



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

AGE-MIL: Anchor-Guided Evidence Learning for Patient-Level Prediction

Jiawei Niu^{1†}, Jian Chen^{2†}, Di Zhang¹, Junbo Lu¹, Zhangcheng Liao³, Xuhao Liu³, Honglin Zhong³, Mireia Crispin-Ortuzar², Chen Li¹, Zeyu Gao^{2(✉)}, and Yi Cai^{3(✉)}

¹ School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China

² Department of Oncology, University of Cambridge, Cambridge, UK

³ Xiangya School of Medicine, Central South University, Changsha, Hunan, China
zg323@cam.ac.uk; cai-yi@csu.edu.cn

Abstract. Existing computational pathology methods predominantly operate within whole-slide image (WSI)-level multiple instance learning (MIL) paradigms, while patient-level modeling remains underexplored. In routine pathological practice, however, pathologists derive diagnostic and prognostic conclusions by integrating evidence across multiple WSIs rather than relying on any single slide. This discrepancy creates a fundamental misalignment when patient-level supervision is directly imposed on conventional MIL frameworks, often leading to unstable optimization and degraded predictive reliability. To address this issue, we propose Anchor-Guided Evidence MIL (AGE-MIL), a weakly supervised framework for patient-level prediction. AGE-MIL constructs a patient-level anchor from slide representations to capture global pathological context and guide the retrieval and integration of diagnostically relevant local patches, enabling robust patient-level modeling. Patient-level risk is further modeled as an evidence accumulation process, promoting stable optimization under weak supervision. AGE-MIL is evaluated on six clinically relevant patient-level prediction tasks from two independent cohorts. Experimental results show that the proposed framework consistently outperforms eight state-of-the-art MIL methods. Code is available at <https://github.com/wodeniua/AGE-MIL>.

Keywords: Multiple instance learning · Patient-level prediction · Whole slide image

1 Introduction

Computational pathology has achieved substantial progress in tumor diagnosis and prognosis prediction [1, 8]. Due to the gigapixel resolution of whole slide images (WSIs) and the scarcity of fine-grained annotations, multiple instance learning (MIL) has become the prevailing modeling paradigm, where WSIs are

[†] Contributed equally to this work.

treated as bags and patches as instances [10, 16, 13, 19]. This formulation has been extensively validated in slide-level prediction tasks, establishing MIL as a powerful paradigm for analyzing large-scale pathological data.

Clinical decision-making, however, is inherently patient-centric [4]. A single patient is typically associated with multiple heterogeneous WSIs, and patient-level conclusions rely on evidence integrated across multiple WSIs rather than any single slide [1, 5]. Directly extending slide-level MIL to patient-level prediction introduces a fundamental mismatch between supervision granularity and evidence distribution [25]. In particular, patch-to-patient aggregation is computationally prohibitive due to the massive number of instances and the sparse distribution of diagnostic evidence, while also disrupting the patch–slide–patient hierarchy and discarding slide-level structural information as shown in Fig. 1. While recent slide-level foundation models provide semantically consistent and transferable representations [22, 24], they primarily operate at the slide level and do not explicitly address patient-level evidence integration. Consequently, directly aggregating slide-level features still risks discarding critical evidence, limiting the ability of existing methods to effectively capture and integrate sparse and distributed patient-level evidence.

To address these challenges, we propose AGE-MIL, an anchor-guided evidence multi-instance learning framework for patient-level prediction. AGE-MIL introduces a patient-level anchor to capture global pathological context and guide the integration of local evidence across multiple slides. Specifically, slide-level representations extracted by the pathology foundation model are aggregated to construct a patient-level anchor with stable semantic characterization. This anchor is then used to guide evidence retrieval from large instance (patch) pools through patient-aware scoring. The retrieved instances are subsequently integrated via an evidence-aware representation learning module, enabling adaptive refinement of patient representations without requiring additional supervision. Furthermore, patient-level risk is modeled as a progressive evidence accumulation process, promoting stable optimization and robust prediction.

