An Attention-Guided Web Framework for Whole-Slide Image Analysis with Foundation Models and Multiple-Instance Learning

Syed Muhammad Ali Abbas
Jožef Stefan Institute
Ljubljana, Slovenia
Syed.Muhammad.Ali.Abbas@ijs.si

Matjaž Gams
Jožef Stefan Institute
Ljubljana, Slovenia
matjaz.gams@ijs.si

Sikandar Ali
King Fahd University
Dhahran, Saudi Arabia
sikandar.ali@kfupm.edu.sa

Abstract

This paper presents an attention-guided prototype developed to support laboratory practitioners in the analysis and review of whole-slide images (WSIs) for breast cancer lymph-node metastasis. The prototype allows a user to submit a WSI, process the slide using an automated analysis pipeline, obtain a slide-level classification result, and inspect spatial attention information together with the original tissue image. The analysis combines tissue segmentation and patch extraction, feature representation using the pretrained UNI foundation model, and clustering-constrained attention multiple-instance learning (CLAM). Patch-level attention values are retained with their spatial coordinates so that they can be mapped back to the corresponding regions of the WSI for visual inspection. On the evaluated CAMELYON16 and CAMELYON17 subsets, UNI+MIL achieved accuracies of 98.40% and 97.60%, respectively. The prototype brings the automated classification result and the corresponding attention information into a single environment for practical slide review.

Keywords

digital pathology, whole-slide images, foundation models, multiple-instance learning, CLAM, attention-guided visualisation

Student paper

1 Introduction

Digital pathology represents glass tissue slides as high-resolution whole-slide images (WSIs), allowing tissue to be examined and analysed computationally. These images contain a large amount of visual information and can reach gigapixel resolution, which makes both automated analysis and interactive viewing computationally demanding. In breast cancer lymph-node assessment, an analysis system therefore needs to process the slide at patch level while still allowing the resulting information to be related back to the original tissue.

This paper presents a prototype developed to support laboratory practitioners during the analysis and review of breast cancer lymph-node WSIs. The prototype provides a practical sequence in which a user submits a digital slide, the slide is processed by the analysis pipeline, and the resulting slide-level classification and spatial attention information are made available for inspection together with the original tissue image. In this way, the system connects automated WSI analysis with the visual examination of the regions that contribute to the model output.

The prototype combines established methods for computational pathology with an interactive slide-review interface. Multiple-instance learning (MIL) represents a WSI as a collection of image patches and enables classification from slide-level labels. Attention-based MIL and CLAM assign relative weights to the patch representations when producing a slide-level prediction [1, 2]. The prototype uses the pretrained UNI foundation model to obtain pathology-specific feature representations from the extracted tissue patches [3]. Related systems such as Cytomine provide broader functionality for multi-gigapixel image management, annotation, and collaborative analysis [4]. The prototype described in this paper has a more focused purpose: combining automated WSI classification with spatial attention information that can be inspected alongside the original tissue image.

The remainder of the paper is organised as follows. Section 2 reviews related work in weakly supervised whole-slide image analysis, multiple-instance learning, and pathology foundation models. Section 3 describes the architecture of the prototype and the processing stages from slide submission to result inspection. Section 4 presents the experimental setup, including the comparison of UNI, CONCH, SimCLR, and ImageNet-pretrained ResNet50 feature representations within the MIL classification setup on subsets of CAMELYON16 and CAMELYON17 [5, 6]. The subsequent sections present the experimental results, discussion, and conclusion.

2 Related Work

Weakly supervised learning has become an established approach for whole-slide image classification because it allows models to learn from slide-level labels without requiring detailed annotations for every tissue region. Campanella et al. demonstrated large-scale weakly supervised classification on whole-slide pathology images [7]. Attention-based multiple-instance learning introduced learnable weighting of individual instances within a bag [1], while CLAM extended this approach with attention-based aggregation and instance-level clustering for computational pathology [2]. Other MIL architectures have explored different ways of modelling relationships between tissue patches, including TransMIL [8] and dual-stream multiple-instance learning [9].

Pathology-specific foundation models have subsequently provided stronger feature representations for downstream slide analysis. UNI was trained on large-scale histopathology data and is used in this work as the primary feature encoder [3]. CONCH combines visual and language-based pretraining for computational pathology [10]. At the whole-slide level, Prov-GigaPath extended foundation-model pretraining to large pathology slides [11]. Recent benchmarking has also compared pathology foundation models as feature extractors for weakly supervised computational pathology tasks [12]. These studies provide the methodological context for the feature-extraction and MIL components incorporated into the prototype presented in this paper.

https://doi.org/10.70314/is.2026.skui.37

3 Prototype Architecture and Processing

Figure 1 summarises the architecture of the prototype and the sequence of operations used to analyse a WSI. The following subsections describe slide submission, tissue and patch preparation, feature extraction, slide-level classification, attention-map generation, and presentation of the resulting information for inspection.

