Machine Learning in Applications
A semi-supervised solution for mesothelioma classification project report

Jiang Peichun, Khattar Reem, Tallone Samuele
Politecnico di Torino

CONTENTS

I Introduction 2

II Background 2
- II-A Multiple Instance Learning (MIL) . . . 2
- II-B Whole Slide Imaging (WSI) . . . 2
- II-C Feature Extractors: ResNet, KimiaNet, and UNI . . . 3
- II-D Feature Aggregator: CLAM . . . 3
- II-E Manual Feature Engineering and the PINS Concept . . . 3

III Methodology 3
- III-A Tissue Segmentation . . . 4
- III-B Patch Extraction . . . 4
- III-C Overlay Generation . . . 4
- III-D Feature Extractors . . . 4
- III-E Manual Feature Extraction . . . 4
- III-F Feature Aggregator . . . 4
- III-G Dataset . . . 5

IV Training Details 5
- IV-A Loss Function . . . 5
- IV-B Optimization and Regularization . . . 6
- IV-C Model Configuration . . . 6
- IV-D Cross-Validation Strategy . . . 6

V Results and Discussion 6
- V-A Confusion Matrixes . . . 6
- V-B PCA Features Visualization . . . 7
- V-C Benchmark with Simple Classifiers . . . 7
- V-D Heatmap Generation . . . 9

VI Conclusions and Future Works 9

References 10

List of Figures

1 Multi-resolution pyramid representation of a Whole Slide Image (WSI). The base level (1:1, $40\times$) contains the full-resolution gigapixel data, while higher pyramid levels store progressively downsampled versions. This structure enables efficient visualization, tiling, and patch-based processing for downstream analysis such as Multiple Instance Learning (MIL). 3
2 Main steps of the tissue segmentation pipeline. From left to right: downsampled thumbnail, grayscale conversion, Laplacian-enhanced image, binary mask after Otsu thresholding, and final refined tissue mask after morphological operations and artifact removal. 5
3 Example of patch selection visualization. Left: downsampled thumbnail of the WSI. Right: semi-transparent overlay showing the locations of selected patches, highlighting coverage of tissue regions while avoiding background. . . 5
4 Architecture of the CLAM (Clustering-constrained Attention Multiple Instance Learning) model: patch-level features extracted by the backbone are weighted through an attention branch to produce a slide-level prediction. The framework also incorporates attention pooling and an instance-level clustering module that separates positive and negative evidence, improving both discriminative power and interpretability. 6
5 Confusion matrices for the four feature extraction strategies. Rows represent ground-truth classes, while columns represent predicted classes. Class 0 corresponds to Biphasic, class 1 to Epithelioid, and class 2 to Sarcomatoid mesothelioma. KimiaNet shows the most balanced classification performance, whereas the inclusion of manually engineered features in ResNet50 (M) leads to reduced discriminative capability. 7
6 Comparison of learned representations and feature quality across different extractors. Top row: PCA visualizations of slide-level features learned by CLAM show how well the model organizes different tumor subtypes in feature space. Bottom row: performance of simple classifiers on aggregated patch features reveals intrinsic discriminative power of each feature extractor, independent of the MIL framework. Class 0: Biphasic, Class 1: Epithelioid, Class 2: Sarcomatoid. 8
7 Heatmap based on CLAM attention for a correct prediction of biphasic cancer. The redder areas indicate regions that the model considers more informative. 9

List of Tables

I Distribution of mesothelioma subtypes in the original and final datasets 5
II Performance comparison of different feature extractors for mesothelioma subtype classification. Reported values correspond to the mean across the cross-validation folds. ResNet50 (M) denotes the model in which manually engineered features were concatenated with deep features extracted by ResNet50. 6

Machine Learning in Applications A semi-supervised solution for mesothelioma classification project report

***Abstract*—Malignant Pleural Mesothelioma (MPM) requires accurate histological subtyping—epithelioid, sarcomatoid, or biphasic—to guide clinical prognosis and treatment. However, manual classification is challenged by tissue heterogeneity and the immense scale of Whole Slide Images (WSIs). This project presents a semi-supervised pipeline utilizing a Multiple Instance Learning (MIL) framework, specifically CLAM (Clustering-constrained Attention Multiple Instance Learning), to classify MPM subtypes from WSIs under weak supervision. We evaluated three deep learning feature extractors—ResNet50, KimiaNet, and UNI—and investigated the integration of manually engineered morphological features (PINS concept) to assess their discriminative value.**

Experimental results on a cohort of 42 cases demonstrated that KimiaNet achieved superior performance with an accuracy of 0.869 and a weighted F1-score of 0.751, significantly outperforming the general-purpose ResNet50 baseline. Notably, the inclusion of handcrafted features (ResNet50 (M)) consistently degraded model performance (F1-score: 0.393) and increased computational overhead. While the UNI foundation model showed high intrinsic discriminative power in simple classifier benchmarks, its performance within the MIL framework was limited by the small dataset size. Our findings suggest that specialized pathology-specific backbones are more effective and scalable than manual feature engineering for MPM subtyping, though larger datasets are required to fully leverage high-capacity foundation models.

I. Introduction

Malignant pleural mesothelioma is a rare and aggressive cancer originating from the mesothelial cells lining the pleura [1]. It is strongly associated with asbestos exposure [1]. Histologically, MPM is classified into three major subtypes: epithelioid, sarcomatoid, and biphasic (or mixed) [2].

The epithelioid subtype is the most common, accounting for approximately 60–80% of cases [2]. It is composed of uniform epithelial-like cells and exhibits a tubulopapillary or solid growth pattern [2]. This subtype generally correlates with a better prognosis and higher treatment responsiveness [3].

