (1)$$

achieving an **ORS of 90.0%**.

As shown in Table IV, the tuned rejection mechanism significantly improved model reliability under distributional shifts, achieving high TRR (100%) while maintaining a good TAR on ID cases.

TABLE IV
OOD DETECTION PERFORMANCE

Metric	ID	OOD
TAR	66.70%	–
TRR	–	100.0%
Accuracy	83.3%	

Qualitative inspection confirmed the system’s ability to reject ambiguous or noisy patches rather than forcing uncertain classifications. This **thresholding strategy**, requiring no additional training or architectural changes, effectively enhances clinical reliability by reducing overconfident predictions on unseen conditions.

2) *Attention Maps for WSI-Level Explainability*: To further improve interpretability, we analyzed **attention maps** generated by the MIL framework (Section III-E). These maps highlight the relative contribution of each patch to the slide-level prediction, providing spatial interpretability aligned with tissue morphology.

Figure 3 shows an example for the *epithelioid WSI M-65*, where the model assigns higher attention scores (warm colors) to tumor regions considered more discriminative for classification, while less relevant or empty regions receive lower scores. The attention distribution is consistent with histopathological patterns: regions with higher attention correspond to denser tumor areas, validating the biological plausibility of the predictions.

This **attention-based explainability** represents a valuable qualitative tool, as it highlights the critical tissue regions influencing automated decisions, potentially helping

pathologists verify and understand the model’s reasoning.

V. CONCLUSIONS AND FUTURE WORKS

In this work, we proposed a semi-supervised approach for the classification of pleural mesothelioma subtypes based on WSIs, leveraging a CLAM attention-based multi-instance learning framework. The model demonstrated promising performance in differentiating epithelioid, sarcomatoid, and biphasic subtypes, while the integration of OOD detection enhanced the system’s robustness under distributional shifts. Furthermore, the attention-based interpretability analysis confirmed the biological plausibility of the model’s decisions, highlighting relevant tissue regions consistent with histopathological patterns.

Overall, the results suggest that combining weakly-supervised learning with explainability mechanisms can provide a valuable decision-support tool for pathologists, improving clinical reliability while maintaining interpretability. Nonetheless, some limitations persist, particularly regarding the misclassification of sarcomatoid samples, which reflects the intrinsic histopathological variability of this subtype.

Future work will focus on expanding the available datasets to improve the generalization of the model and systematically exploring different parameter configurations. In particular, we plan to adjust the *backbone loss temperature*, modify the number of attention heads in the *mhaMIL* model, experiment with alternative backbone architectures, and refine the attention mechanism. We expect that improved training strategies and fine-tuning will lead to higher accuracy, more stable OOD rejection, and even better interpretability, ultimately bringing the system closer to practical clinical adoption.

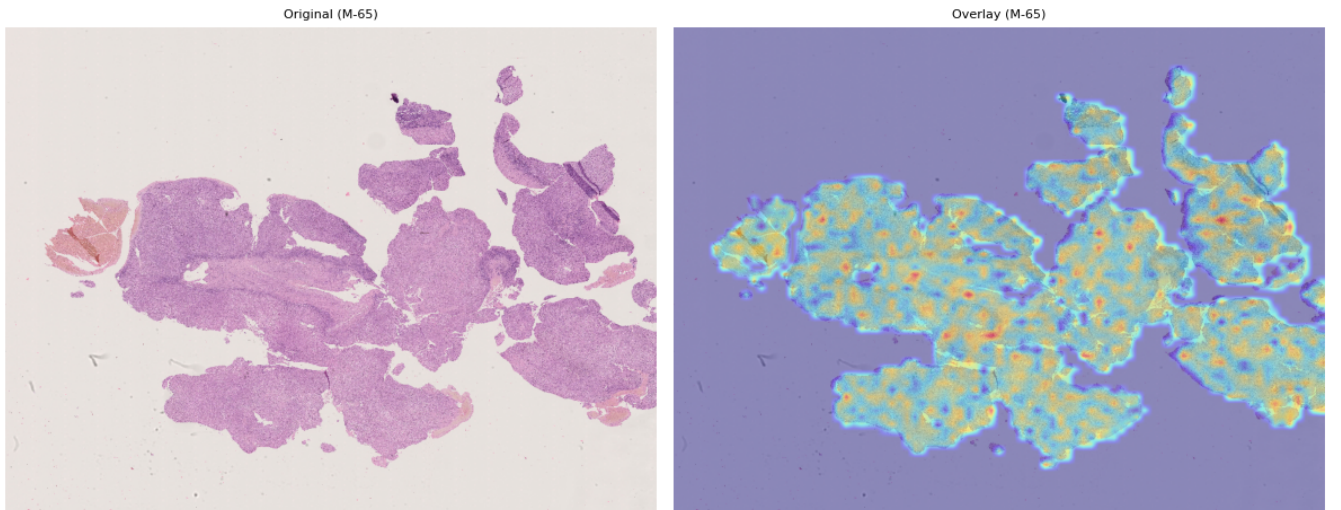


Fig. 3. Attention map for an epithelioid WSI (M-65). Warmer colors indicate patches with higher attention scores, corresponding to regions considered more relevant for the classification. The blue background indicates an absence of data in this region, resulting from patches being discarded during pre-processing.

REFERENCES

- [1] L. Brcic and I. Kern, "Impact of the jenkyns event on shallow-marine carbonates and coeval emerged paleoenvironments: the plitvice lakes region, croatia," *Palaeogeography, Palaeoclimatology, Palaeoecology*, vol. 655, p. 112519, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S003101822400508X>
- [2] M. Gadermayr and M. Tschuchnig, "Multiple instance learning for digital pathology: A review on the state-of-the-art, limitations & future potential," *arXiv preprint arXiv:2206.04425*, 2022. [Online]. Available: <https://arxiv.org/abs/2206.04425>
- [3] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.
- [4] —, "Data efficient and weakly supervised computational pathology on whole slide images," 2020. [Online]. Available: <https://arxiv.org/abs/2004.09666>
- [5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: <https://arxiv.org/abs/2002.05709>
- [6] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "Multi-head attention: Collaborate instead of concatenate," 2021. [Online]. Available: <https://arxiv.org/abs/2006.16362>
- [7] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," 2021. [Online]. Available: <https://arxiv.org/abs/2011.08939>
- [8] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and Y. Zhang, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," 2021. [Online]. Available: <https://arxiv.org/abs/2106.00908>
- [9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2020. [Online]. Available: <https://arxiv.org/abs/1911.05722>
- [10] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," 2021. [Online]. Available: <https://arxiv.org/abs/2011.13971>