Machine Learning in Applications
FP-01 Project Report

Luca Ianniello, Raffaele Martone, Antonio Sirica
s327313, s324807, s326811
Politecnico di Torino

CONTENTS

List of Tables

Machine Learning in Applications FP-01 Project Report

***Abstract*—Malignant pleural mesothelioma represents a rare but aggressive neoplasm that requires accurate diagnosis for the classification of histological subtypes. This study presents an innovative approach based on semi-supervised machine learning techniques for the automatic classification of mesothelioma subtypes (epithelioid, sarcomatoid, and biphasic) using high-resolution digital histological preparation images (WSI). A subset of 22 WSI from the 123 WSI dataset from San Luigi Hospital in Turin was used, implementing two different frameworks, CLAM and DSMIL, with different feature extractors (ResNet50, UNI, UNIv2, Phikon-v2), data augmentation techniques, and bag/instance loss (CE, WCE, Focal, Contrastive). The study demonstrates the effectiveness of semi-supervised learning approaches for mesothelioma classification, highlighting the crucial role of data augmentation in improving model performance and offering a significant contribution to clinical practice and research in the field of digital pathology.**
All the material, experimental results, and detailed data from this project are available at the following repository: https://github.com/LucaIanniello/MLIAProject

I. Introduction

A. Malignant pleural mesothelioma

Malignant pleural mesothelioma is a rare but aggressive tumor that arises from the mesothelial cells lining the pleura, the thin membrane enveloping the lungs. Its epidemiology is closely linked to **asbestos exposure**, with a notable latency period of approximately 40 years between initial exposure and disease onset. The incidence of pleural mesothelioma has shown a gradual increase in countries with historical asbestos use, reflecting industrial trends and regulatory changes. For example, in Japan, deaths from mesothelioma rose from 500 in 1995 to 953 in 2004, paralleling the peak of asbestos imports decades earlier. The disease predominantly affects older adults, with a male-to-female ratio of about 3:1, and the pleura is the site of origin in the vast majority (84%) of cases [1].

The **clinical importance** of pleural mesothelioma lies in its challenging diagnosis, poor prognosis, and significant public health implications. Early symptoms are often nonspecific, such as chest pain or dyspnea, leading to delayed diagnosis and limited treatment options. Accurate pathological identification is crucial not only for guiding therapy but also because a confirmed diagnosis may entitle patients to compensation, especially when asbestos exposure is implicated. However, diagnostic accuracy remains a concern, with an estimated 10–15% of patients receiving an inadequate diagnosis, underscoring the need for improved clinical and pathological collaboration [2].

From a histopathological perspective, **mesothelioma is classified into several subtypes**. The three major histological types are:

- **Epithelioid mesothelioma** (approximately 60% of cases): Characterized by cells resembling normal mesothelial cells, often forming papillary or tubular structures. This subtype generally has a better prognosis compared to others.
- **Sarcomatoid mesothelioma** (about 20%): Composed of spindle-shaped cells mimicking true sarcoma, it is associated with a more aggressive clinical course and poorer outcomes.
- **Biphasic mesothelioma** (about 20%): Contains both epithelioid and sarcomatoid components within the same tumor.

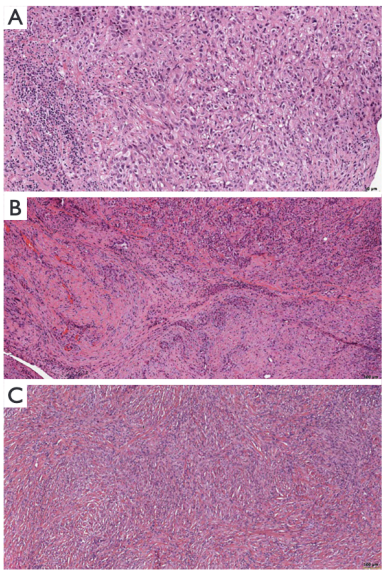


Fig. 1: Histologic presentation of 3 mesothelioma types; epithelioid (A), biphasic (B) and sarcomatoid (C)

Other rare variants include *desmoplastic mesothelioma* (1–2%), lymphohistiocytoid, deciduoid, anaplastic, and well-differentiated papillary mesothelioma. The diversity in histological presentation complicates differential diagnosis, especially in small biopsy samples, and necessitates the use of immunohistochemical markers for accurate classification and distinction from other malignancies such as lung adenocarcinoma or sarcomas [1] [3].

Motivation for research and clinical focus on pleural mesothelioma stems from its increasing incidence, diagnostic challenges, and the need for effective therapeutic strategies. The disease remains a sentinel marker of occupational and environmental health risks associated with asbestos, highlighting the ongoing relevance of preventive measures, early detection, and multidisciplinary management.

B. Potential of Semi-Supervised Learning

1) Theoretical Foundations of Semi-Supervised and Self-Supervised Learning: Before analyzing the specific advantages in the histopathological field, it is essential to clarify the conceptual distinction between semi-supervised and self-supervised learning. Semi-supervised learning represents a machine learning paradigm that strategically combines labeled and unlabeled data to train artificial intelligence models, positioning itself between traditional supervised learning and fully unsupervised learning. This approach is particularly advantageous when collecting labeled data is prohibitively expensive or difficult to obtain, but large amounts of unlabeled data are available [4], [5]. Self-supervised learning, on the other hand, uses the data itself to generate supervisory signals without relying on externally provided labels. In this paradigm, models learn meaningful representations by designing auxiliary tasks (*pretext tasks*) that allow the "ground truth" to be inferred directly from the unlabeled data. This approach more closely mimics the way humans learn to classify objects, developing robust representations through observation of intrinsic structures and relationships within the data [5], [6].

2) Advantages of the Semi-Supervised Approach: Semi-supervised learning represents a strategic solution to address the scarcity of annotated data in the histopathological domain, where the labeling process requires highly specialized expertise and considerable time [4]. [5] conducts an in-depth comparative study between semi-supervised and self-supervised learning methods in the field of computational pathology, highlighting how these approaches can significantly reduce annotation effort while maintaining competitive performance. The research compared three state-of-the-art methods: PAWS as a semi-supervised approach, SimCLR as a contrastive self-supervised method, and SimSiam as a non-contrastive self-supervised method. The results showed that pretraining with semi- and self-supervised methods generally has a positive impact on the performance of histopathological classifiers, particularly evident in scenarios with limited data [5]. However, contrary to expectations, PAWS showed the weakest performance despite explicitly using label information during pretraining, and was also the most sensitive to hyperparameter settings and weight initialization. In contrast, SimSiam demonstrated the best overall performance and the greatest stability when encoder weights are updated during fine-tuning [5], [6].

3) Benefits of WSI and Patch Segmentation: The segmentation of Whole Slide Images (WSI) into patches represents a fundamental computational strategy for the analysis of ultra-high-resolution histopathological images. [5] used patches of size 96×96 pixels to handle gigapixel datasets, demonstrating how this approach enables computational scalability and greater data representativeness [4]. Dividing images into patches allows very large images to be processed with standard hardware resources, facilitating both large-scale training and inference. The study also highlighted that models trained on patches from different WSIs tend to be more robust to variations in staining and sample preparation, improving generalization. However, an important limitation emerged: the features learned from a particular tissue type are transferable in-domain only to a limited extent, suggesting the need for diversified datasets that include various tissue types to develop encoders applicable across the entire histopathological domain [5], [6].

4) Challenges in Whole Slide Image (WSI) Analysis and Supervised Learning: The dataset used for the classification of different mesothelioma cells is structured as Whole Slide Images (WSI). WSIs are extremely large digital images of histological slides, often containing gigapixels of data. [5] Their analysis in digital pathology presents several significant technical and methodological challenges:

- **Data Size and Computational Complexity:** WSIs require substantial computational resources for storage, visualization, and processing. Efficient handling and tiling strategies are necessary to make analysis feasible.
- **Limited and Weak Annotations:** Manual annotation of WSIs is time-consuming and expensive, often resulting in datasets with sparse, weak, or only slide-level (bag-level) labels. This limits the effectiveness of fully supervised learning approaches.
- **Learning from Weak Labels:** Since detailed, instance-level (patch-level) annotations are rarely available, models must learn from weak supervision, where only global labels are provided. This motivates the use of Multiple Instance Learning (MIL) and related approaches.
- **Class Imbalance:** Medical datasets are frequently imbalanced, with some classes (e.g., rare variations) underrepresented. This can bias model training and reduce performance on minority classes.

These challenges drive the development and adoption of advanced machine learning techniques such as Multiple Instance Learning (MIL), semi-supervised learning, and data augmentation strategies to improve model robustness and generalization in WSI analysis.

The standard approach to analyzing WSIs is based on supervised learning, with fragmentation of the WSI into patches. Supervised learning involves training a model on a labeled dataset, where each input (e.g., a patch from a WSI) is associated with a corresponding label (e.g., presence or absence of a specific cell type) [5]. The model learns to map inputs to outputs based on these labels, allowing it to make predictions on new, unseen data. However, this approach is often limited by the availability of labeled data, which can be scarce and expensive to obtain in the medical domain. In addition, obtaining a large number of samples for a specific class can be challenging if that class represents a rare pathology, leading to class imbalance. As a result, many researchers are exploring alternative methods such as weakly supervised learning, self-supervised learning, and semi-supervised learning to leverage unlabeled data effectively.