We evaluate the proposed framework on two large-scale prostate cancer biopsy datasets comprising over 700 patients and approximately 10,000 WSIs from two independent clinical cohorts, covering six clinically relevant patient-level prediction tasks, including lymph node metastasis and prognosis prediction. We further conduct comprehensive patient-level benchmarking against eight state-of-the-art MIL methods. Experimental results show that AGE-MIL improves AUC and accuracy by approximately 2% on average across the six tasks, establishing state-of-the-art performance.

2 Related Work

Multiple Instance Learning. MIL has become a dominant framework in computational pathology [10, 27], largely motivated by the gigapixel resolution of WSIs and the lack of fine-grained annotations [20]. Within this paradigm, WSIs are modeled as bags and image patches as instances, with embedding-level

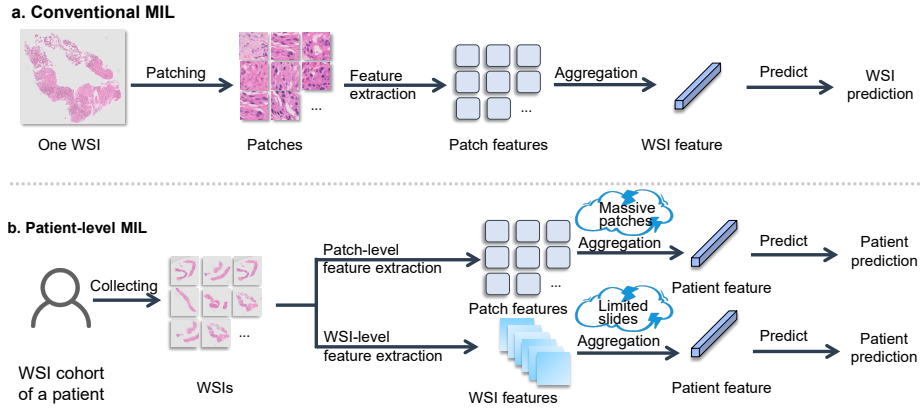


Fig. 1. Conventional WSI-level MIL vs. patient-level MIL. Patient-level aggregation can be performed either via patch-to-patient or slide-to-patient.

and attention-based aggregation strategies [10] widely adopted for slide-level prediction. However, conventional MIL formulations typically rely on slide-level supervision, implicitly assuming that each bag corresponds to a single WSI [13, 14, 17]. This assumption becomes problematic in patient-level prediction settings. Assigning patient-level labels to individual slides or instances often introduces supervision ambiguity. Moreover, diagnostically relevant pathological evidence is frequently sparse, creating a structural mismatch with aggregation mechanisms that depend on densely informative instances. Such discrepancies may lead to noisy gradients and the dominance of spurious instances. Although hierarchical MIL variants have been explored [9, 2, 28], challenges related to supervision noise and evidence sparsity remain insufficiently addressed.

Pathology Foundation Models. Foundation models have emerged as a dominant paradigm in computational pathology, largely driven by large-scale pre-training at the patch and slide levels. These approaches typically combine high-capacity visual encoders with slide-level aggregation mechanisms to learn transferable representations from extensive WSI collections. Representative directions include attention-based aggregation [22], coordinate-aware modeling [6, 26], long-sequence architectures [24], and cross-modal alignment [18, 11, 7]. Pretrained representations have been shown to improve semantic consistency and generalization across downstream tasks. Existing studies primarily focus on encoder design and pretraining strategies, while the role of pretrained features in downstream inference remains less explored. Under weak supervision and patient-level prediction settings, semantically consistent representations naturally enable alternative modeling strategies that emphasize evidence selection and information aggregation rather than instance-level representation learning.