3.1 Slide Submission and Background Processing

Analysis begins when a WSI is submitted through the prototype's Next.js web interface. The file is passed to a Flask REST API, which coordinates the analysis tasks. Celery workers handle preprocessing, feature extraction, inference, and output generation, while Redis is used for task queuing and status tracking. The frontend displays the processing status until the analysis is complete. Separating these tasks from the initial web request allows the longer image-processing steps to run in the background.

3.2 Tissue Detection and Patch Preparation

The uploaded WSI is read with OpenSlide [13]. A low-resolution representation is converted from RGB to HSV, median-filtered, and segmented using Otsu thresholding. Dilation and erosion are applied to refine the tissue mask, and OpenCV contour detection is used to remove connected regions smaller than 5,000 pixels.

Candidate patch locations are checked against the tissue mask, and patches with at least 70% tissue coverage are retained. The tissue is divided into 256×256-pixel patches. Patch coordinates, masks, and feature vectors are stored in compressed HDF5 files. Retaining the coordinates is important because the model output can later be mapped back to the corresponding locations in the original slide.

3.3 Patch Feature Extraction

The retained patches are encoded using the pretrained `MahmoodLab/UNI` checkpoint available on Hugging Face [3, 14]. UNI produces a 1024-dimensional feature vector for each patch. These vectors provide the patch-level representations used by the MIL classifier. The UNI encoder is used for feature extraction, while the downstream classifier is trained for the metastasis classification task described in Section 4.

3.4 Slide-Level Classification

CLAM aggregates the patch features to produce a slide-level representation and prediction [2]. Its attention mechanism assigns a relative weight to each patch, while the classifier produces scores for the metastatic and non-metastatic classes. During training, the instance-clustering objective provides an additional constraint on selected patch features.

The slide-level prediction and the patch attention values are retained as separate outputs. The prediction represents the classifier result for the WSI, while the attention values indicate the relative contribution of patches within the MIL model. The stored patch coordinates are then used to place these attention values back onto the slide.

Table 1: Dataset composition used in the experiments

Dataset	Metastatic	Non-metastatic	Total
CAMELYON16	100	170	270
CAMELYON17 subset	140	274	414

3.5 Attention Map and Result Inspection

After inference, the attention values are associated with their stored patch coordinates and arranged into a spatial heatmap. Warmer colours indicate higher attention values and cooler colours indicate lower values. This map provides a visual representation of how attention is distributed across the analysed tissue regions.

The original WSI and the attention map are represented as Deep Zoom Image pyramids. OpenSeadragon requests the image tiles needed for the current position and zoom level, allowing the large slide to be explored interactively in the browser. The interface presents the tissue image and attention map in adjacent views, so a user can move between the overall slide and specific regions while comparing the image content with the model's attention pattern.

3.6 Detailed Model and Heatmap Processing

The detailed model-processing stages implemented in the prototype are summarised in Figure 2. After tissue segmentation and patch extraction, the image patches are processed through the UNI-based feature-extraction stage to produce feature embeddings. The pipeline uses multiple views of the tissue regions together with alignment and reconstruction objectives during feature processing.

The resulting feature representations are passed through attention-based classification branches and subsequently used within the multiple-instance learning stage. Attention weights are assigned to the patch representations, while instance-level clustering is used as part of the CLAM-based processing. The aggregated features are then used to obtain the slide-level classification output.

Because the patch coordinates are retained throughout preprocessing, the resulting attention values can be associated with their corresponding locations in the original WSI. These values are used to construct spatial attention and prediction maps that can be viewed together with the tissue image in the browser.

4 Experimental Setup

The classification experiments used the 270 labelled CAMELYON16 training WSIs, comprising 100 metastatic and 170 normal slides. A binary slide-level subset of CAMELYON17 contained 414 WSIs, comprising 140 metastatic and 274 non-metastatic slides (Table 1). The source slides were scanned at 40× magnification [5, 6]. For CAMELYON17, the evaluation is performed at the slide-classification level.

An 80:20 train-test split was used for the principal experiments. Training used Adam with a learning rate of 2×10^{-4}, weight decay of 10^{-5}, dropout of 0.25, a maximum of 150 epochs, and early stopping based on validation performance. Experiments were implemented in Python 3.10 and PyTorch 2.0.1 with CUDA 11.7 and OpenSlide 1.2.0. Training used Linux systems with NVIDIA RTX A5000 GPUs and an RTX 3090-class workstation with 128 GB RAM.