II. Background

A. Computer Vision in Histopathology & Multiple Instance Learning in Medical Imaging

The evolution of computational analysis in histopathology has undergone a radical transformation in recent years, shifting from traditional microscopy-based analysis to advanced digital systems that integrate artificial intelligence and deep learning.

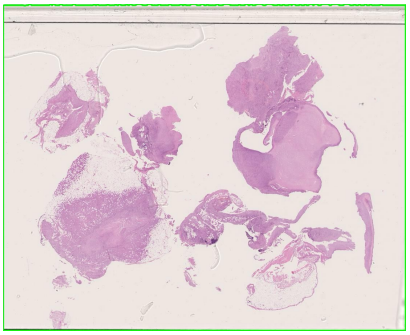


Fig. 2: Whole Slide Image Sample

The integration of computer vision into pathology through slide digitization represents a transformative leap in the field's evolution, offering consistent, reproducible, and objective results with ever-increasing speed and scalability [4], [5].

Deep learning techniques for WSI (Whole Slide Image) analysis face unique computational challenges due to the gigapixel size of these images, which can range from 100 million to 10 billion pixels. The most effective approaches do not use the entire image as input, but instead extract and utilize only a small number of patches, typically ranging in size from 32×32 to 10,000×10,000 pixels, with the majority of approaches using patches of approximately 256×256 pixels. This approach to reducing the high dimensionality of WSIs can be seen as a form of human-guided feature selection [4], [6].

The principles of Multiple Instance Learning (MIL) are based on the paradigm that each WSI is treated as a *bag* and the patches extracted from it as *instances* of the bag. In a positive context (bag), there is at least one positive instance, while in a negative context (bag), all instances are negative. This framework is particularly effective in the domain of digital pathology, as labels for whole slide images are often routinely captured, whereas labels for patches, regions, or pixels are not. The most recent MIL methods account for both global and local dependencies among instances, using attention mechanisms and transformer architectures to aggregate patch features.

Semi-supervised approaches in medical imaging have proven particularly effective in addressing the scarcity of labeled data in the medical domain. Semi-supervised learning can help by providing a strategy to pre-train a neural network with unlabeled data, followed by fine-tuning for a downstream task with limited annotations. Self-supervised and contrastive learning represent especially promising approaches for extracting meaningful representations from vast histological archives [5].

Recent foundation models such as UNI and CONCH have marked a significant breakthrough in the field. UNI and CONCH are notable as the first foundation models trained on diverse internal pathology datasets spanning infectious, inflammatory, and neoplastic diseases, and made openly accessible to the research community. CONCH, a visual-language foundation model developed using various sources of histopathological images, biomedical text, and over 1.17 million image-caption pairs, has demonstrated state-of-the-art performance on 14 different benchmarks.

However, significant gaps remain in the current literature. The main challenges include the lack of labeled data, pervasive variability across tissues and staining, the non-Boolean nature of diagnostic tasks, and the need for regulatory approval [4]. MIL-based methods show effectiveness for histopathological classification and segmentation but require improvements for instance-level variability and small region recognition, often necessitating additional supervision constraints or being prone to overfitting [5]. The lack of standardized evaluation benchmarks and prospective validation protocols represents a significant limitation for clinical adoption. Furthermore, the challenge of interpretability and transparency of algorithms remains, which is essential for gaining the trust of pathologists and facilitating integration into clinical workflows [4].

B. CLAM: Clustering-constrained Attention Multiple-instance Learning

Deep learning has greatly advanced computational pathology, especially for whole-slide image (WSI) analysis. Traditional methods often require exhaustive pixel-level annotations or assume that slide-level labels apply uniformly to all regions, which limits scalability and accuracy. Multiple-instance learning (MIL) offers a weakly supervised alternative by using slide-level labels, but classical MIL approaches (e.g., max-pooling) only leverage the most salient patch per slide, underutilizing the available data.

To address these limitations, [7] introduced the CLAM (Clustering-constrained Attention Multiple-instance learning) model. CLAM uses attention-based pooling to aggregate patch-level features into slide-level representations, automatically identifying diagnostically relevant regions. It also incorporates instance-level clustering to refine the feature space and improve classification. During training and inference, CLAM assigns an attention score to each patch, indicating its importance for the slide-level prediction. The model uses multiple parallel attention branches to generate class-specific slide representations, which are then classified to produce the final slide-level predictions. This approach enables accurate WSI classification without detailed region annotations and improves interpretability by highlighting the most relevant tissue regions.

C. Trident for Segmentation and Feature Extraction

For the processing of histopathological images and feature extraction, Trident [8] was employed—a Python package specifically designed for handling Whole Slide Images (WSI) using pretrained foundation models. Trident implements a robust tissue-background segmentation pipeline based on DeepLabV3 pretrained on the COCO dataset, which surpasses the limitations of traditional Otsu thresholding or binary thresholding methods, ensuring better generalization beyond H&E staining and more effective separation of tissue from noise and artifacts.

The processing workflow initially involved automatic tissue segmentation to remove background regions and minimize unnecessary processing, followed by the subdivision of tissue

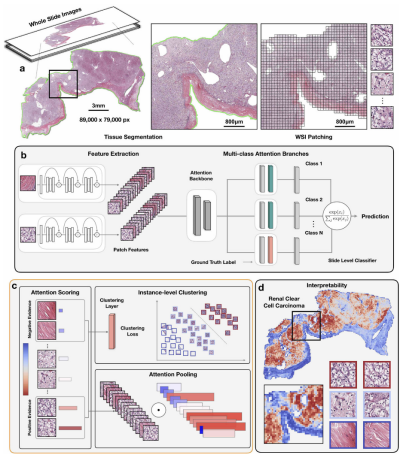


Fig. 3: Overview of the CLAM conceptual framework, architecture and interpretability.

areas into individual patches of specified size. For feature extraction, Trident provides unified model factories that allow for easy loading of various pretrained patch encoders. In this study, four distinct models were used: ResNet50-ImageNet as a traditional baseline, UNI and UNIv2 as general foundation models for computational pathology, and Phikon-v2 as a model specialized for biomarker prediction. The modular architecture of Trident enabled the standardization of the inference process across all models, facilitating direct performance comparison and ensuring reproducibility in the extraction of patch-level representations for subsequent downstream analyses.

D. DSMIL

DSMIL builds upon weakly supervised multiple instance learning frameworks like CLAM but introduces key innovations to improve classification in whole-slide image (WSI) analysis. Its core contribution is a dual-stream architecture that explicitly separates instance-level and bag-level reasoning. One stream identifies the most critical patch via a max-pooling mechanism, while the other computes attention weights by measuring the similarity between each patch and the critical instance using a trainable non-local operation. This allows the model to softly aggregate relevant patches based on learned contextual similarity, rather than relying on a single instance.

To enhance feature quality, DSMIL integrates self-supervised contrastive learning (SimCLR), which helps generate robust and discriminative patch embeddings without requiring detailed annotations. Additionally, it employs a pyramidal fusion mechanism to combine features across multiple magnifications, capturing both coarse and fine tissue structures—mirroring how pathologists interpret slides.

Compared to CLAM, DSMIL simplifies the architecture by removing clustering and pseudo-labeling, while improving performance through its two-stream design and multiscale integration. These innovations make DSMIL both effective and easy to train in weakly supervised settings, particularly in the context of computational pathology.

III. Materials and Methods

A. Dataset

The **"Mesothelioma San Luigi"** dataset consists of whole-slide images (WSIs) obtained from real patients diagnosed with malignant pleural mesothelioma. This dataset is strictly confidential and was granted by San Luigi Hospital to Politecnico di Torino for research purposes; therefore, it is not publicly available.

It includes cases from all three major histological subtypes of mesothelioma: epithelioid, sarcomatoid, and biphasic.

The dataset comprises 123 cases, each associated with a unique patient ID, histologic number, provenance (origin of the sample), and a confirmed diagnosis indicating the specific histological subtype (epithelioid, biphasic, or sarcomatoid).

Among these examples, epithelioid mesothelioma is the most represented, with 96 cases, confirming its prevalence in clinical populations. Biphasic mesothelioma accounts for 22 cases, while the sarcomatoid subtype is the least frequent, with only 5 instances recorded.

This distribution mirrors known clinical patterns, where epithelioid mesothelioma is not only the most common form but also tends to have a better prognosis and a more favorable response to treatment. Biphasic mesothelioma, which combines features of both epithelioid and sarcomatoid cell types, is observed less frequently and typically exhibits more aggressive behavior. Sarcomatoid mesothelioma, the rarest and most aggressive subtype, presents the greatest diagnostic and therapeutic challenges and is correspondingly underrepresented in the dataset.

The dataset thus exhibits a degree of **class imbalance**, with a substantial majority of cases belonging to the epithelioid class. This imbalance should be carefully considered in any computational modeling effort. Without adequate adjustment models trained on this dataset may become biased toward the majority class and fail to generalize well to underrepresented subtypes, particularly sarcomatoid cases.