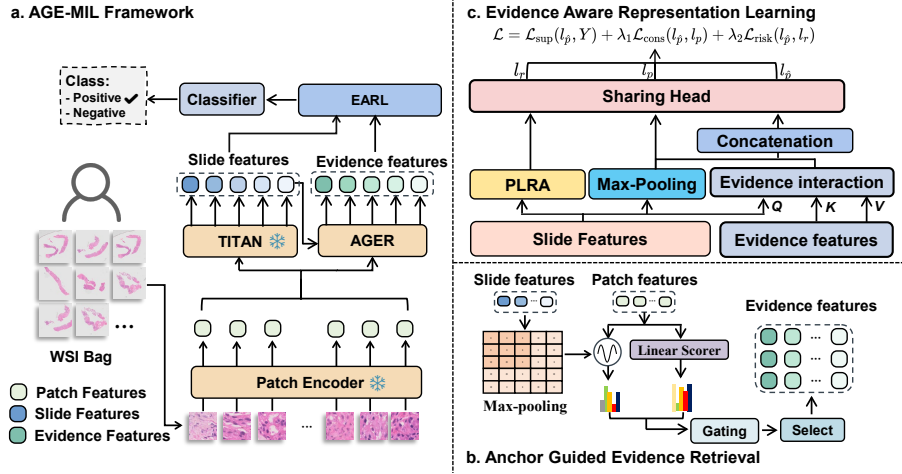


Fig. 2. Overview of AGE-MIL. a) AGE-MIL Framework: Patch features are aggregated into slide-level features using TITAN, then processed via AGER for evidence feature retrieval or EARL for patient-level interaction with evidence features. b) AGER: Guided by the anchor, selects evidence features from a large pool of patches. c) EARL: Patient-conditioned evidence interaction enhances patient-level modeling.

3 Method

The overall framework is illustrated in Fig. 2 and consists of three modules: (1) Anchor Guided Evidence Retrieval (AGER), which identifies diagnostically relevant patches; (2) Evidence Aware Representation Learning (EARL), which integrates patient-level and evidence representations; and (3) Patient Level Risk Accumulation (PLRA), which estimates patient-level risk.

3.1 Problem Formulation

Consider a patient-level dataset containing N patients, denoted as $\mathcal{D} = \{(\mathcal{S}, y)\}$, where $y \in \{0, 1\}$ indicates the patient label and $\mathcal{S} = \{S\}$ represents the set of WSIs associated with a patient. Each slide is decomposed into image patches, yielding an instance set $\mathcal{X} = \{x \in \mathbb{R}^D\}$.

The objective is to learn a mapping $f(\mathcal{S}, \mathcal{X}) \rightarrow \hat{y}$ that predicts the patient label by leveraging both slide-level context and patch-level evidence. This formulation captures the hierarchical nature of pathological data, where diagnostically relevant patterns may be sparsely distributed across slides and instances.

3.2 Anchor Guided Evidence Retrieval

For each patient i , patch-level features are first extracted using a patch-level foundation model, followed by slide-level feature aggregation using a slide-level

foundation model. We infer a patient-level anchor $p_i \in \mathbb{R}^{D_p}$ as a latent global state from the set of slide features. In our implementation, this inference is realized by a max aggregation operator, which serves as a robust estimator of the latent state in small-bag patient settings, where attention-based aggregation may exhibit sensitivity to instance variability. This choice is further supported by prior findings that informative instances often correspond to larger feature magnitudes [3, 12], making max aggregation a simple yet effective mechanism for capturing salient evidence.

Instead of aggregating all patches within a patient’s WSIs, we formulate evidence selection as a patient-conditioned retrieval process and extract a compact subset of task-relevant patches as diagnostic evidence. Each patch embedding x is assigned a retrieval score with two complementary components: an unconditional term $s^u = w_u^\top x$ that captures intrinsic instance saliency, and a conditional term $s^c = \langle \phi(x), \psi(p) \rangle / \tau$ that measures anchor–instance compatibility. Here, $\phi(\cdot)$ and $\psi(\cdot)$ are learnable projections and τ is a temperature parameter. The final score is defined as:

$$s = (1 - \alpha)s^u + \alpha s^c, \quad \alpha = \sigma(g(p)), \quad (1)$$

where α adaptively trades off generic saliency and anchor-consistent relevance. Top- K patches ranked by $\{s\}$ form the evidence set $\mathcal{E}$.