The sarcomatoid variant is a less frequent (10–20%) but more aggressive variant consisting of spindle-shaped cells [2]. It is associated with poor therapeutic response and a worse overall prognosis [3]. The biphasic subtype (10–15%) contains both epithelioid and sarcomatoid components [2]. The clinical behavior of biphasic mesothelioma lies between the two, but its prognosis worsens as the sarcomatoid component increases [2]. Accurate and reproducible subtype classification is essential, as it directly informs treatment decisions, eligibility for surgery, and enrollment in clinical trials [4].

The advent of digital pathology and the availability of high-resolution Whole Slide Images (WSIs) have opened the way for developing AI-driven tools to assist in diagnostic workflows [5]. Deep learning models, particularly those based on convolutional neural networks (CNNs), have shown promise in capturing complex morphological patterns within histological images [6]. However, the application of such models to mesothelioma subtype classification remains underexplored, especially considering the challenges posed by histological heterogeneity and the presence of multiple tissue types within a single slide [4].

A significant challenge in developing automated classification models for mesothelioma subtypes is the nature of available annotations [5]. Labels are typically assigned at the WSI level, indicating the predominant subtype present [5]. However, each WSI may contain a mixture of tissue types, including non-tumorous regions, leading to a weakly supervised learning scenario [7], [8].

In such cases, Multiple Instance Learning (MIL) frameworks have been employed effectively [8]. MIL treats each WSI as a bag of instances, with the slide-level label applying to the entire bag [8]. Attention-based MIL approaches have been particularly successful in identifying diagnostically relevant regions within WSIs, facilitating accurate classification despite the lack of detailed annotations [9].

This project aims to develop a robust deep learning pipeline for the classification of mesothelioma subtypes from WSIs under weak supervision [4]. We implement and evaluate a novel framework that uses both manually extracted features and features extracted from a pretrained CNN and works on the entire WSI. We also provide a comparison with current state of the art features extractors, to evaluate if manual features extraction actually provides substantial benefits, over current base of foundation models.

II. Background

A. Multiple Instance Learning (MIL)

In digital pathology, Whole Slide Images (WSIs) are too large to be processed directly by standard deep learning architectures. Multiple Instance Learning (MIL) provides a weakly supervised solution by treating each WSI as a ”bag” of many small image patches, or ”instances”. Instead of requiring expensive pixel-level annotations, MIL allows the model to learn from slide-level labels by identifying the most discriminative instances within the bag that support the final diagnosis.

B. Whole Slide Imaging (WSI)

Whole Slide Imaging (WSI) is the process of digitizing traditional glass histology slides into high-resolution ”gigapixel” images. Unlike standard digital photographs, WSIs capture

tissue at an extreme level of detail, often achieving a resolution of 0.25 μm per pixel when scanned at $40\times$ magnification.

Because of this high resolution, WSIs result in massive data files. As shown in the provided documentation, uncompressed files can occupy approximately 48 megabytes for every square millimeter of scanned tissue. A single slide can easily exceed 50 GB in size, making it impossible to load the entire image into a computer's Random Access Memory (RAM) or a GPU's VRAM for direct processing.

To address these computational constraints, WSIs are organized in a **multi-resolution pyramid format** [8]. This structure allows for efficient data handling by storing the image at multiple scales:

- **Level N (Thumbnail):** At the apex of the pyramid is the lowest resolution version, providing a global view of the entire tissue sample.
- **Intermediate Levels:** These represent "downsampled" versions (e.g., $5\times, 10\times, 20\times$), which balance context and detail.
- **Level 0 (Base Level):** The bottom of the pyramid contains the raw, highest-resolution data. This level is where critical cellular details and fine tissue structures—essential for cancer subtyping—are most discernible.

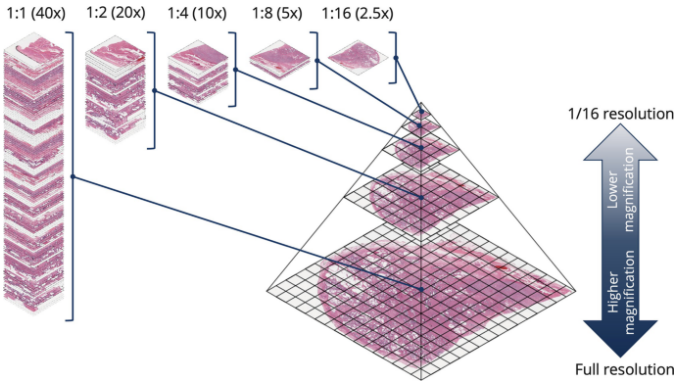


Fig. 1: Multi-resolution pyramid representation of a Whole Slide Image (WSI). The base level (1:1, $40\times$) contains the full-resolution gigapixel data, while higher pyramid levels store progressively downsampled versions. This structure enables efficient visualization, tiling, and patch-based processing for downstream analysis such as Multiple Instance Learning (MIL).

By using this pyramidal architecture, researchers can "tile" the image, extracting only small, high-resolution patches from Level 0 to be processed individually. This hierarchical organization is the foundational requirement for the Multiple Instance Learning (MIL) approach discussed in the following sections.

C. Feature Extractors: ResNet, KimiaNet, and UNI

To reduce the dimensionality of gigapixel slides, each patch is converted into a compact feature vector using a pretrained backbone. In this project, we evaluate three distinct extractors:

- **ResNet50:** A standard residual network pretrained on ImageNet, serving as a general-purpose baseline for morphological feature extraction [10].
- **KimiaNet:** A specialized DenseNet architecture fine-tuned specifically on over 240,000 histopathology images from The Cancer Genome Atlas (TCGA). It is designed to capture domain-specific features of human tissue that general models may miss [6].
- **UNI:** A state-of-the-art foundation model for pathology, utilizing a Vision Transformer (ViT) architecture trained on over 100 million histological patches. UNI aims to provide a universal representation for diverse tissue types and rare pathologies like mesothelioma [11].