Despite this imbalance, the dataset remains clinically meaningful. Its distribution of histological subtypes reflects the epidemiology of malignant pleural mesothelioma and provides a realistic foundation for developing and validating machine learning models intended for use in diagnostic pathology or digital slide analysis. Nevertheless, strategies to mitigate bias and ensure robust performance across all subtypes will be essential when applying this dataset to supervised learning tasks.

To conduct our analysis while addressing time constraints related to the execution and training of all implemented methodologies, we adopted a strategy based on selecting a limited subset of whole-slide images (WSIs). Specifically, we selected 9 WSIs for the biphasic subtype, 8 for the epithelioid subtype, and 5 for the sarcomatoid subtype. This selection aimed to preserve, as closely as possible, the original distribution of histological labels within the dataset, ensuring that the reduced sample remains representative of the broader cohort.

B. Preprocessing

All the implemented methodologies operate not on the raw whole-slide images (WSIs) themselves, but on smaller patches extracted from them using various techniques. In the context of WSIs, patches are defined as manageable, smaller sections or tiles derived from these extremely large images. This strategy significantly reduces the computational load and enables models to concentrate on specific histological features that are essential for accurate diagnosis or classification. Patch extraction involves dividing the slide into a grid of tiles and selecting only those that contain meaningful tissue, while excluding areas that consist of background or irrelevant content.

To work with CLAM, various feature extractors have been employed, each taking the extracted image patches as input. To balance efficiency and quality, a patch size of 256×256 pixels was used for most extractors. The only exception is PhikonV2, which uses a patch size of 224×224 pixels.

For the DSMIL methodology, patches of size 256 by 256 pixels are extracted from each whole-slide image (WSI). These patches are obtained at two different magnification levels to capture varying levels of detail. The first magnification level, referred to as level 0, corresponds to a 10x magnification, which serves as the base magnification for the analysis. The second magnification level, level 1, is set at 5x magnification, which is effectively half the base magnification, achieved by dividing the base magnification of 10x by 2. This multi-scale approach allows the model to incorporate both fine and coarser tissue features during the learning process.

C. CLAM

CLAM (Clustering-constrained Attention Multiple-instance learning) is a deep learning framework designed for efficient and interpretable analysis of Whole Slide Images (WSIs) using only slide-level labels, without requiring region-of-interest (ROI) extraction or patch-level annotations. The model is particularly suited for multi-class subtyping problems in computational pathology.

1) Architecture Overview:

- **Patch Extraction and Feature Representation:** WSIs are first segmented and divided into smaller patches. Instead of saving all image patches, CLAM saves only the coordinates and extracts features from each patch using a pretrained encoder (e.g., ResNet50, UNI, or CONCH). This results in a set of feature vectors representing each slide. The patches can be extracted using customizable parameters, such as patch size and stride, to balance computational efficiency and feature richness. The feature extractor is a ResNet50-based encoder, which is pretrained on ImageNet to capture rich visual features.
- **Multiple Instance Learning (MIL):** The slide is treated as a bag of instances (patches), and only the slide-level label is used for supervision. CLAM leverages an attention-based pooling mechanism to aggregate patch-level features into a slide-level representation.
- **Attention Mechanism:** Each patch receives an attention score, indicating its importance for the slide-level prediction. The attention module enables the model to focus on diagnostically relevant regions, improving both accuracy and interpretability.
- **Clustering Constraint:** To further refine the feature space, CLAM introduces an instance-level clustering constraint. Highly attended patches are clustered, and the model is encouraged to separate positive and negative evidence within each slide. This helps the model distinguish between different tissue types or diagnostic patterns.
- **Multi-Branch Attention:** For multi-class problems, CLAM uses multiple parallel attention branches, each specializing in identifying evidence for a specific class. Each branch produces a class-specific slide representation, which is then classified to obtain the final prediction.

2) Loss Functions:

- **Bag Loss:** The primary supervision comes from the slide-level (bag-level) label. The CLAM project presents only SVM and Cross-Entropy (CE) as possible bag loss.
- **Instance Loss:** To enforce the clustering constraint, an additional instance-level loss is applied to the most and least attended patches, encouraging the model to separate positive and negative evidence within each slide. In details, the project uses SVM loss. This loss is estimates by taking the patches with the highest attention scores and applying a loss function to these patches. This helps the model to learn the characteristics that distinguish between different classes, even when only slide-level labels are available.

3) Training and Workflow:

- **Segmentation and Patching:** Tissue regions are segmented, and patches are extracted using customizable parameters. The process is efficient, as only patch coordinates are saved and features are extracted on-the-fly.
- **Feature Extraction:** Features are extracted from each patch using a pretrained encoder. The resulting feature files are used for downstream training.
- **Dataset Preparation:** Datasets are organized with slide-level labels and split into training, validation, and test sets, ensuring that slides from the same patient do not appear in multiple splits.
- **Training:** The model is trained using the attention-based MIL framework, optimizing both bag-level and instance-level losses. Training can be performed for binary or multi-class classification tasks, with support for cross-validation and hyperparameter tuning.
- **Evaluation and Visualization:** Trained models can be evaluated on independent test sets. CLAM also provides heatmap visualizations, highlighting the most relevant regions for each prediction, aiding interpretability.

4) Key Advantages:

- Data-efficient: Requires only slide-level labels, reducing annotation burden.
- Interpretable: Attention scores and heatmaps provide insight into model decisions.
- Flexible: Supports multiple pretrained encoders and is adaptable to various WSI datasets and classification tasks.

- Scalable: Efficient patching and feature extraction pipelines enable high-throughput analysis.

In summary, CLAM combines attention-based MIL with clustering constraints to enable accurate, interpretable, and data-efficient WSI classification, making it a powerful tool for computational pathology and semi-supervised learning scenarios.

D. Training Strategy: Loss function

The training strategy for CLAM involves several key components and considerations to ensure effective learning from weakly labeled data. The model is trained using a combination of bag-level and instance-level losses, which are crucial for optimizing the attention mechanism and clustering constraints. The bag loss is used to supervise the overall slide-level classification, while the instance loss focuses on the most and least attended patches to enforce clustering constraints. This dual-loss approach helps the model learn to differentiate between positive and negative evidence within each slide, enhancing its ability to classify WSIs accurately.

In details, the instance loss is defined as a support vector machine (SVM) loss, which is applied to the patches with the highest attention scores. This loss function encourages the model to learn the characteristics that distinguish between different classes, even when only slide-level labels are available. The training process involves iterating over the dataset, updating the model parameters based on the computed losses, and optimizing the attention scores to highlight diagnostically relevant regions.

The bag loss is implemented using cross-entropy loss, which is suitable for multi-class classification tasks, and a SVM loss. The model is trained to minimize the difference between the predicted slide-level probabilities and the true labels, ensuring that the attention mechanism focuses on the most informative patches. However, we have tested different losses functions, including Contrastive Loss, Focal Loss, and Weighted Cross Entropy (WCE), to evaluate their impact on model performance. Each loss function has its own strengths and weaknesses, and the choice of loss can significantly affect the model's ability to generalize across different datasets and tasks.

The focal loss is particularly useful in scenarios with class imbalance, as it down-weights the loss contribution from well-classified examples, allowing the model to focus more on hard-to-classify instances. Weighted Cross Entropy (WCE) is another approach that adjusts the loss function to account for class imbalance by assigning different weights to different classes based on their prevalence in the dataset. Both these losses are helpful in our project due to the class imbalance present in the dataset, where some classes are underrepresented compared to others.

We have implemented a supervised contrastive loss function, which encourages the model to learn feature representations where samples belonging to the same class are close together in the embedding space, while samples from different classes are pushed apart. This loss is particularly effective in scenarios where the model needs to learn discriminative features from weakly labeled data. Our implementation is inspired by the supervised contrastive learning framework [6]. For a batch of feature vectors, we compute the pairwise cosine similarities, scaled by a temperature parameter. For each anchor patch, the loss considers all other patches with the same label as positives and all others as negatives, excluding self-comparisons. The objective is to maximize the log-probability that an anchor is similar to its positive pairs, while minimizing similarity to negatives. The choice of the patches to be used is done considering the 1000 patches with the highest attention scores. This approach leverages label information to structure the feature space, making it easier for the model to distinguish between different tissue types or diagnostic categories.

E. DSMIL

1) Architecture Overview: DSMIL introduces a dual-stream architecture that explicitly models both instance-level and bag-level information.

The first stream operates in a manner similar to traditional max-pooling, identifying the "critical instance"—the patch with the highest predicted relevance to the bag label. The second stream measures the similarity between each instance and the critical instance using a trainable, non-local operation: for each patch, the model generates two vectors: a "query" and an "information" vector. The similarity between each patch and the critical instance is computed by taking the inner product (dot product) of their query vectors, passing this through a softmax function so that all similarities in the slide sum to one. The weights resulting from this process determine how much each patch should contribute to the overall slide representation, based on its learned similarity to the critical instance.