3.3 Evidence Aware Representation Learning

To perform patient-conditioned evidence interaction, we employ a cross-attention operator defined as:

$$\text{Attn}(\mathcal{H}, \mathcal{E}) = \text{softmax} \left(\frac{\mathcal{H}\mathcal{E}^\top}{\sqrt{d}} \right) \mathcal{E}, \quad (2)$$

where $\mathcal{H}$ denotes the set of slide-level features for a patient, $\mathcal{E}$ represents the retrieved evidence embeddings, and d is the dimension of the evidence feature space. This operation yields context-conditioned evidence representations.

To regulate evidence reliability, we introduce a learnable gating mechanism applied to each evidence token, $\hat{e} = \gamma \cdot e$, where the gate γ is predicted from the evidence feature itself. The gated evidence tokens are aggregated via a maxpooling operator to obtain a compact evidence summary z^{evi} .

Finally, the anchor and evidence representations are jointly integrated. They are concatenated as $[p, z^{\text{evi}}]$, followed by a linear transformation: $\hat{p} = f_{\text{fuse}}([p, z^{\text{evi}}])$, where f_{fuse} denotes the fusion mapping.

Gradients are detached from the anchor to prevent retrieval-induced anchor drift. This design preserves a stable global patient context while allowing retrieved evidence to refine the final representation.

3.4 Patient Level Risk Accumulation

The patient-level risk logits r are derived from slide-level logits $\{l_j\}_{j=1}^m$ using a Log-Mean-Exp (LME) aggregation:

$$r = \log \sum_{j=1}^m \exp(l_j) - \log m, \quad (3)$$

This formulation admits a natural interpretation from a log-evidence accumulation perspective. Specifically, $\exp(l_j)$ can be understood as a non-negative evidence contribution associated with slide j , reflecting its relative support for patient-level risk. Under this view, the aggregated risk corresponds to the logarithm of accumulated evidence across slides. The normalization term $-\log m$ compensates for variations in slide counts, thereby reducing bias induced by differing numbers of observations.

Compared with hard max pooling, LME yields a smooth approximation that retains sensitivity to dominant high-risk slides while preserving stable gradient behavior. The logarithmic form further improves numerical stability and is consistent with standard probabilistic interpretations of logits.

The overall training objective combines supervised learning, representation consistency, and risk alignment:

$$\mathcal{L}_{\text{sup}} = \text{CE}(l_{\hat{p}}, Y), \mathcal{L}_{\text{cons}} = \text{KL}(l_{\hat{p}}, l_p), \mathcal{L}_{\text{risk}} = T^2 \text{KL}\left(\frac{l_r}{T}, \frac{\text{stopgrad}(l_{\hat{p}})}{T}\right), \quad (4)$$

where $\mathcal{L}_{\text{sup}}$ denotes the supervised classification loss, $\mathcal{L}_{\text{cons}}$ enforces predictive consistency between the evidence-aware representation and the patient anchor, and $\mathcal{L}_{\text{risk}}$ regularizes the risk branch by aligning it with the primary prediction. The coefficients λ_1 and λ_2 control the relative contributions of each term. We use $l_{\hat{p}}$ for the evidence-aware prediction logits, l_p for the patient anchor logits, l_r for the risk logits, and Y for the patient label. Thus, the overall loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_1 \mathcal{L}_{\text{cons}} + \lambda_2 \mathcal{L}_{\text{risk}}. \quad (5)$$

4 Experiments and Results

Feature Extraction. Both slide-level and patch-level features are extracted at $20\times$ magnification with an input resolution of 512×512 using the TRIDENT tool [21]. The patch-level features are extracted using CONCH v1.5 [15], and these features are aggregated into slide-level representations using TITAN [6].

Dataset. We evaluate the proposed framework on two independent patient-level pathology datasets collected from the same medical institution. The datasets consist of biopsy WSIs acquired from routine clinical practice, with patient-level labels defined according to corresponding clinical outcomes. All data were

Table 1. Performance comparison on three metastasis prediction tasks.