UNI+MIL was compared with CONCH+MIL [10], SimCLR+MIL [15], and a CLAM baseline using an ImageNet-pretrained ResNet50

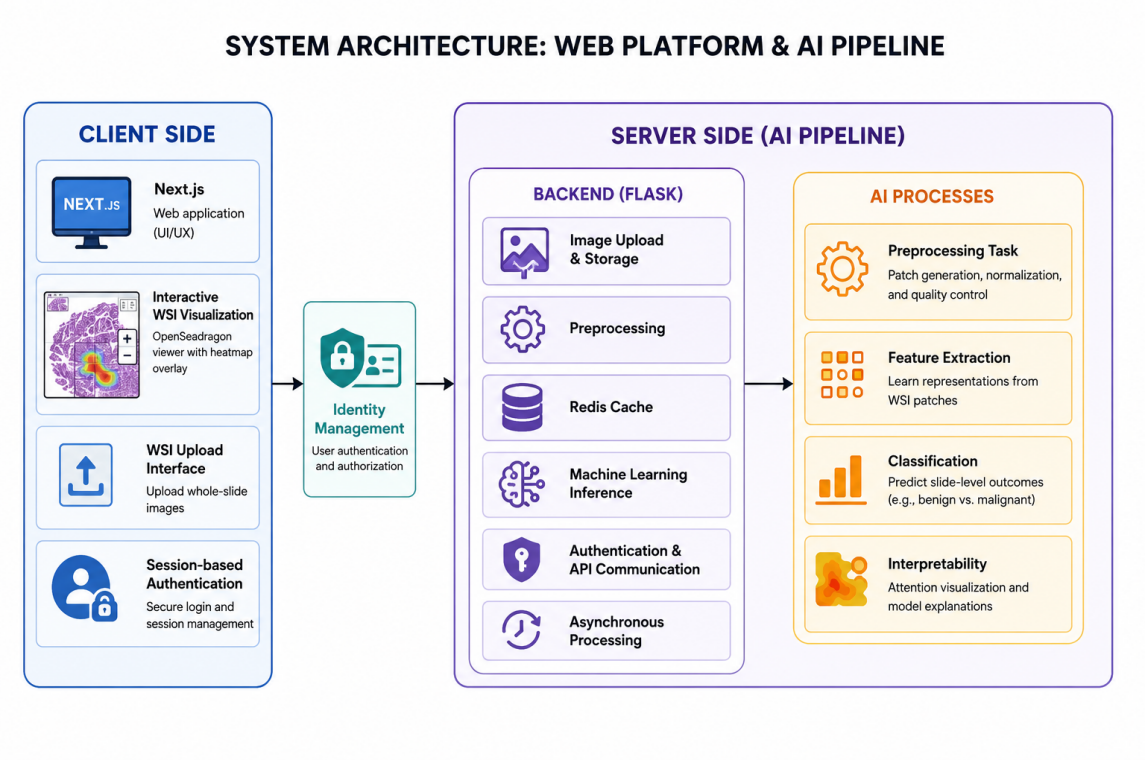

Figure 1: Architecture and processing stages of the prototype, from WSI submission through tissue preparation, feature extraction, slide-level classification, and generation of spatial attention information for review.

Table 2: Reported classification performance on the evaluated subsets (%)

Model	CAMELYON16		CAMELYON17	
	Acc.	AUC	Acc.	AUC
UNI+MIL	**98.40**	**100.00**	**97.60**	**98.44**
CONCH+MIL	97.01	97.02	95.89	94.45
SimCLR+MIL	76.66	76.16	74.09	74.09
CLAM baseline	87.20	92.20	88.29	86.81

encoder. The metrics were accuracy and area under the receiver operating characteristic curve (AUC). Sensitivity and specificity were additionally reported for UNI+MIL.

5 Results

For the evaluated subsets, UNI+MIL produced the highest accuracy and AUC among the four configurations shown in Table 2. Its accuracy was 98.40% on CAMELYON16 and 97.60% on CAMELYON17, compared with 97.01% and 95.89% for CONCH+MIL. The corresponding UNI+MIL AUC values were 100.00% and 98.44%.

UNI+MIL sensitivity and specificity were 97.50% and 99.20% on CAMELYON16, and 96.80% and 98.30% on CAMELYON17. Together, these results describe the performance of the evaluated feature representations within the slide-level MIL classification setup used in this study.

The classification output is connected to the prototype's visual inspection component through the stored patch coordinates and attention values. After inference, these values are mapped back to their corresponding locations in the WSI and used to generate the spatial attention representation described in Section 3.5.