This mechanism assigns higher weights to patches that are more similar to the critical instance, allowing the model to softly select and aggregate relevant patches rather than relying on a single instance. By averaging the outputs of these two streams, DSMIL achieves a more nuanced and robust aggregation of information, leading to improved classification performance.

2) Self-supervised Contrastive Learning feature extractor: A further innovation in DSMIL is its use of self-supervised contrastive learning for feature extraction using SimCLR as embedder [9]. Recognizing the difficulty of learning effective patch representations under weak supervision, DSMIL leverages contrastive pre-training to produce robust and discriminative embeddings for each patch. This approach enables the model to better capture the variability and complexity of tissue morphology, even in the absence of detailed annotations.

3) Pyramidal fusion mechanism: DSMIL also incorporates a pyramidal fusion mechanism, which combines features extracted at multiple magnification levels. This multiscale approach reflects the way pathologists assess tissue slides, considering both coarse and fine-grained structures. By fusing information from different scales, DSMIL is able to leverage contextual cues that are critical for accurate classification and localization.

4) Comparison with CLAM: DSMIL builds directly on the foundational ideas introduced by CLAM, retaining core aspects such as the use of patch-level features extracted through a pretrained convolutional neural network, attention-based pooling mechanisms for aggregating patch information, and a weakly supervised learning framework that relies solely on slide-level labels. These elements allow both models to function effectively in settings where fine-grained annotations are unavailable, a common scenario in digital pathology.

However, DSMIL introduces several meaningful innovations that distinguish it from CLAM. One of the most significant changes is its dual-stream architecture, which separates the learning of instance-level and bag-level features. While CLAM focuses primarily on attention-based pooling, DSMIL incorporates both a max-pooling stream to identify the most discriminative patch and an attention-based stream to capture contextual slide-level information. This design allows DSMIL to leverage both localized and global features, improving its robustness across varied cases.

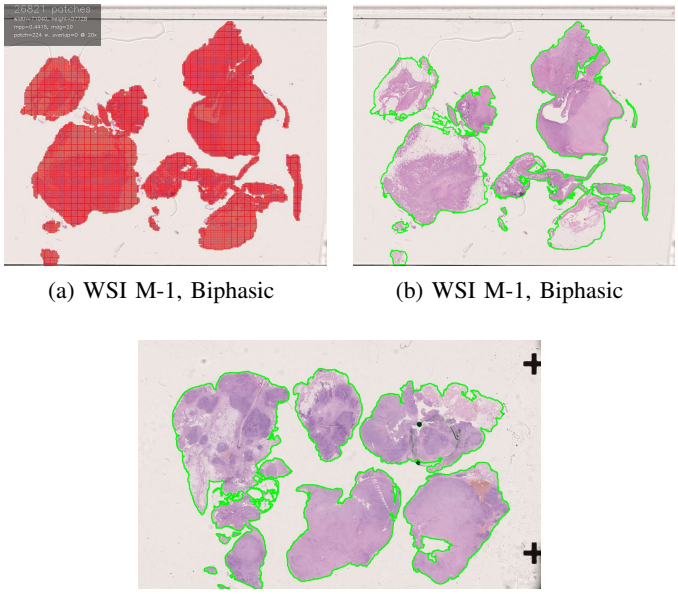
Another important distinction lies in how DSMIL fuses these two streams to produce final predictions. By integrating information from both pathways, it mitigates the limitations of relying solely on either highly localized or entirely global representations, leading to more reliable performance in challenging slides. Additionally, DSMIL skips the clustering and pseudo-labeling mechanisms used in CLAM, which not only simplifies the model architecture but also makes it easier to train and tune without sacrificing interpretability. This streamlined approach enhances efficiency while maintaining strong performance, making DSMIL a more accessible and scalable solution in weakly supervised WSI classification tasks.

F. Segmentation and Feature Extraction

1) Segmentation and Feature Extraction Architecture: For feature extraction from histopathological images, a processing pipeline based on the **CLAM** and **Trident** frameworks was implemented. The feature extraction process is a fundamental step to convert high-resolution patches into compact and informative numerical representations, significantly reducing the computational dimensionality required for Whole Slide Image (WSI) analysis [7].

The extraction pipeline initially performs **automatic tissue segmentation** to remove background regions and minimize unnecessary processing, followed by the subdivision of tissue areas into individual patches of specified size (see Fig. 4). CLAM uses a segmentation approach based on thresholding in the HSV color space, while Trident implements a more robust segmentation based on **DeepLabV3 pretrained on the COCO dataset**, which surpasses the limitations of traditional Otsu thresholding methods, ensuring better generalization beyond H&E staining [7], [8].

For **feature extraction**, four different encoder models were used: **ResNet50-ImageNet** as a traditional baseline, **UNI** and **UNIv2** as general foundation models for computational pathology, and **Phikon-v2** as a model specialized for biomarker prediction. ResNet50 pretrained on ImageNet was employed via the CLAM framework to convert 256×256 pixel patches into 1024-dimensional feature representations using adaptive mean-spatial pooling after the third residual block of the network [7]. Trident provides unified model factories that allow easy loading of various pretrained patch encoders, facilitating direct performance comparison and ensuring reproducibility in patch-level feature extraction [8].



(a) WSI M-1, Biphasic (b) WSI M-1, Biphasic

(c) WSI M-108, Biphasic

Fig. 4: Blockmaps showing segmentation and separation in patches over different WSIs.

2) Feature-Level Data Augmentation: Given the gigapixel nature of WSIs and the resulting computational limitations in applying traditional data augmentation techniques directly to whole images, a strategy of **feature space augmentation** was adopted. This approach avoids repeated feature extraction and enables online augmentation during MIL model training, overcoming the limitations of offline image-level augmentation [10], [11].

3) Extrapolation Methods: Among the implemented feature space augmentation techniques, particular attention was given to **extrapolation methods** between samples in feature space. As demonstrated by DeVries and Taylor [10], extrapolation between samples in feature space can be used to augment datasets and improve the performance of supervised learning algorithms. The extrapolation process generates new feature vectors according to the formula:

$$\mathbf{c}' = (\mathbf{c}_j - \mathbf{c}_k)\lambda + \mathbf{c}_j$$

where $\mathbf{c}'$ is the synthetic context vector, $\mathbf{c}_i$ and $\mathbf{c}_j$ are neighboring context vectors, and λ is a variable in the range $\{0, \infty\}$ controlling the degree of extrapolation. Extrapolation is particularly advantageous for generating samples with greater variability than those already common in the dataset, an essential characteristic for improving model robustness when training data is limited [10].

4) Diffusion Model-Based Augmentation Framework: Inspired by the **AugDiff** framework, a feature augmentation strategy based on **Variational Autoencoder (VAE)**, **U-Net**,

and **diffusion models** was implemented to generate new feature-level samples from extracted features [11]. This approach leverages the generative diversity of diffusion models to improve the quality of feature augmentation and the step-by-step generation property to control the preservation of semantic information.

The implemented framework involves a **Denoising AutoEncoder (DAE) training process** that includes adding noise in the diffusion process and predicting the noise by the DAE. To preserve the original semantic information, the sampling process is divided into two phases: **K-step Diffusion** and **K-step Denoising**. In the first phase, a K-step diffusion process is applied to the original features, where K is less than the total number of steps T. In the second phase, the trained DAE is used to denoise the input features for K steps, generating augmented versions of the original features that retain the fundamental semantic characteristics [11].

This diffusion model-based augmentation methodology has proven, in principle, to be superior to traditional methods such as Mixup, which often produce unrealistic features, offering a more efficient and effective framework for MIL training with online generation capabilities and fine control over semantic preservation [11].

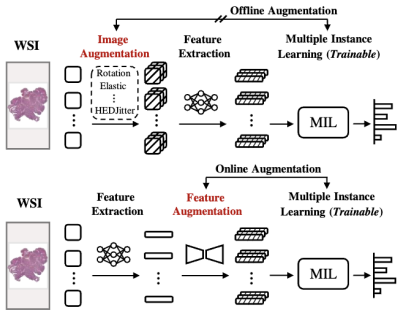


Fig. 5: Overview of the AugDiff conceptual framework.

G. Materials

All implementations are publicly available in our GitHub repository.

https://github.com/LucaIanniello/MLIAProject

The experiments for CLAM and DISMIL models are organized into clearly structured folders to ensure reproducibility and ease of use.

For the CLAM experiments, the Script folder contains all necessary Jupyter notebooks and .sh scripts. These scripts are designed to support a wide range of configurations, including various loss functions, feature extractors, and augmentation techniques. Users can easily specify parameters such as the choice of model, training loss, and data augmentation strategy directly within the scripts. This modular and flexible design allows for straightforward adaptation to different datasets and classification tasks.

For the DSMIL experiments, the corresponding DSMIL folder includes both shell scripts and notebooks required to execute the full experimental pipeline. Similar to the CLAM setup, these scripts are highly configurable, allowing users to modify experimental parameters and settings to explore different configurations and validate performance under various conditions.

This flexible scripting infrastructure enables reproducible and extensible experimentation across multiple weakly supervised multiple instance learning frameworks in computational pathology.