Method	LNM (Any)		LNM (Distant)		LNM (Region)	
	ACC	AUC	ACC	AUC	ACC	AUC
MeanMIL	78.10±6.21	81.06±5.71	81.02±1.69	80.67±4.35	78.06±4.35	77.44±10.58
MaxMIL	77.57±5.03	82.13±4.63	79.94±2.58	81.62±3.88	76.98±4.63	76.12±10.68
ABMIL	78.35±6.66	81.60±6.62	79.95±0.95	80.87±4.64	76.74±4.91	78.68±10.19
CLAM-SB	77.04±5.04	81.36±6.20	81.29±1.28	80.85±5.04	76.48±6.19	76.01±14.65
CLAM-MB	79.15±5.81	81.67±5.78	81.28±1.94	79.93±5.52	77.26±4.01	78.77±8.24
DSMIL	73.09±2.33	80.35±6.09	77.81±1.55	73.81±9.49	73.81±1.06	78.49±9.70
TransMIL	78.28±3.02	81.87±4.02	81.56±2.78	80.39±4.81	77.32±6.54	76.73±10.12
ILRA	77.29±4.97	80.51±3.78	81.54±3.99	79.42±6.81	74.61±2.74	74.18±6.62
AGE-MIL	80.99±5.54	82.09±5.12	82.35±2.27	82.24±2.30	78.58±5.42	80.64±7.73

collected under approval from the Institutional Review Board (IRB) of the institution, and the requirement for informed consent was waived due to the retrospective and anonymized nature of the study.

Prostate Metastasis Prediction Dataset: This dataset is designed for lymph node metastasis (LNM) risk modeling and is formulated as three independent binary classification tasks, defined according to metastasis patterns: distant metastasis, regional metastasis, and the presence of either distant or regional metastasis. It comprises 379 patients, corresponding to 6,476 WSIs. At the instance level, approximately 1.26M patches are sampled from tissue regions for model training and evidence modeling.

Prostate Prognostic Prediction Dataset: This dataset is constructed for prognosis modeling associated with androgen receptor signaling inhibitors (ARSIs) and is similarly organized into three independent binary classification tasks. The tasks are defined based on three clinically relevant PSA thresholds (0.2, 2.0, and 4.0 ng/mL), corresponding to biochemical recurrence, disease progression, and high-risk disease, respectively. The dataset contains 306 patients and 3,402 WSIs, from which approximately 1.20M patches are extracted.

For both datasets, patients serve as the primary prediction unit, with labels defined at the patient level. Each patient may correspond to multiple WSIs. All slides undergo tissue-aware patch sampling to construct instance sets for weakly supervised modeling.

Comparison with MIL methods. We compare the proposed method against a diverse set of representative MIL approaches, including MeanMIL, MaxMIL, ABMIL [10], CLAM-SB/MB [16], DSMIL [13], TransMIL [19], and ILRA [23]. To adapt these MIL baseline models to the patient-level prediction scenario, we aggregate the slide features of each patient extracted by TITAN.

Performance is evaluated using classification accuracy (ACC) and the area under the receiver operating characteristic curve (AUC). To ensure robust and unbiased assessment, we adopt a five-fold cross-validation on all datasets.

From Table 1 and Table 2, our method consistently outperforms state-of-the-art approaches across both datasets and tasks. In the LN dataset, our method

Table 2. Performance comparison on three prognostic prediction tasks.