D. Feature Aggregator: CLAM

The Clustering-constrained Attention Multiple Instance Learning (CLAM) framework is used to aggregate the patch-level features into a slide-level prediction. CLAM utilizes an attention mechanism to assign importance scores to each patch, allowing the model to focus on tumor-rich regions while ignoring background artifacts. Additionally, CLAM introduces instance-level clustering, which encourages the model to learn distinct representations for different tissue classes, thereby improving both accuracy and interpretability.

E. Manual Feature Engineering and the PINS Concept

While deep features are powerful, they are often considered "black boxes" that may overlook specific geometric or spatial relationships between cells. The PINS (Pathology-Image-based Numerical Signatures) approach [4] suggests that manually engineered features, such as nuclear size, cell density, and spatial distribution, can provide a robust numerical signature for cancer subtyping. By extracting these handcrafted features via QuPath and concatenating them with ResNet50 deep features, we aim to investigate whether traditional morphological indicators can enhance the performance of modern MIL pipelines on small datasets. Since the PINS paper lacked this comparison, we also aim to investigate whether manually extracted features offer any significant advantages over modern foundation models used as feature extractors. Given that these models are pre-trained on large-scale histopathological data, they may reduce, or potentially eliminate, the need for manual feature engineering.

III. Methodology

For our methodology, we start by segmenting the WSI into patches, since it is unfeasible to pass the entire slide to the network due to its substantial size. The pipeline itself is divided into tissue segmentation and patch extraction. After we explain the models that we used both for features extraction and features aggregations.

A. Tissue Segmentation

The tissue was segmented following standard practices commonly used in state-of-the-art models for WSI classification. To reduce processing times and, especially, RAM usage, tissue segmentation was carried out on a downsampled representation of each WSI rather than at full resolution (16× from level 0).

The first step was to extract a thumbnail of the original image and convert it to grayscale. When enabled, a Laplacian filter was applied to enhance structural information such as tissue boundaries and cellular regions.

Binary tissue masks were then obtained using Otsu thresholding, an automatic image segmentation technique that determines an optimal global threshold to separate pixels into two classes by maximizing inter-class variance from a grayscale image histogram.

To increase the robustness of this solution and mitigate segmentation failures caused by staining artifacts, if the fraction of detected tissue fell outside the interval [8%, 70%], the segmentation was repeated without the Laplacian filter. This also occurred if the images were listed in a specific CSV file.

Afterward, morphological closing was performed using a circular structuring element with a radius of 50 μm (converted to pixels according to the slide resolution and downsampling factor). Morphological closing consists of a dilation followed by an erosion. This operation fills small holes inside tissue regions and smooths irregular boundaries.

Next, connected components in the binary mask were identified, and objects and holes smaller than 100,000 μm^2 were removed, since they were mostly segmentation errors or dust. In this way, the final tissue mask used in the patch extraction process was obtained. The steps are all visible in the (Figure 2).

B. Patch Extraction

Once the tissue mask was obtained at downsampled resolution, it was used to guide patch selection at full WSI resolution.

Instead of scanning the original gigapixel image directly, a sliding window was applied to the downsampled tissue mask, significantly reducing memory usage. Only regions with a tissue fraction greater than a predefined threshold (90%) were retained. This filtering step ensures that selected patches contain predominantly tissue and minimizes the inclusion of background or glass regions.

Although the selection was performed in the downsampled coordinate system, the final patch coordinates were mapped back to the native resolution of the slide. All patches were then extracted at level 0, corresponding to the highest available spatial resolution of the WSI and with a patch size and stride of 512×512 pixels. The valid patch coordinates were stored in an HDF5 file for efficient access during training.

C. Overlay Generation

To visually verify patch selection, semi-transparent overlays of the selected regions were projected onto a downsampled version of each WSI. This allows quick inspection of tissue coverage and ensured that patches avoid background or glass regions. The examples are in the figure below (Figure 3):

D. Feature Extractors

Since we are working within the MIL paradigm and do not have labels for individual patches to supervise training, a combination of feature extractors and feature aggregators must be used to successfully classify the data.

We decided to utilize two state-of-the-art feature extractors, UNI and KimiaNet, which are specifically trained on histopathological images and are expected to provide a better representation of the patches. To these, we added a standard feature extractor, ResNet50, to serve as a baseline for performance evaluation and, furthermore, to assess whether manual feature extraction can provide enough distinction between patches to improve the performance of a standard classifier compared to state-of-the-art models, especially on smaller datasets. All of these models extracted features vectors of dimension 1024.

E. Manual Feature Extraction

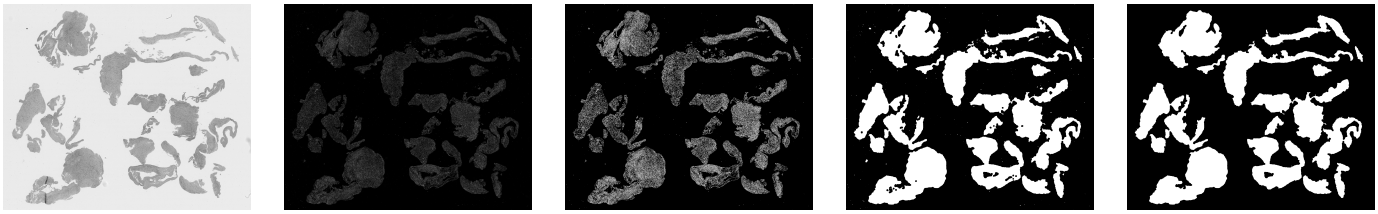
The concept of manual feature extraction is not novel in the machine learning field, as it formed the foundation of early machine learning approaches. However, to the best of our knowledge, the only study that applies it to H&E WSI mesothelioma subtype classification is the PINS paper. That work focused on applying manual feature extraction via the software QuPath on TMA cores, which are notably smaller and easier to handle than full WSIs. Additionally, it did not provide a comparison with current state-of-the-art feature extractors, which are trained on large-scale datasets and therefore capable of capturing a wide variety of morphological tissue patterns.