IV. Results and Discussion

A. Implementation Set-up Details

1) Hardware specifications: Most of the experiments were conducted on the Legion Server of the HPC cluster at Politecnico di Torino, equipped with an NVIDIA Tesla V100 SXM2 GPU featuring 32 GB of memory and 512 CUDA cores.

The use of this high-performance machine was essential due to the computational demands imposed by the large number of patches extracted from high-resolution whole slide images (WSIs), as well as the need to run multiple jobs in parallel efficiently.

In addition to the HPC cluster, Google Colab was also utilized, leveraging its NVIDIA T4 GPU to support experiments. This provided a flexible and accessible environment for running smaller-scale or exploratory tasks, complementing the resources available on the Legion Server.

2) Train/validation/test splits: For all the implementation we adopted a single-split strategy to partition the dataset into training, validation, and test sets, with test fractions of 0.2 (CLAM) and 0.3 (DSMIL). Given the limited size of the available data, this approach was selected to maximize the amount of information available for training while still preserving separate subsets for model selection and unbiased performance evaluation.

A single split was preferred over k-fold cross-validation primarily due to the constraints imposed by the dataset size. In scenarios with a small number of samples, dividing the data into multiple folds can lead to high variance across runs and insufficient data per fold, ultimately affecting the stability and learning capacity of the model. By contrast, a well-constructed single split offers a fixed, consistent evaluation protocol and allows a larger portion of the dataset to be used for learning, which is particularly valuable when working with whole slide images that are computationally expensive to process.

3) Chosen evaluation metrics: For model evaluation, we chose to report both accuracy and AUC (Area Under the ROC Curve), as they offer complementary insights into performance—particularly important in medical image classification tasks.

Accuracy provides a straightforward measure of how often the model's predictions match the ground truth labels. It is intuitive and useful when class distributions are balanced, as it reflects the overall proportion of correct predictions. However, accuracy alone can be misleading in imbalanced datasets, where a model might perform well by favoring the majority class.

To address this, we also report AUC, which evaluates the model's ability to distinguish between classes across different

decision thresholds. AUC is particularly valuable in clinical contexts because it captures the trade-off between sensitivity and specificity, independent of a fixed threshold. This makes it more robust in imbalanced settings and better suited to assess how confidently the model separates positive and negative cases.

By combining both metrics, we gain a more nuanced understanding of the model's reliability and discriminative power, which is essential for assessing its potential utility in real-world diagnostic scenarios.

B. Feature Extractors results

The following tables summarize the performance of various feature extractors, CLAM, ResNet50 Trident, Univ1 Trident, Univ2 Trident, and Phikon Trident, evaluated across different loss functions, including Cross Entropy, Contrastive, Focal, and Weighted Cross Entropy. Metrics reported include Test and Validation AUC, as well as Accuracy, providing a comparative overview of how each feature extractor performs under different training objectives.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	CE	SVM	0.91	0.61	0.4	0.6
ResNet50 Trident	CE	SVM	0.44	0.44	0.4	0.2
Univ1 Trident	CE	SVM	0.83	0.27	0.6	0.4
Univ2 Trident	CE	SVM	0.55	0.77	0.4	0.6
Phikon Trident	CE	SVM	0.66	0.36	0.6	0.4

TABLE I: Performance between different feature extractors, considering Cross Entropy Loss

Table I summarizes the performance of different feature extractors when trained with Cross Entropy as the bag loss and SVM as the instance loss. The CLAM Feature Extractor achieves the highest test AUC (0.91), indicating strong discriminative ability on the test set, and also shows a balanced validation AUC (0.61). Univ1 Trident and Phikon Trident both reach the highest test accuracy (0.6), but their AUC values are lower than those of the CLAM Feature Extractor, suggesting that while they can correctly classify a similar proportion of samples, their overall ranking ability is less robust. ResNet50 Trident and Univ2 Trident display lower test AUC and accuracy, highlighting the importance of feature extractor choice in this context. Overall, these results demonstrate that the CLAM Feature Extractor is particularly effective for this task, but some alternative extractors can achieve comparable accuracy, albeit with reduced AUC. This comparison underscores the impact of feature representation on WSI classification performance.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	Contr	SVM	0.91	0.66	0.4	0.4
ResNet50 Trident	Contr	SVM	0.58	0.44	0.4	0.4
Univ1 Trident	Contr	SVM	0.47	0.30	0.6	0.2
Univ2 Trident	Contr	SVM	0.41	0.77	0.4	0.4
Phikon Trident	Contr	SVM	0.44	0.33	0.6	0.4

TABLE II: Performance between different feature extractors, considering Contrastive Loss

Table II reports the results obtained by the different feature extractors when using the supervised contrastive loss as the bag loss and SVM as the instance loss. The CLAM Feature Extractor again achieves the highest test AUC (0.91) and a strong validation AUC (0.66), confirming its effectiveness in extracting discriminative features even under the contrastive loss regime. Univ1 Trident and Phikon Trident reach the highest test accuracy (0.6), but their AUC values are considerably lower, indicating that their predictions are less reliable in ranking positive and negative cases. ResNet50 Trident and Univ2 Trident show moderate performance, with lower AUC and accuracy values compared to CLAM. Notably, the validation AUC for Univ2 Trident is the highest among all extractors (0.77), suggesting some generalization ability on the validation set, but this does not translate into higher test accuracy. Overall, these results highlight the robustness of the CLAM Feature Extractor across different loss functions, while also illustrating that the choice of feature extractor and loss function can significantly impact model performance in WSI classification tasks.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	Focal	SVM	0.83	0.77	0.4	0.6
ResNet50 Trident	Focal	SVM	0.44	0.44	0.4	0.2
Univ1 Trident	Focal	SVM	0.77	0.41	0.6	0.4
Univ2 Trident	Focal	SVM	0.55	0.63	0.4	0.6
Phikon Trident	Focal	SVM	0.72	0.36	0.6	0.2

TABLE III: Performance between different feature extractors, considering Focal Loss

Table III shows the results for the different feature extractors when using Focal Loss as the bag loss and SVM as the instance loss. The CLAM Feature Extractor achieves the highest validation AUC (0.77), indicating good generalization to the validation set, and maintains a balanced test AUC (0.83). Univ1 Trident and Phikon Trident reach the highest test accuracy (0.6), but their AUC values are lower, suggesting that their predictions are less reliable in ranking positive and negative cases. ResNet50 Trident and Univ2 Trident display lower AUC and accuracy values, highlighting the challenges these extractors face under class imbalance, even with Focal Loss. Overall, these results confirm the robustness of the CLAM Feature Extractor, while also demonstrating that Focal Loss can help mitigate class imbalance, but the choice of feature extractor remains a critical factor in achieving optimal performance in WSI classification.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	WCE	SVM	0.91	0.61	0.4	0.6
ResNet50 Trident	WCE	SVM	0.44	0.44	0.4	0.2
Univ1 Trident	WCE	SVM	0.83	0.27	0.6	0.4
Univ2 Trident	WCE	SVM	0.55	0.77	0.4	0.6
Phikon Trident	WCE	SVM	0.66	0.36	0.6	0.4

TABLE IV: Performance between different feature extractors, considering Weighted Cross Entropy (WCE) Loss

Table IV presents the results for the different feature extractors when using Weighted Cross Entropy (WCE) Loss as the bag loss and SVM as the instance loss. The CLAM Feature Extractor achieves the highest test AUC (0.91) and a balanced validation AUC (0.61), confirming its strong performance even when class imbalance is addressed through WCE. Univ1 Trident and Phikon Trident reach the highest test accuracy (0.6), but their AUC values are lower, indicating that their

ability to rank positive and negative cases is less robust compared to CLAM. ResNet50 Trident and Univ2 Trident show lower AUC and accuracy values, suggesting that WCE alone does not fully compensate for the limitations of these feature extractors. Overall, these results reinforce the effectiveness of the CLAM Feature Extractor and highlight that, while WCE can help mitigate class imbalance, the choice of feature extractor remains a key determinant of classification performance in WSI analysis.

Across all four loss functions, Cross Entropy, Contrastive, Focal, and Weighted Cross Entropy (WCE), the CLAM Feature Extractor consistently achieves the highest or among the highest test AUC values, demonstrating robust discriminative capability regardless of the loss function employed. While Univ1 Trident and Phikon Trident occasionally match or exceed CLAM in terms of test accuracy, their AUC values are generally lower, indicating less reliable ranking of positive and negative cases. ResNet50 Trident and Univ2 Trident typically show lower performance across both AUC and accuracy metrics, highlighting the importance of feature extractor choice. The impact of the loss function is evident in the variation of validation AUC and accuracy, but the overall ranking of feature extractors remains stable, with CLAM outperforming alternatives in most scenarios. These results underscore that, although loss function selection can influence generalization and class imbalance handling, the choice of feature extractor is the primary determinant of WSI classification performance in this study.