Method	ARSI (0.2)		ARSI (2)		ARSI (4)	
	ACC	AUC	ACC	AUC	ACC	AUC
MeanMIL	66.80±8.90	63.24±11.77	74.80±1.79	56.08±9.69	82.00±2.45	62.58±8.08
MaxMIL	66.80±8.67	62.41±14.27	74.40±1.67	62.48±6.72	81.20±2.28	61.93±11.97
ABMIL	66.40±9.32	61.68±13.02	74.40±1.67	59.04±8.83	82.40±2.19	62.48±10.67
CLAM-SB	67.20±8.44	64.11±10.89	75.20±1.10	55.73±8.41	82.00±1.41	58.85±10.28
CLAM-MB	66.40±9.10	65.56±12.63	73.60±1.67	60.23±3.75	82.00±2.45	60.17±10.79
DSMIL	65.60±6.39	63.31±9.19	73.20±1.79	58.80±6.38	80.40±0.89	60.86±8.36
TransMIL	62.80±5.22	63.40±6.21	74.00±1.41	53.02±10.77	81.20±1.79	58.28±11.50
ILRA	62.80±3.35	64.46±7.58	75.20±3.35	56.11±12.26	82.00±1.41	62.18±8.52
AGE-MIL	66.40±4.28	67.67±8.82	75.40±1.72	62.57±7.75	80.80±2.15	62.71±8.32

Table 3. Ablation study of the proposed method on three LNM datasets. I: Evidence concatenation, II: Cond-Sim, III: StopGrad, IV: $\mathcal{L}_{\text{risk}}$, V: $\mathcal{L}_{\text{cons}}$.

Method	LNM (Any)		LNM (Distant)		LNM (Region)	
	ACC	AUC	ACC	AUC	ACC	AUC
w/o I	77.10±6.50	78.30±6.20	79.05±3.05	78.80±3.00	75.20±6.30	76.90±8.90
w/o II	78.20±6.25	79.40±6.00	80.05±2.90	79.85±2.85	76.35±6.10	78.00±8.60
w/o III	79.85±5.90	81.15±5.40	81.60±2.48	81.45±2.43	77.70±5.70	79.70±8.00
w/o IV	79.30±6.05	80.55±5.60	81.10±2.60	80.95±2.55	77.20±5.85	79.10±8.25
w/o V	79.75±5.95	81.05±5.45	81.55±2.50	81.35±2.45	78.65±5.75	79.60±8.05
AGE-MIL	80.99±5.54	82.09±5.12	82.35±2.27	82.24±2.30	78.58±5.42	80.64±7.73

improves AUC by 1-3% over the best competitor, MaxMIL, while in the ARSI dataset, we achieve a notable AUC improvement of over 4% for ARSI1 and ARSI2 compared to CLAM-MB. These results demonstrate the superior performance and robustness of our approach across various tasks and datasets.

Ablational Studies. We perform an ablation study to assess the contribution of the proposed components, as reported in Table 3. The analysis considers variants without the evidence concatenation, the conditional similarity retrieval mechanism (Cond-Sim), the stop-gradient strategy, and the introduced loss terms.

The observations indicate that all components positively influence model performance. Excluding the evidence pathway leads to clear degradation across datasets, highlighting the benefit of incorporating retrieved evidence. Removing the conditional similarity mechanism likewise reduces accuracy, suggesting that patient-aware retrieval is important for identifying relevant instances. Notably, disabling the stop-gradient strategy results in inferior performance, implying its role in maintaining stable retrieval behavior. In addition, eliminating the proposed loss terms further weakens the results, reflecting their complementary effects during optimization.

5 Conclusion

In this paper, we presented AGE-MIL, a multiple-instance learning framework for patient-level prediction. By explicitly modeling patient-level context and evidence integration, AGE-MIL addresses key challenges in patient-level pathology analysis, including sparse diagnostic evidence, multi-slide heterogeneity, and weak supervision. Comprehensive benchmarking against eight state-of-the-art MIL methods across six patient-level tasks on two independent datasets demonstrates the superior performance and strong generalizability of AGE-MIL.