For this research project, manual features were extracted using QuPath. Specifically, for each WSI, we uploaded the previously extracted patches and performed cell detection on each patch. After cell detection, spatial features as well as morphological and intensity-based features were extracted for each detected cell. Since these features corresponded to individual cells, they were then averaged across all cells in the patch, and both the mean and standard deviation were saved as the final patch-level features. In total, a feature vector of approximately 100 elements was obtained for each patch. This vector was then added to the one extracted from ResNet50. The list of all the manual features that were kept is avaiable in the file avg_std.py in our GitHub repository.

F. Feature Aggregator

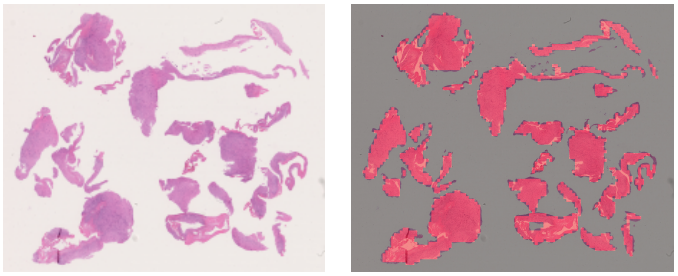
After extracting features from individual patches, these need to be aggregated to obtain a slide-level representation suitable for classification under the Multiple Instance Learning (MIL) paradigm. In this work, we employed CLAM (Clustering-constrained Attention Multiple Instance Learning) [9], a state-of-the-art MIL framework designed for WSI classification.

CLAM extends the standard attention-based MIL approach by introducing a clustering constraint on the instance-level embeddings, encouraging the model to learn discriminative subgroups of patches that are most informative for the slide-level label. This allows the model to assign attention scores to individual patches, effectively weighing their contribution to the final classification.



(a) Grayscale (b) Laplacian (c) Otsu mask (d) Morph closing (e) Final mask

Fig. 2: Main steps of the tissue segmentation pipeline. From left to right: downsampled thumbnail, grayscale conversion, Laplacian-enhanced image, binary mask after Otsu thresholding, and final refined tissue mask after morphological operations and artifact removal.



(a) Thumbnail (b) Overlay

Fig. 3: Example of patch selection visualization. Left: downsampled thumbnail of the WSI. Right: semi-transparent overlay showing the locations of selected patches, highlighting coverage of tissue regions while avoiding background.

Formally, given a bag of patch-level feature vectors $\{x_i\}_{i=1}^{N}$, CLAM computes attention scores a_i for each patch and aggregates them into a slide-level representation:

$$z = \sum_{i=1}^{N} a_i x_i$$

where z is the slide-level embedding used for classification. The attention mechanism enables interpretability, as an heatmap can be generated allowing for the visualization of the most important patches in the diagnosis.

CLAM has been proven to be a reliable model and is still used as a benchmark for newer models now, it's architechture is displayed in (Figure 4).

G. Dataset

The dataset consisted of approximately 120 patient slides, with only one case available per patient. The distribution of histological subtypes was highly imbalanced: the majority of samples were Epithelioid, 19 cases were Biphasic, and only 5 were Sarcomatoid.

To obtain a more balanced subset for our experiments, we selected a reduced cohort. In preliminary test runs, the models tended to focus primarily on the epithelioid class while largely ignoring the other histological subtypes. Additionally, the high computational cost of manual feature extraction on the available CPUs limited the number of slides that could be processed. The final dataset used in this study consisted of 19 Epithelioid, 19 Biphasic, and 4 Sarcomatoid slides. One Sarcomatoid case was excluded because it required substantially more processing time and computational resources than the others. The summary of the datasts used can be visualized in (Table I).

Subtype	Total Dataset	Used Dataset
Epithelioid	96	19
Biphasic	19	19
Sarcomatoid	5	4
Total	120	42

TABLE I: Distribution of mesothelioma subtypes in the original and final datasets

The csv of the dataset used, along with the coordinates of the patches extracted, the features extracted both manually and using state of the art feature extractors are avaiable at this link: Google Drive Repository.

IV. Training Details

The CLAM model was trained under the Multiple Instance Learning (MIL) framework using slide-level labels only. Each whole-slide image was represented as a bag of patch-level feature vectors, and the model learned to aggregate instance information through an attention-based pooling mechanism.

A. Loss Function

Training was driven by a combination of bag-level and instance-level supervision, as originally proposed in CLAM. The bag-level classification loss was computed using cross-entropy between the slide-level logits and the ground-truth label. In parallel, an instance-level clustering loss was applied to encourage separation between high-attention positive and negative instances within each bag. The total loss was defined as:

$$\mathcal{L}_{total} = \lambda\, \mathcal{L}_{bag} + (1 - \lambda)\, \mathcal{L}_{inst}$$

where $\mathcal{L}_{bag}$ is the slide-level cross-entropy loss supervising the global slide prediction, $\mathcal{L}_{inst}$ is the instance-level clustering loss that enforces separation between high- and low-attention patches, and $\lambda = 0.7$ is a weighting hyperparameter that balances slide-level classification accuracy and the discriminative structuring of patch-level feature representations.