C. Multidimensionality Reduction

We have tested the performance of different feature extractors using PCA (Principal Component Analysis) to reduce dimensionality, while applying various loss functions. The tables below summarize the results for each loss function across different feature extractors, including CLAM Feature Extractor, ResNet50 Trident, Univ1 Trident, Univ2 Trident, and Phikon Trident.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	CE	SVM	0.69	0.66	0.6	0.6
ResNet50 Trident	CE	SVM	0.58	0.58	0.4	0.4
Univ1 Trident	CE	SVM	0.69	0.50	0.6	0.2
Univ2 Trident	CE	SVM	0.77	0.88	0.6	0.6
Phikon Trident	CE	SVM	0.50	0.58	0.2	0.2

TABLE V: Performance between different feature extractors with Cross Entropy Loss and PCA

Table V presents the performance of different feature extractors with Cross Entropy Loss after applying PCA for dimensionality reduction. Univ2 Trident achieves the highest test AUC (0.77) and validation AUC (0.88), indicating strong discriminative ability and generalization. The CLAM Feature Extractor and Univ2 Trident both reach the highest test accuracy (0.6), with CLAM also showing balanced validation accuracy (0.6). Univ1 Trident matches the top test accuracy but has lower validation accuracy (0.2), suggesting less stable generalization. ResNet50 Trident and Phikon Trident display lower performance across both AUC and accuracy metrics, highlighting the importance of feature extractor choice. Overall, these results suggest that, when using PCA and Cross Entropy Loss, Univ2 Trident and CLAM Feature Extractor provide the most robust performance, while other extractors are less effective in this setting.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	Contrastive	SVM	0.83	0.36	0.6	0.4
ResNet50 Trident	Contrastive	SVM	0.50	0.70	0.6	0.4
Univ1 Trident	Contrastive	SVM	0.69	0.50	0.6	0.4
Univ2 Trident	Contrastive	SVM	0.83	0.75	0.4	0.6
Phikon Trident	Contrastive	SVM	0.72	0.58	0.6	0.2

TABLE VI: Performance between different feature extractors with Contrastive Loss and PCA

Table VI presents the results of different feature extractors using Contrastive Loss with PCA. The CLAM Feature Extractor and Univ2 Trident both achieve the highest test AUC (0.83), but Univ2 Trident stands out with the highest validation AUC (0.75) and validation accuracy (0.6), indicating better generalization to unseen data. CLAM, ResNet50 Trident, Univ1 Trident, and Phikon Trident all reach the highest test accuracy (0.6), but only CLAM and Univ1 Trident maintain moderate validation accuracy (0.4). Phikon Trident, despite matching the top test accuracy, has the lowest validation accuracy (0.2), suggesting less stable performance. Overall, Univ2 Trident demonstrates the best balance between test and validation metrics, while CLAM and Univ1 Trident provide competitive but less consistent results. These findings highlight the importance of considering both test and validation performance when evaluating feature extractors with Contrastive Loss and PCA.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	Focal	SVM	0.83	0.44	0.6	0.4
ResNet50 Trident	Focal	SVM	0.58	0.77	0.6	0.4
Univ1 Trident	Focal	SVM	0.69	0.50	0.6	0.4
Univ2 Trident	Focal	SVM	0.88	0.75	0.6	0.8
Phikon Trident	Focal	SVM	0.77	0.58	0.8	0.2

TABLE VII: Performance between different feature extractors with Focal Loss and PCA

Table VII presents the comparative performance of different feature extractors using Focal Loss as the bag loss and SVM as the instance loss, with PCA applied for dimensionality reduction. Among the models, Univ2 Trident achieved the highest test AUC (0.88) and validation accuracy (0.8), indicating strong generalization and discriminative capability. Phikon Trident reached the highest test accuracy (0.8), but its validation accuracy was notably lower (0.2), suggesting less consistent performance on unseen data. The CLAM Feature Extractor and ResNet50 Trident both showed balanced results, with CLAM achieving a test AUC of 0.83 and ResNet50 Trident a validation AUC of 0.77. Univ1 Trident demonstrated moderate performance across all metrics. Overall, these results highlight that the choice of feature extractor significantly influences classification outcomes, with Univ2 Trident and CLAM Feature Extractor providing the most robust and reliable results under Focal Loss and PCA, while Phikon Trident excels in test accuracy but lacks stability in validation performance.

Extractor	B.Loss	I.Loss	Test AUC	Val AUC	Test ACC	Val ACC
CLAM Extractor	WCE	SVM	0.83	0.36	0.6	0.4
ResNet50 Trident	WCE	SVM	0.50	0.78	0.6	0.4
Univ1 Trident	WCE	SVM	0.70	0.50	0.6	0.4
Univ2 Trident	WCE	SVM	0.83	0.75	0.4	0.6
Phikon Trident	WCE	SVM	0.72	0.58	0.6	0.2

TABLE VIII: Performance between different feature extractors with Weighted Cross Entropy (WCE) Loss and PCA

Table VIII reports the performance of different feature extractors when using Weighted Cross Entropy (WCE) Loss as the bag loss and SVM as the instance loss, with PCA applied for dimensionality reduction. Univ2 Trident stands out with the highest validation AUC (0.75) and validation accuracy (0.6), indicating strong generalization to unseen data. Both CLAM Feature Extractor and Univ2 Trident achieve the highest test AUC (0.83), demonstrating robust discriminative ability. Phikon Trident attains the highest test accuracy (0.6), but its validation accuracy is the lowest (0.2), suggesting less stable performance. ResNet50 Trident shows a high validation AUC (0.78) but only moderate accuracy, while Univ1 Trident provides balanced but not outstanding results across all metrics. Overall, these findings highlight that Univ2 Trident and CLAM Feature Extractor are the most effective under WCE Loss with PCA, while Phikon Trident excels in test accuracy but lacks consistency in validation, and ResNet50 Trident offers good validation AUC but lower accuracy.

Across all four loss functions with PCA, Cross Entropy, Contrastive, Focal, and Weighted Cross Entropy (WCE), the results show that no single feature extractor consistently dominates in every metric, but certain trends emerge. The CLAM Feature Extractor and Univ2 Trident often achieve the highest test AUC and validation AUC, indicating strong discriminative power and generalization, especially for Cross Entropy and Focal Loss. Univ1 Trident and Phikon Trident sometimes reach the highest test accuracy, but their AUC values are generally lower, suggesting less reliable ranking of cases. ResNet50 Trident typically shows lower performance, particularly in test accuracy. The choice of loss function and the application of PCA both influence the relative performance, but the overall pattern remains: feature extractors like CLAM and Univ2 Trident tend to provide more robust results, while the impact of the loss function is secondary to the choice of feature representation. These findings highlight the importance of both dimensionality reduction and feature extractor selection in optimizing WSI classification performance.

D. Augmentation Results

The following tables present the results obtained by applying different data augmentation techniques to the Trident feature extractors. Initially, we evaluated the impact of augmentation using extrapolation-based methods, followed by experiments with the AugDiff approach. All augmentation experiments in this section were conducted exclusively on the Trident feature extractors to assess their effectiveness in enhancing classification performance.

Extractor	B.Loss	I.Loss	PCA	Test AUC	Val AUC	Test ACC	Val ACC
Aug Resnet50	CE	SVM	No	0.91	0.89	0.90	0.90
Aug Univ1	CE	SVM	No	1.00	1.00	0.90	0.90
Aug Univ2	CE	SVM	No	1.00	1.00	1.00	1.00
Aug Phikon	CE	SVM	No	0.98	0.98	0.80	0.90
Aug Resnet50	CE	SVM	Si	0.98	0.98	0.80	0.80
Aug Univ1	CE	SVM	Si	1.00	1.00	0.90	0.90
Aug Univ2	CE	SVM	Si	1.00	1.00	1.00	1.00
Aug Phikon	CE	SVM	Si	0.97	0.85	0.80	0.80

TABLE IX: Performance comparison between different feature extractors, with and without PCA, using Cross Entropy Loss and data augmentation.

Table IX presents the performance of various feature extractors under the combined effect of data augmentation and Cross Entropy Loss, both with and without the application of PCA for dimensionality reduction. The results demonstrate a substantial improvement across all metrics compared to non-augmented settings. Notably, Aug Univ2 Trident achieves perfect scores in both AUC and accuracy for test and validation sets, regardless of PCA usage, indicating exceptional discriminative power and generalization. Aug Univ1 Trident also attains perfect AUCs and high accuracy (0.90) in all configurations. Aug ResNet50 Trident and Aug Phikon Trident show strong performance, with AUCs close to or above 0.90 and high accuracy, though slightly lower than the Univ models. The application of PCA does not significantly diminish performance; in some cases, it maintains or even slightly improves the results, suggesting that dimensionality reduction can be beneficial or at least non-detrimental when combined with robust feature extraction and augmentation. Overall, these findings highlight the effectiveness of data augmentation in boosting classification performance and demonstrate that, with appropriate augmentation, all tested feature extractors can achieve high accuracy and AUC, with Univ2 Trident standing out as the most robust and consistent model.