Acknowledgments. This work has been supported in part by the Key Research and Development Program of Shaanxi Province (2025SF-YBXM-363, 2024SF-GJHX-32), the Outstanding Youth Fund Project of Hunan Province (2024JJ2090), the Medical Discipline Construction Project of Hunan Province, the National Natural Science Foundation of China (82272907, 81974397), the Natural Science Foundation of Hunan Province (2025JJ60616), the National Science and Technology Major Project (2025ZD0544802, 2024ZD0527700). All clinical data used in this study were approved by the institutional review board, with the ethical approval number: 2026010083.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25**(8), 1301–1309 (2019)
2. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 16144–16155 (2022)
3. Chikontwe, P., Kim, M., Jeong, J., Sung, H.J., Go, H., Nam, S.J., Park, S.H.: Fr-mil: Distribution re-calibration-based multiple instance learning with transformer for whole slide image classification. *IEEE Transactions on Medical Imaging* **44**(1), 409–421 (2024)
4. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al.: Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine* **25**(10), 1519–1525 (2019)
5. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* **89**(1-2), 31–71 (1997)
6. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: A multimodal whole-slide foundation model for pathology. *Nature medicine* pp. 1–13 (2025)
7. Gao, Z., He, K., Su, W., Machado, I.P., McGough, W., Jimenez-Linan, M., Rous, B., Wang, C., Li, C., Pang, X., et al.: Alpaca: Adapting llama for pathology context analysis to enable slide-level question answering. *medRxiv* pp. 2025–04 (2025)

8. Gao, Z., Mao, A., Dong, Y., Clayton, H., Wu, J., Liu, J., Wang, C., He, K., Gong, T., Li, C., et al.: Smmile enables accurate spatial quantification in digital pathology using multiple-instance learning. *Nature Cancer* pp. 1–17 (2025)
9. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: H²-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 36, pp. 933–941 (2022)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *International conference on machine learning*. pp. 2127–2136. PMLR (2018)
11. Jaume, G., Oldenburg, L., Vaidya, A., Chen, R.J., Williamson, D.F., Peeters, T., Song, A.H., Mahmood, F.: Transcriptomics-guided slide representation learning in computational pathology. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9632–9644 (2024)
12. Lee, P., Wang, J., Lu, Y., Byun, H.: Weakly-supervised temporal action localization by uncertainty modeling. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 35, pp. 1854–1862 (2021)
13. Li, B., Li, Y., Eliceiri, K.W.: 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*. pp. 14318–14328 (2021)
14. Liu, J., Mao, A., Niu, Y., Zhang, X., Gong, T., Li, C., Gao, Z.: Pamil: Prototype attention-based multiple instance learning for whole slide image classification. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 362–372. Springer (2024)
15. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., et al.: A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024)
16. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**(6), 555–570 (2021)
17. Niu, Y., Liu, J., Zhan, Y., Shi, J., Chen, J., Zhang, D., Li, C., Gao, Z.: Learning heterogeneous embedding with prototype-aware graph attention for whole slide image classification. In: *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. pp. 2671–2678. IEEE (2025)
18. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., et al.: Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254* (2024)
19. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* **34**, 2136–2147 (2021)
20. Tellez, D., Litjens, G., Van der Laak, J., Ciompi, F.: Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence* **43**(2), 567–578 (2019)
21. Vaidya, A., Zhang, A., Jaume, G., Song, A.H., Ding, T., Wagner, S.J., Lu, M.Y., Doucet, P., Robertson, H., Almagro-Perez, C., et al.: Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652* (2025)
22. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**(8035), 970–978 (2024)

23. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023)
24. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**(8015), 181–188 (2024)
25. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis* **65**, 101789 (2020)
26. Zhang, D., Gong, Z., Pang, X., Liu, J., Lu, J., Cui, H., Ge, J., Zeng, Z., Yi, K., Li, Y., Liu, S., Yu, T., Wang, H., Crispin-Ortuzar, M., eimiao Yu, Li, C., Gao, Z.: Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis (2026), <https://arxiv.org/abs/2602.21637>
27. Zhang, Y., Gao, Z., He, K., Li, C., Mao, R.: From patches to wsis: A systematic review of deep multiple instance learning in computational pathology. *Information Fusion* **119**, 103027 (2025)
28. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kollachalama, V.B.: A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging* **41**(11), 3003–3015 (2022)