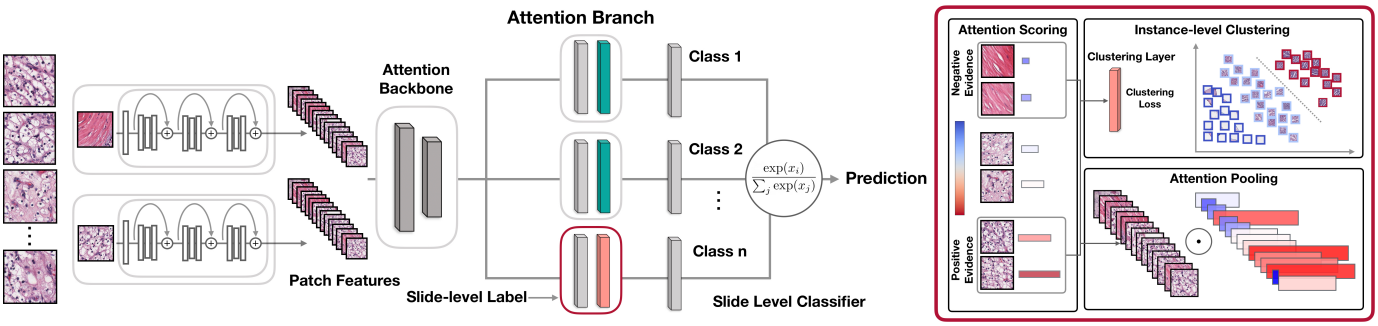


Fig. 4: Architecture of the CLAM (Clustering-constrained Attention Multiple Instance Learning) model: patch-level features extracted by the backbone are weighted through an attention branch to produce a slide-level prediction. The framework also incorporates attention pooling and an instance-level clustering module that separates positive and negative evidence, improving both discriminative power and interpretability.

B. Optimization and Regularization

CLAM was trained using the Adam optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. A dropout rate of 0.25 was applied within the CLAM architecture to reduce overfitting, which is particularly important given the limited dataset size. Training was performed for 20 epochs per fold.

C. Model Configuration

We adopted the single-branch CLAM variant (CLAM-SB) with gated attention and a small model configuration. The embedding dimension was set to 1024 or 1114, matching the dimensionality of the extracted patch-level features. During training, the model sampled $k = 8$ high- and low-attention instances per bag for the instance-level clustering objective, as suggested in the original implementation.

D. Cross-Validation Strategy

To ensure robust evaluation on the limited dataset, we employed stratified 4-fold cross-validation. For each fold, the model achieving the lowest validation loss was selected as the best checkpoint. Performance was assessed using accuracy, weighted F1-score, AUC, and confusion matrices at the slide level.

V. Results and Discussion

This section presents the experimental results obtained using the pipeline described in Section III. The evaluation has two main objectives: (i) to assess the contribution of manually extracted features to mesothelioma subtype classification, and (ii) to compare whether manually engineered features are more or less informative than representations extracted using state-of-the-art pathology-specific feature extractors.

The classification results for the three histological subtypes are summarized in Table II.

From Table II, it is evident that KimiaNet achieves the best overall performance across all evaluation metrics. In contrast, the use of manually engineered features does not appear to provide a benefit for subtype classification. When concatenated with deep features (ResNet50 (M)), these handcrafted features lead to lower performance compared to using deep features alone, suggesting that they may introduce redundant or less discriminative information into the model. Strange was also the performance of UNI model that should be able to provide richer features then other features extractors since it has been trained on the bigger and more diverse dataset.

Feature Extractor	F1-score	Accuracy	AUC
ResNet50	0.516	0.559	0.709
UNI	0.603	6340	0.713
KimiaNet	0.751	0.777	0.869
ResNet50 (M)	0.393	0.464	0.684

TABLE II: Performance comparison of different feature extractors for mesothelioma subtype classification. Reported values correspond to the mean across the cross-validation folds. ResNet50 (M) denotes the model in which manually engineered features were concatenated with deep features extracted by ResNet50.