Extractor	B.Loss	I.Loss	PCA	Test AUC	Val AUC	Test ACC	Val ACC
Aug Resnet50	Contrastive	SVM	No	0.94	0.88	0.90	0.90
Aug Univ1	Contrastive	SVM	No	1.00	1.00	1.00	1.00
Aug Univ2	Contrastive	SVM	No	1.00	1.00	0.88	0.88
Aug Phikon	Contrastive	SVM	No	0.99	0.96	0.90	0.90
Aug Resnet50	Contrastive	SVM	Si	0.98	0.97	0.80	0.80
Aug Univ1	Contrastive	SVM	Si	0.97	0.97	0.80	0.80
Aug Univ2	Contrastive	SVM	Si	1.00	1.00	1.00	1.00
Aug Phikon	Contrastive	SVM	Si	0.97	0.90	0.80	0.80

TABLE X: Performance comparison between different feature extractors, with and without PCA, using Contrastive Loss and data augmentation.

Table X summarizes the performance of different feature extractors trained with Contrastive Loss and data augmentation, both with and without PCA. The results indicate that data augmentation, in combination with contrastive learning, leads to outstanding classification performance across all feature extractors. Aug Univ1 Trident and Aug Univ2 Trident achieve perfect or near-perfect AUC and accuracy on both test and validation sets, regardless of PCA application, demonstrating exceptional generalization and discriminative ability. Aug ResNet50 Trident and Aug Phikon Trident also perform very well, with AUCs above 0.94 and high accuracy, though slightly lower than the Univ models. The use of PCA

does not significantly reduce performance; in some cases, it maintains or only slightly decreases the metrics, suggesting that dimensionality reduction is compatible with strong feature extraction and augmentation. Overall, these findings confirm that the combination of data augmentation and contrastive loss enables all tested feature extractors to reach high or perfect classification metrics, with Univ1 and Univ2 Trident models standing out for their robustness and consistency.

Extractor	B.Loss	I.Loss	PCA	Test AUC	Val AUC	Test ACC	Val ACC
Aug Resnet50	Focal	SVM	No	0.91	0.89	0.90	0.90
Aug Univ1	Focal	SVM	No	1.00	1.00	0.90	0.90
Aug Univ2	Focal	SVM	No	1.00	1.00	1.00	1.00
Aug Phikon	Focal	SVM	No	0.98	1.00	0.90	0.90
Aug Resnet50	Focal	SVM	Si	0.98	0.97	0.80	0.80
Aug Univ1	Focal	SVM	Si	0.98	1.00	0.90	1.00
Aug Univ2	Focal	SVM	Si	1.00	1.00	1.00	1.00
Aug Phikon	Focal	SVM	Si	0.98	0.88	0.80	0.80

TABLE XI: Performance comparison between different feature extractors, with and without PCA, using Focal Loss and data augmentation.

Table XI presents the results of different feature extractors trained with Focal Loss and data augmentation, both with and without PCA. The results show that all models achieve very high performance, with Aug Univ2 Trident reaching perfect AUC and accuracy on both test and validation sets, regardless of PCA application. Aug Univ1 Trident also demonstrates excellent results, with perfect AUCs and high accuracy, and even improves validation accuracy to 1.00 when PCA is applied. Aug ResNet50 Trident and Aug Phikon Trident perform strongly as well, with AUCs above 0.90 and high accuracy, though their metrics are slightly lower than those of the Univ models. The use of PCA does not substantially reduce performance; in some cases, it maintains or even enhances validation accuracy, particularly for Aug Univ1 Trident. Overall, these findings confirm that the combination of Focal Loss, data augmentation, and robust feature extraction leads to consistently high classification performance, with Univ2 Trident and Univ1 Trident standing out for their robustness and generalization.

Feature Extractor	B.Loss	I.Loss	PCA	Test AUC	Val AUC	Test ACC	Val ACC
Aug Resnet50	WCE	SVM	No	0.91	0.89	0.90	0.90
Aug Univ1	WCE	SVM	No	1.00	1.00	0.90	0.90
Aug Univ2	WCE	SVM	No	1.00	1.00	1.00	1.00
Aug Phikon	WCE	SVM	No	0.98	0.98	0.80	0.90
Aug Resnet50	WCE	SVM	Si	0.98	0.97	0.80	0.80
Aug Univ1	WCE	SVM	Si	0.97	0.97	0.80	0.80
Aug Univ2	WCE	SVM	Si	1.00	1.00	1.00	1.00
Aug Phikon	WCE	SVM	Si	0.97	0.90	0.80	0.80

TABLE XII: Performance comparison between different feature extractors, with and without PCA, using Weighted Cross Entropy (WCE) Loss and data augmentation.

Table XII shows the results of different feature extractors trained with Weighted Cross Entropy (WCE) Loss and data augmentation, both with and without PCA. The results indicate that all models achieve excellent performance, with Aug Univ2 Trident reaching perfect AUC and accuracy on both test and validation sets, regardless of PCA application. Aug Univ1 Trident also performs at a very high level, with perfect AUCs and high accuracy, and maintains strong results even when PCA is applied. Aug ResNet50 Trident and Aug Phikon Trident demonstrate robust performance as well, with AUCs above 0.90 and high accuracy, though their metrics are slightly lower than those of the Univ models. The use of PCA does not significantly reduce performance; in some cases, it maintains or only slightly decreases the metrics, suggesting that dimensionality reduction is compatible with strong feature extraction and augmentation. Overall, these findings confirm that the combination of WCE Loss, data augmentation, and robust feature extraction leads to consistently high classification performance, with Univ2 Trident and Univ1 Trident standing out for their reliability and generalization.

The results obtained with the application of augmentation techniques demonstrate a remarkable improvement in classification performance for all feature extractors when data augmentation is applied, regardless of the loss function or the use of PCA. Compared to the preceding tables without augmentation, where test and validation AUCs and accuracies were generally moderate and varied significantly across feature extractors, the augmented models achieve near-perfect or perfect scores, especially for Aug Univ2 Trident and Aug Univ1 Trident, which consistently reach AUC and accuracy values of 1.00. Even feature extractors that previously showed lower or less stable performance, such as Aug ResNet50 Trident and Aug Phikon Trident, benefit substantially from augmentation, with their metrics rising close to the top-performing models. This dramatic increase can be attributed to the effect of data augmentation, which enriches the training set with diverse and representative samples, thereby reducing overfitting and enabling the models to generalize better to unseen data. Augmentation introduces variability and simulates a broader range of real-world scenarios, allowing the feature extractors and classifiers to learn more robust and discriminative representations. As a result, the models are less sensitive to noise and class imbalance, leading to the consistently high results observed across all metrics and configurations.

The following table shows the results obtained with the AugDiff method. We have tested it only on the Resnet50 Trident feature extracted.

Extractor	B.Loss	I.Loss	PCA	Test AUC	Val AUC	Test ACC	Val ACC
AugDiff R50	CE	SVM	No	0.44	0.66	0.22	0.55
AugDiff R50	CE	SVM	Si	0.50	0.67	0.22	0.33
AugDiff R50	Contr	SVM	No	0.40	0.61	0.22	0.55
AugDiff R50	Contr	SVM	Si	0.52	0.71	0.22	0.44
AugDiff R50	Focal	SVM	No	0.44	0.62	0.22	0.55
AugDiff R50	Focal	SVM	Si	0.50	0.67	0.22	0.33
AugDiff R50	WCE	SVM	No	0.40	0.62	0.22	0.55
AugDiff R50	WCE	SVM	Si	0.50	0.67	0.22	0.33

TABLE XIII: Performance comparison between different feature extractors, with and without PCA, using all Losses and AugDiff data augmentation.

Table XIII presents the results for the AugDiff ResNet50 Trident model under various loss functions (Cross Entropy, Contrastive, Focal, and Weighted Cross Entropy), with and without PCA, in the context of data augmentation. Across all configurations, the model's performance is notably lower than

that of other augmented feature extractors reported previously. Test AUC values range from 0.40 to 0.52, and test accuracies remain at 0.22, indicating limited discriminative ability and classification accuracy. Validation AUCs are somewhat higher, reaching up to 0.71, with validation accuracies ranging between 0.33 and 0.55. However, these results still fall short of the near-perfect scores observed with other feature extractors.

The application of PCA does not lead to substantial improvements. In some cases, it slightly increases the validation AUC but reduces validation accuracy. The relatively poor performance may be attributed to the low dimensionality of the dataset, the reduced number of time steps, and the subsampling inherent in the Gaussian diffusion model. These factors likely contribute to the lower-than-expected results.

Overall, these findings suggest that, despite the use of data augmentation and various loss functions, the AugDiff ResNet50 Trident model struggles to achieve high performance. Since these results are inferior to those obtained with the Extrapolation method, we did not extend testing to other feature extractors.