6 Discussion

The experiments compare several feature representations within the same MIL-based slide-classification setup. The reported metrics differ across the tested representations, with UNI producing the highest accuracy and AUC values on the evaluated CAMELYON16 and CAMELYON17 subsets.

An important part of the prototype is the connection between the slide-level classification result and the spatial information retained during patch-based analysis. Because patch coordinates are preserved during preprocessing, the corresponding attention values can be returned to their original positions in the WSI. The resulting attention map allows the model output to be inspected in relation to the underlying tissue rather than being presented only as a slide-level class label.

The prototype separates computationally intensive analysis from interactive slide inspection. Background tasks perform tissue preparation, feature extraction, and model inference, while the Deep Zoom representation allows the interface to retrieve only the image tiles required for the current view. This architecture allows

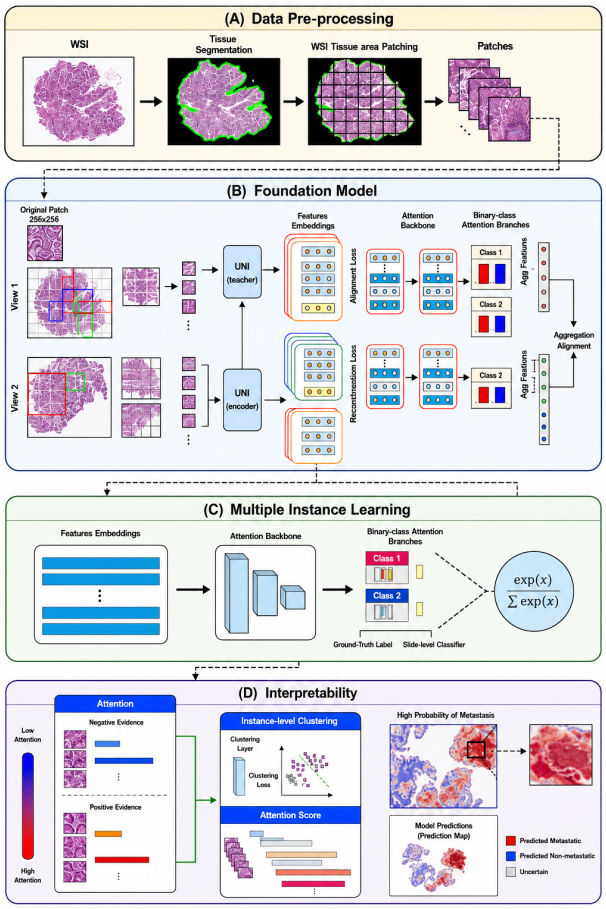

Figure 2: Model-processing and attention-visualisation pipeline: (A) WSI preprocessing and patch extraction, (B) UNI-based feature processing, (C) multiple-instance learning, and (D) attention and prediction-map generation.

large WSIs and their corresponding attention maps to be explored without loading the complete slide image into the interface at once.

The reported results are based on the specified CAMELYON subsets and the experimental split described in Section 4. Future work can evaluate the prototype using additional datasets and acquisition conditions and can examine how the presentation of classification and attention information supports laboratory practitioners during slide review.

7 Conclusion

This paper presented an attention-guided prototype developed to support the analysis and review of breast cancer lymph-node whole-slide images. The prototype integrates slide submission, background processing, tissue and patch preparation, UNI-based feature extraction, CLAM-based slide classification, and spatial attention visualisation within a common system. In the reported experiments, UNI+MIL achieved 98.40% accuracy on the evaluated CAMELYON16 subset and 97.60% on the evaluated CAMELYON17 subset. By preserving the spatial coordinates of analysed patches, the prototype can map model attention back to the original WSI and present this information together with the tissue image. The resulting system provides laboratory practitioners with both the slide-level classification result and spatial model information that can be inspected during digital slide review.

References

[1] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136, 2018.

[2] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021.

[3] Richard J. Chen et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024.

[4] Raphaël Marée et al. Collaborative analysis of multi-gigapixel imaging data using cytomine. *Bioinformatics*, 32(9):1395–1401, 2016.

[5] Babak Ehteshami Bejnordi et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[6] Peter Bandi et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019.

[7] Gabriele Campanella et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25:1301–1309, 2019.

[8] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer-based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021.

[9] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, Nashville, TN, USA, 2021. IEEE.

[10] Ming Y. Lu et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[11] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024.

[12] Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, et al. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature Biomedical Engineering*, 10(6):1113–1123, 2026.

[13] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, 4:27, 2013.

[14] Mahmood Lab. UNI: Pretrained pathology foundation model (model card and weights). Hugging Face model repository, https://huggingface.co/MahmoodLab/UNI. Accessed: 2026-09-24.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607, 2020.