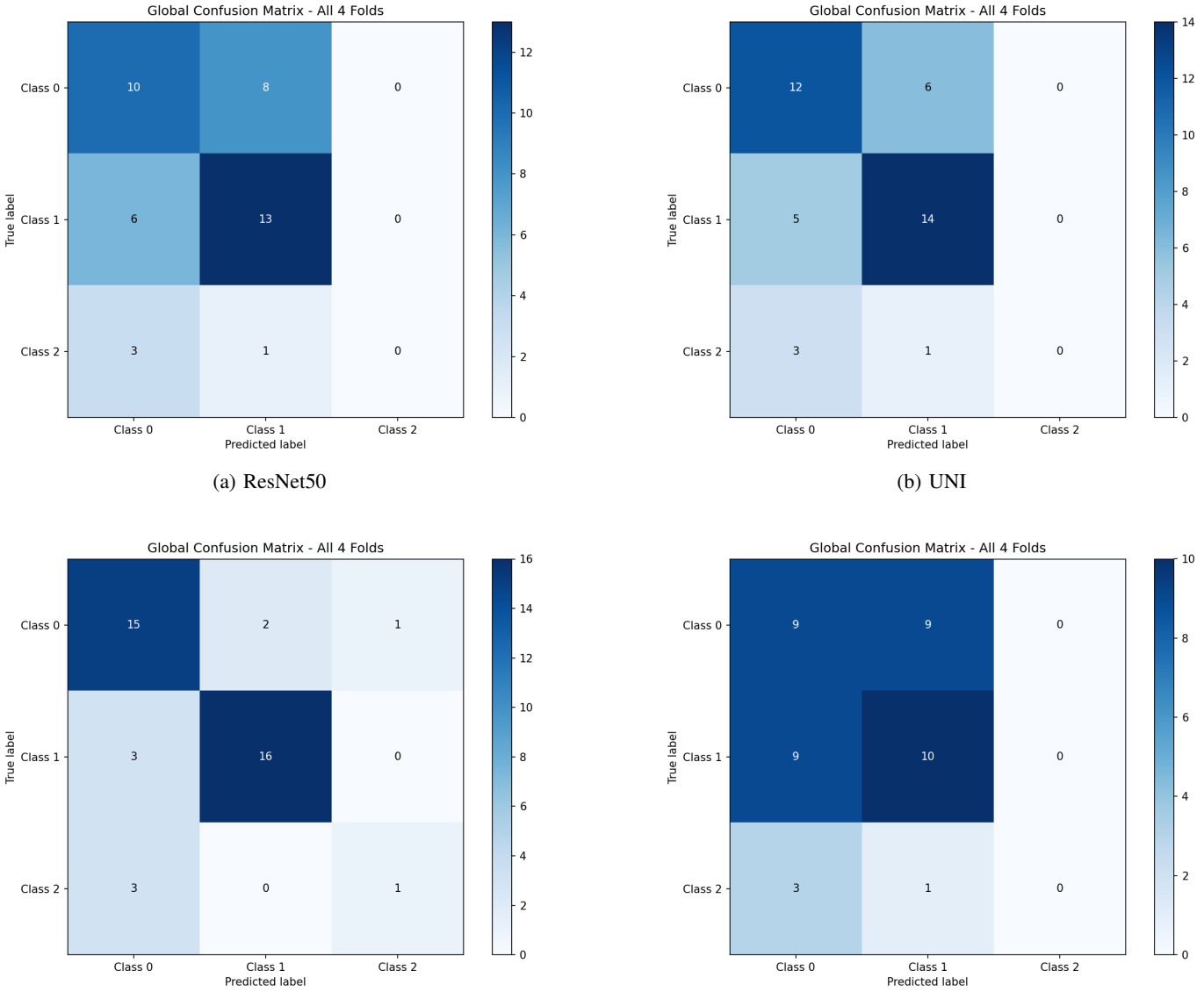
A. Confusion Matrixes

From the aggregated confusion matrices obtained by summing the results across all cross-validation folds (Figure 5), it can be observed that the primary challenge lies in distinguishing the sarcomatoid subtype from the other classes. This difficulty is likely due to the very limited number of sarcomatoid samples in the dataset, which reduces the model's ability to learn representative features for this class and leads to frequent misclassifications.

It is also evident that Kiminet performs much better than the other networks, as it is the only one that manages to correctly classify a sarcomatoid sample, and it also achieves significantly better discrimination between Bifasic and Epithelioid subtypes.

The performance of UNI is surprising, as it is similar to ResNet50 despite being trained on a much larger histopathological dataset. This likely indicates that the CLAM model is unable to fully leverage the more complex feature representations of UNI on a dataset of this scale.

Once again, we can see that manually extracted features were insufficient and did not aid in classification, creating



(a) ResNet50

(b) UNI

(c) KimiaNet

(d) ResNet50 (M)

Fig. 5: Confusion matrices for the four feature extraction strategies. Rows represent ground-truth classes, while columns represent predicted classes. Class 0 corresponds to Biphasic, class 1 to Epithelioid, and class 2 to Sarcomatoid mesothelioma. KimiaNet shows the most balanced classification performance, whereas the inclusion of manually engineered features in ResNet50 (M) leads to reduced discriminative capability.

more confusion between Bifasic and Epithelioid than using ResNet50 features alone.

B. PCA Features Visualization

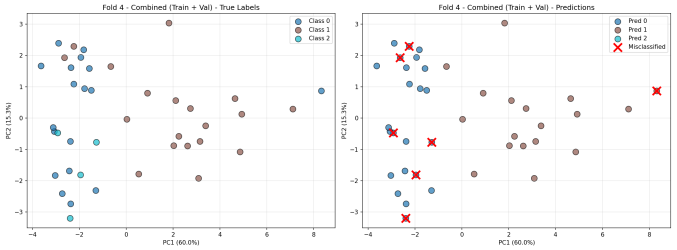
We performed a visualization analysis to understand how the model organizes slides in its learned feature space. For each slide, we extracted the final feature vector produced after attention-based pooling of patch-level features. Since these vectors are high-dimensional, we used Principal Component Analysis (PCA) to project them into two dimensions while preserving the main patterns of variation.

We created separate PCA plots for training and validation sets. In each plot, we colored points according to their true class labels in one view and predicted labels in another view. Misclassified samples were highlighted with red markers. This approach allows us to visually assess whether the model learned to separate different tumor classes into distinct regions of the feature space, and to identify which classes are most frequently confused.

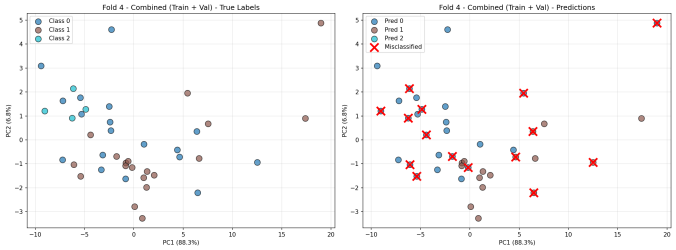
We included here only some of the more meaningful where it shows that KimiNet is able to correclty separate the classes in space (Figure 6a), wherease ResNet50 (M) fails to do so (Figure 6b). All the other analysis and tables are avaiable at this link: Google Drive Repository.

C. Benchmark with Simple Classifiers

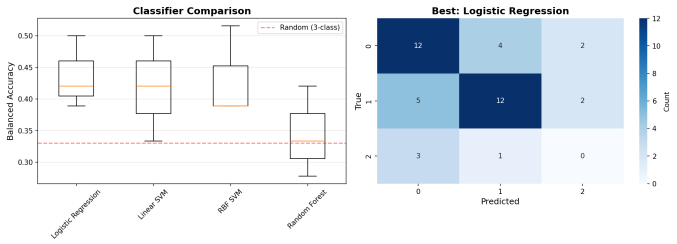
However, the previous analyses only tell part of the story. To assess whether the features extracted from different feature



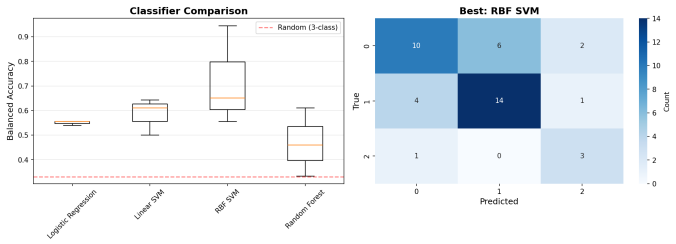
(a) PCA visualization of CLAM-learned features using KimiaNet backbone. Left: true labels. Right: predictions with misclassifications marked. The three classes form relatively distinct clusters, demonstrating good class separability.