E. DSMIL results

Split	Epochs	Dropout	LR	WD	Test AUC	Val AUC	Test ACC	Val ACC
0	350	0.8	5e-5	1e-5	0.2778	1.0000	0.6	1.0
0	400	0.5	1e-5	1e-5	0.2222	1.0000	0.6	1.0
0	600	0.8	1e-4	1e-5	0.2778	1.0000	0.6	1.0
0	600	0.8	5e-5	1e-5	0.3333	1.0000	0.6	1.0
0	400	0.8	1e-5	1e-5	0.2222	1.0000	0.6	1.0
0	350	0.5	1e-5	1e-5	0.7222	0.9167	0.6	0.8
0	600	0.8	1e-5	1e-5	0.6111	0.9444	0.6	0.8
0	600	0.5	1e-4	1e-5	0.3889	0.9444	0.6	0.8
0	600	0.8	5e-5	1e-4	0.3333	0.9444	0.6	0.8
0	300	0.8	5e-5	1e-4	0.2778	1.0000	0.6	1.0

TABLE XIV: Top DSMIL-SimCLR (ResNet50 as backbone) results ranked by test accuracy.

For the DSMIL, SimCLR-based, two separate feature extractors are trained for 20 epochs, with Adam scheduler (1e-5), weight decay, as scheduler cosine annealing and ResNet50 as backbone. One embedder is trained on patches at low magnification (5x) and the other on patches at higher magnification (10x). After training, features from these two zoom levels are extracted and used as input to DSMIL, allowing an exploration of how contrastive learning influences the model's ability to learn from multi-scale histological information.

DSMIL is trained with different combinations of parameters to find the best configurations: epochs from 350 to 600, dropout from 0.5 to 0.8, learning rate from 1e-5 to 1e-4, weight decay from 1e-5 to 1e-4.

Table XIV presents the top DSMIL results, used SimCLR with ResNet50 as backbone, ranked by test accuracy. Across all configurations, the test accuracy remains constant at 0.6, while the validation accuracy is either 1.0 or 0.8, indicating that the model generalizes well to the validation set but does not achieve high discriminative power on the test set. Notably, the test AUC values are consistently low (ranging from 0.22 to 0.72), whereas the validation AUCs are much higher (0.92 to 1.00), suggesting a significant gap between test and validation performance. This discrepancy may be due to overfitting or to differences in data distribution between the test and validation sets. The hyperparameters (epochs, dropout, learning rate, and weight decay) do not appear to have a strong impact on test accuracy, as all configurations yield the same value. Overall, these results indicate that while DSMIL ResNet50 can achieve perfect or near-perfect validation performance, its ability to generalize to unseen test data is limited, as reflected by the low test AUC and moderate test accuracy. This highlights the need for further optimization or regularization to improve the robustness and generalization of the DSMIL approach in this context.

F. Some Interpretability Insights

To gain intuitive and interpretable insights into the model's behavior, we leverage the blockmap generation module of CLAM, as illustrated in Figure 6. Blockmaps are spatial visualizations of attention scores computed over instance-level patches, highlighting regions within a whole slide image (WSI) that contribute most significantly to the model's prediction. In the context of medical imaging—particularly computational pathology—blockmaps play a crucial role in enhancing interpretability. They provide a window into the model's decision-making process within Multiple Instance Learning (MIL) frameworks like CLAM, enabling the localization of diagnostically relevant tissue regions without requiring pixel-level annotations. This is especially valuable given the practical constraints of histopathology, where exhaustive expert labeling is often impractical or unavailable.

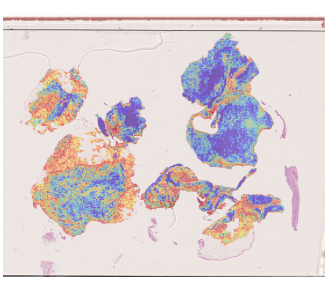
The generation of blockmaps in this study was carried out using attention weights derived from the augmented CLAM model, with patch-level features extracted via the UNIv1 Trident architecture. A segmentation threshold (*sthresh*) of 12 was used to define foreground tissue, and only patches passing this filter were retained for further analysis. The resulting attention scores were normalized and mapped back to their spatial coordinates to form 2D attention heatmaps.

Blockmaps produced in this way are not only useful for understanding model behavior but are also of practical significance in clinical settings. By highlighting regions of interest within diagnostic slides, they provide a basis for trust and validation in AI-assisted diagnosis and can potentially assist pathologists in focusing their review on the most informative regions of tissue.

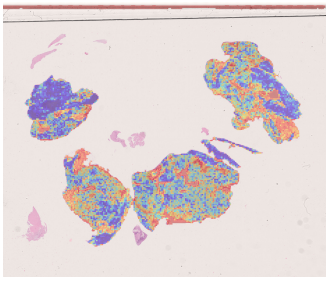
V. Conclusions and future works

This work provides a comprehensive evaluation of weakly supervised learning strategies for whole slide image (WSI) classification, focusing on the interplay between feature extraction, loss function selection, dimensionality reduction, and data augmentation. The results demonstrate that the choice of feature extractor is the most influential factor in determining classification performance. In particular, the Univ2 Trident and CLAM feature extractors consistently deliver superior results across a range of loss functions and experimental settings, with Univ2 Trident achieving perfect or near-perfect AUC and accuracy when combined with data augmentation.

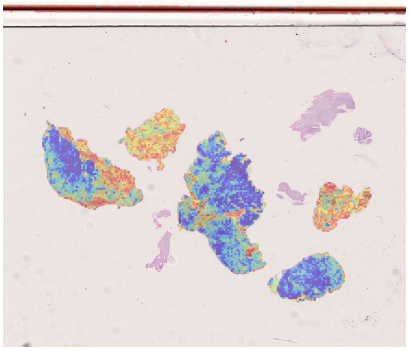
Loss function selection also plays a significant role, especially in the presence of class imbalance. Focal Loss and



(a) WSI M-1, Biphasic



(b) WSI M-10, Epithelioid



(c) WSI M-101, Sarcomatoid

Fig. 6: Blockmaps showing attention over different WSIs.

Weighted Cross Entropy (WCE) are shown to improve robustness, particularly for the CLAM extractor, while supervised contrastive loss further enhances discriminative capability in some settings. However, the impact of loss function is generally secondary to the choice of feature extractor.

Dimensionality reduction via PCA is found to be compatible with strong feature extraction and augmentation, sometimes even improving generalization, but it does not fundamentally alter the ranking of model performance. The application of feature-level data augmentation, including both extrapolation-based and diffusion model-based methods, leads to dramatic improvements in classification metrics for most extractors. Augmentation not only increases the diversity of the training set but also mitigates overfitting and enhances the models' ability to generalize to unseen data. Notably, all Trident-based models benefit substantially from augmentation, with Univ1 and Univ2 Trident achieving perfect scores in several configurations. In contrast, the AugDiff ResNet50 Trident model does not show similar gains, highlighting that the effectiveness of augmentation is contingent on the underlying feature extractor and its compatibility with the augmentation strategy.

The DSMIL ResNet50 experiments reveal a persistent gap between validation and test performance, suggesting overfitting or sensitivity to data distribution shifts. While DSMIL can achieve high validation accuracy, its generalization to independent test data remains limited, indicating the need for further regularization or architectural refinement.

Overall, this study underscores the critical importance of selecting appropriate feature extractors and loss functions, and demonstrates that advanced data augmentation techniques can substantially elevate WSI classification performance. The findings advocate for a holistic approach that combines robust feature representation, tailored loss functions, and effective augmentation to address the challenges of weak supervision and class imbalance in computational pathology.

Future work should explore the integration of more advanced self-supervised learning methods, domain adaptation strategies to address distributional shifts, and the application of these frameworks to larger and more diverse datasets. Additionally, further investigation into the interpretability and clinical relevance of model predictions will be essential for translating these advances into practical diagnostic tools.

References

[1] K. Inai, "Pathology of mesothelioma," *Environmental Health and Preventive Medicine*, vol. 13, no. 2, pp. 60–64, 2008.

[2] G. Alì, R. Bruno, and G. Fontanini, "The pathological and molecular diagnosis of malignant pleural mesothelioma: a literature review," *Translational Lung Cancer Research*, vol. 10, no. 1, pp. 72–83, 2021.

[3] L. Brcic and I. Kern, "Clinical significance of histologic subtyping of malignant pleural mesothelioma," *Translational Lung Cancer Research*, vol. 7, no. 5, pp. 556–569, 2018.

[4] M. Gadermayr and M. Tschuchnig, "Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations future potential," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1121–1135, 2022.

[5] B. Voigt, O. Fischer, B. Schilling, C. Krumnow, and C. Hertaba, "Investigation of semi- and self-supervised learning methods in the histopathological domain," *Computers in Biology and Medicine*, vol. 144, p. 105377, 2022.

[6] L. Qu, Y. Ma, X. Luo, Q. Guo, M. Wang, and Z. Song, "Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1217–1226.

[7] M. Y. Lu, D. F. 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.

[8] A. Zhang, G. Jaume, A. Vaidya, T. Ding, and F. Mahmood, "Accelerating data processing and benchmarking of ai models for pathology," *Nature Communications*, vol. 14, no. 1, p. 1456, 2023.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 1597–1607.

[10] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," in *International Conference on Learning Representations (ICLR)*, 2017.

[11] Z. Shao, L. Dai, Y. Wang, H. Wang, and Y. Zhang, "Augdiff: Diffusion based feature augmentation for multiple instance learning in whole slide image," *arXiv preprint arXiv:2309.07935*, 2023.