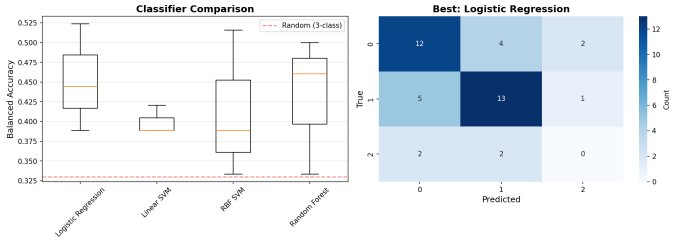
(b) PCA visualization of CLAM-learned features using ResNet50 with manually engineered features. Left: true labels. Right: predictions with misclassifications marked. Greater overlap between classes indicates reduced separability compared to deep features alone.



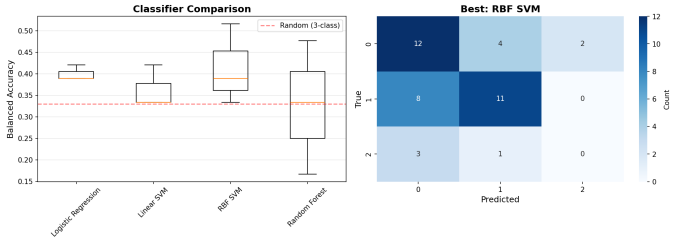
(c) Simple classifier benchmark on ResNet50 features. Left: balanced accuracy distribution. Right: confusion matrix for best classifier (Logistic Regression). ResNet50 features show moderate discriminative power.



(d) Simple classifier benchmark on UNI features. Left: balanced accuracy distribution showing superior performance. Right: confusion matrix for best classifier (RBF SVM). UNI features prove most informative even with simple aggregation.



(e) Simple classifier benchmark on KimiaNet features. Left: balanced accuracy distribution. Right: confusion matrix for best classifier (Logistic Regression). KimiaNet features demonstrate good discriminative power across all subtypes.



(f) Simple classifier benchmark on ResNet50 with manual features. Left: balanced accuracy near random baseline. Right: confusion matrix for best classifier (RBF SVM). Manual features degrade performance, suggesting they introduce noise.

Fig. 6: Comparison of learned representations and feature quality across different extractors. Top row: PCA visualizations of slide-level features learned by CLAM show how well the model organizes different tumor subtypes in feature space. Bottom row: performance of simple classifiers on aggregated patch features reveals intrinsic discriminative power of each feature extractor, independent of the MIL framework. Class 0: Biphasic, Class 1: Epithelioid, Class 2: Sarcomatoid.

extractors are truly informative, independently of potential biases introduced by the CLAM model, we implemented a benchmark using simpler classifiers applied directly to the slide-level representations obtained by aggregating patch-level features. This analysis allows us to evaluate the discriminative power of the features regardless of the complexity of the MIL model, and to compare the effectiveness of state-of-the-art deep learning extractors with manually engineered features.

Specifically, we considered several standard classifiers, including Logistic Regression, Support Vector Machines (linear and RBF kernels), and Random Forests, applied on features aggregated through methods such as mean, max, combinations of statistical moments, and attention-based pooling. Performance was assessed via stratified cross-validation using accuracy, balanced accuracy, and AUC, along with confusion matrices to identify the classes most frequently misclassified.

Balanced accuracy was computed as the average recall across classes, thus assigning equal weight to each class independently of its frequency.We included this metric because the dataset is class-imbalanced, and standard accuracy may be biased toward the majority class.

From this analysis, we can see that the features extracted by UNI (Figure 6d) are the most informative of all, as they are more easily separable even with a standard classifier. KimiaNet also performs well (Figure 6e), whereas ResNet50 (Figure 6c) and ResNet50 (M) (Figure 6f) show the lowest performance.

This suggests that the primary limitation for the features extracted by UNI likely lies in the feature aggregator itself, which may require a bigger dataset in order to take advantage of the richer features produced by UNI, rather than in the feature extractor failing to produce informative representations. In contrast, the manually engineered features in our case were not informative at all and actually hindered the performance of ResNet50 when combined, creating additional confusion between the classes.

D. Heatmap Generation

We chose an attention-based model like CLAM because it is easy to generate a heatmap of the predictions (Figure 7), which we considered a valuable resource for expert pathologists in real-world applications. Unfortunately, we are not able to evaluate the quality of these predictions ourselves, as we are not experts in this field.

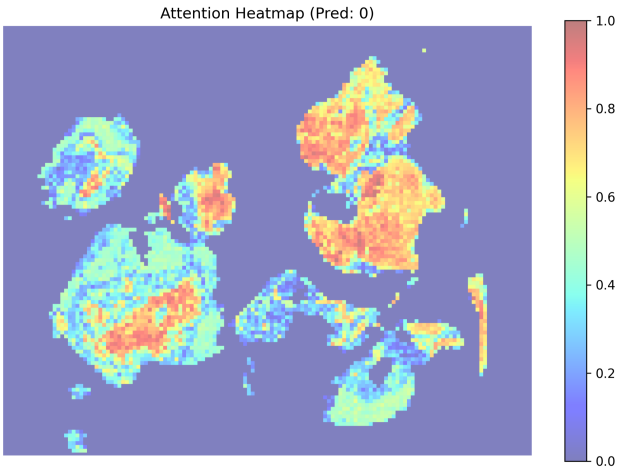


Fig. 7: Heatmap based on CLAM attention for a correct prediction of biphasic cancer. The redder areas indicate regions that the model considers more informative.

VI. Conclusions and Future Works

In this project we investigated the classification of malignant pleural mesothelioma subtypes using a Multiple Instance Learning framework, with two primary objectives: (i) assessing the contribution of manually engineered features to mesothelioma subtype classification, and (ii) comparing the discriminative power of handcrafted morphological features against representations extracted by state-of-the-art deep learning-based feature extractors.

Our experimental results provide clear answers to both questions. First, manually extracted features did not improve classification performance. When concatenated with ResNet50 deep features, these handcrafted features consistently degraded performance across all evaluation metrics compared to using ResNet50 features alone. This was evident both in the CLAM framework (F1-score dropping from 0.516 to 0.393) and in the simple classifier benchmark, where performance approached random baseline levels. The PCA visualization further confirmed that manual features introduced noise rather than discriminative information, leading to greater overlap between tumor subtypes in the learned feature space.

This finding contrasts with the PINS approach, which successfully applied manual feature extraction to TMA cores for mesothelioma classification. However, several factors may explain this discrepancy. First, TMA cores are significantly smaller and more homogeneous than full WSIs, making manual feature extraction more reliable and less prone to artifacts caused by tissue heterogeneity. Second, since we were limited by our hardwere we have missed some handcrafted features included in their pipeline, as they reported extracting approximately 320 additional features on top of those derived from ResNet.

However, even if manual feature extraction had resulted in improved performance, its practical disadvantages would still outweigh the potential benefits. The process proved to be extremely demanding in terms of computational resources, placing a heavy burden on CPU usage and memory consumption. Moreover, the extraction pipeline was highly time-consuming, requiring extensive processing time for each slide, which significantly limited scalability. These constraints make manual feature engineering difficult to justify in real-world clinical or large-scale research settings, where efficiency, reproducibility, and automation are essential.

In contrast, deep learning based feature extractors offer a more scalable and standardized solution, requiring minimal manual intervention once the pipeline is established, even more now that models pretrained on hystopatological data are becoming more common. Although they may not always capture task-specific morphological cues as explicitly as handcrafted descriptors, their ability to generalize across diverse datasets and their substantially lower computational overhead during feature extraction make them more suitable for practical deployment.

Regarding the second objective, the benchmark with simple classifiers provided an important complementary perspective on the intrinsic quality of the extracted features. By evaluating slide-level representations independently of the CLAM architecture, we were able to directly assess how informative each feature extractor was for subtype discrimination. This analysis confirmed that manually engineered features were not sufficiently discriminative in our setting, as their inclusion led to near-random performance and increased class overlap.

In contrast, features extracted by UNI proved to be highly informative when used with simple classifiers, achieving the best separability among the considered representations. This suggests that the limitation observed with UNI inside the CLAM framework is likely not due to poor feature quality, but rather to the small dataset size, which may prevent the MIL aggregator from fully exploiting the richer representations produced by such a large foundation model. Overall, these findings reinforce the idea that modern pathology-specific deep feature extractors hold strong potential, but their effective use in complex MIL models may require larger and more diverse training cohorts.

In this field, future work should focus on developing more

effective methods for manually extracting features that are also computationally efficient, as well as defining a set of the most informative and descriptive handcrafted features for distinguishing between these pathologies.

Other research directions may include improving feature aggregation strategies, exploring hybrid approaches that combine deep features with carefully selected handcrafted descriptors, and evaluating these methods on larger and more diverse datasets to better assess their generalizability.

Another possible direction is to move away from handcrafted features and instead focus on stronger foundation models that can be adapted to a wide range of tasks, not limited to classification.

REFERENCES

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

[2] L. Brcic and I. Kern, "Clinical significance of histologic subtyping of malignant pleural mesothelioma," *Translational Lung Cancer Research*, vol. 9, no. 3, pp. 924–933, 2020.

[3] G. Ali, R. Bruno, and G. Fontanini, "The pathological and molecular diagnosis of malignant pleural mesothelioma: a literature review," *Journal of Thoracic Disease*, vol. 10, no. Suppl 2, pp. S276–S284, 2018.

[4] M. Eastwood, S. T. Marc, X. Gao, H. Sailem, J. Offman, E. Karteris, A. M. Fernandez, D. Jonigk, W. Cookson, M. Moffatt, S. Popat, F. Minhas, and J. L. Robertus, "Malignant mesothelioma subtyping via sampling driven multiple instance prediction on tissue image and cell morphology data," *Artificial Intelligence in Medicine*, vol. 143, p. 102628, 2023.

[5] L. Qu, S. Liu, X. Liu, M. Wang, and Z. Song, "Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis," *Physics in Medicine & Biology*, vol. 67, no. 20, p. 20TR01, 2022.

[6] A. Riasatian, M. Babaie, D. Maleki, S. Kalra, M. Valipour, S. Hemati, M. Zaveri, A. Safarpoor, S. Shafiei, M. Afshari, M. Rasoolijaberi, M. Sikaroudi, M. Adnan, S. Shah, C. Choi, S. Damaskinos, C. J. V. Campbell, P. Diamandis, L. Pantanowitz, H. Kashani, A. Ghodsi, and H. R. Tizhoosh, "Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides," *arXiv preprint arXiv:2101.07903v1*, 2021.

[7] B. Voigt, O. Fischer, B. Schilling, C. Krumnow, and C. Herta, "Investigation of semi- and self-supervised learning methods in the histopathological domain," *Journal of Pathology Informatics*, vol. 14, p. 100305, 2023.

[8] M. Gadermayr and M. Tschuchnig, "Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential," *arXiv preprint arXiv:2206.04425v2*, 2023.

[9] 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.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] R. J. Chen, T. Ding, M. Y. Lu *et al.*, "A general-purpose foundation model for computational pathology," *Nature Medicine*, 2024.