

Weakly Supervised Multiple Instance Learning for Benign–Malignant Classification of Parotid Gland Tumors on Whole-Slide Images

Puihei Lee^{1a}, Qisheng Rao^{1a*}, Fungying Li^{1b*}, Yingying Feng^a, Tianxin Wang^a, Guo quan Zhang^{a*}, Hongtao Jin^{c*}

Affiliations:

^a Medical Center of Stomatology, Shenzhen People's Hospital (The First Affiliated Hospital, Southern University of Science and Technology; The Second Clinical Medical College, Jinan University), Shenzhen 518020, China.

^b School of Software Engineering, Sun Yat-sen University, 519082, Zhuhai, China.

^c Department of Pathology, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen 518020, Guangdong, China.

^{1*}: These authors contributed equally to this work and should be considered co-first authors.

* Corresponding authors. E-mail addresses:

Guoquan Zhang : zhang.guoquan@szhospital.com;

Hongtao Jin : 807697204@qq.com

Declaration of Competing Interest

This study was supported by by SanmingProject of Medicine in Shenzhen (No.SZSM202311009).

Acknowledgment

None declared.

Competing interests

The authors declare no competing interests.

Weakly Supervised Multiple Instance Learning for Benign–Malignant Classification of Parotid Gland Tumors on Whole-Slide Images

Abstract:

Background: Accurate differentiation between benign and malignant parotid gland tumors (PGTs) is important for diagnosis and treatment planning. Although histopathologic examination of hematoxylin and eosin (H&E)-stained slides remains the diagnostic reference standard, whole-slide image (WSI) assessment is labor-intensive and may be challenging in histologically heterogeneous lesions. Weakly supervised multiple instance learning (MIL) enables WSI-level classification using case-level diagnostic labels without requiring region-level annotation.

Methods: We retrospectively analyzed 101 patients with PGTs, including 58 benign and 43 malignant cases, represented by 189 digitized WSIs. Patient-level stratified five-fold cross-validation was performed to ensure that all WSIs from an individual patient were assigned exclusively to one fold. Following automated tissue detection and quality-controlled patch extraction, 256×256 -pixel patches were encoded using either ResNet-50 or the pathology-pretrained CTransPath backbone. Three MIL strategies—CLAM-SB, attention-based MIL (ABMIL), and mean pooling—were evaluated. Model performance was assessed using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, F1-score, and area under the precision–recall curve.

Results: CTransPath + CLAM-SB yielded the highest mean AUC among the evaluated configurations (0.83), with an accuracy of 0.81 sensitivity of 0.82, specificity of 0.80, and F1-score of 0.80. Under the CLAM-SB framework, CTransPath showed a higher mean AUC than ResNet-50 (0.83 vs. 0.81). Using CTransPath features, CLAM-SB achieved higher mean AUC than ABMIL (0.80) and mean pooling (0.77). However, pairwise differences between models did not reach statistical significance. The optimized CLAM-SB configuration with a bag loss weight of 0.3 further increased the mean AUC to 0.86.

Conclusions: Weakly supervised MIL enabled benign–malignant classification of PGTs from routine H&E WSIs without region-level annotation. In this retrospective single-center cohort, CTransPath combined with CLAM-SB showed the most favorable overall performance, with a consistent trend toward better discrimination across models. Given the limited sample size, these findings should be regarded as preliminary and warrant validation in larger, multicenter cohorts with external validation to confirm their generalizability and clinical utility.

Keywords: Parotid gland tumors (PGT); Whole Slide Imaging (WSI); Artificial intelligence; Multiple instance learning (MIL); CTransPath ; CLAM-SB

Introduction

Parotid gland tumors (PGTs) are highly heterogeneous head and neck neoplasms that pose considerable diagnostic challenges^[1]. Benign lesions carry favorable outcomes after complete excision, whereas malignant subtypes (e.g., adenoid cystic carcinoma, mucoepidermoid carcinoma) are prone to perineural invasion, recurrence, and distant metastasis^[2-4]. Thus, accurate discrimination between benign and malignant PGTs is essential for treatment planning and prognostic evaluation^[5-7].

Given the marked histological diversity of PGTs, which encompasses over 20 subtypes according to the WHO classification, their biological behavior is difficult to determine^[8, 9]. Malignant transformation and histopathological patterns differ greatly across subtypes. Although pleomorphic adenoma is the most common benign lesion, borderline lesions and rare malignant subtypes often share overlapping cytological and architectural features^[10]. Consequently, histological differentiation of benign from malignant PGTs remains heavily reliant on clinical expertise.

Traditional H&E-stained slide evaluation is subjective and labor-intensive, with significant inter-observer variability, particularly in borderline or low-grade lesions^[11]. Therefore, automated WSI-based assessment offers potential for standardizing diagnosis and reducing pathologist workload.

With the digital transformation of pathology, deep learning has emerged as a powerful tool in computational pathology^[12], achieving performance in common cancers such as breast, lung, and prostate^[13-15]. However, AI applications in PGTs remain limited, largely due to the remarkable histological diversity and diagnostic complexity of these neoplasms. Existing studies predominantly rely on preoperative radiomics rather than histopathological WSIs^[16], and the few WSI-based studies have several limitations: (1) reliance on small-scale, manually annotated regions of interest (ROIs) requiring pixel-level labeling^[17]; (2) single-center cohorts with limited coverage of rare subtypes^[18]; and (3) dependence on standard CNN architectures without addressing spatial context or global feature aggregation in WSIs.

Given the critical role of accurate histological differentiation in dictating PGT treatment strategies, our study addresses these limitations by leveraging weak-supervised multiple instance learning (MIL)^[19, 20]. Using only slide-level or case-level diagnostic labels, we develop the deep learning framework designed to classify PGTs on full-field H&E-stained WSIs. Furthermore, we systematically evaluate the impact of various pathology-specific feature extractors and MIL aggregation strategies, aiming to provide an accurate, interpretable, and clinically applicable tool for PGT diagnosis.

Materials and Methods

Ethical approval

This study was reviewed and approved by the Ethics Committee of Shenzhen People's Hospital (Approval No.: LL-KY-2025066-01). All procedures were performed in compliance with the Declaration of Helsinki and applicable national legal requirements.

Study cohort and histopathologic reference standard

Patients with histopathologically diagnosed parotid gland tumors who underwent surgical treatment at Shenzhen People's Hospital between 2021 and 2025 were retrospectively identified. Archived H&E-stained histologic slides were retrieved from the institutional pathology archive.

All eligible cases were independently reviewed by two senior pathologists, each with at least 10 years of experience. Cases were included when the histopathologic diagnoses were concordant between the two reviewers. Slides with severe tissue folding, extensive artifacts, substantial tissue loss, inadequate staining, or other technical defects that could compromise histopathologic interpretation were excluded. Patient identifiers were removed before digital image processing and computational analysis.

The final cohort comprised 101 patients, including 58 benign and 43 malignant cases, represented by 189 digitized WSIs. The detailed histopathologic composition of the cohort is summarized in Table 1.

Table 1. Summary of pathological types of PGT

Benign parotid tumors	Number of cases
Pleomorphic adenoma	29
Warthin tumor	18
Others	11
Total	58
Malignant parotid tumors	Number of cases
Adenoid cystic carcinoma	7
Mucoepidermoid carcinoma	14
Acinic cell carcinoma	5
Sarcoma	3
Carcinoma ex pleomorphic adenoma	3
Others	11
Total	43

Other benign lesions included basal cell adenoma, lymphoid hyperplasia, reactive hyperplasia, branchial cleft cyst, lymphoepithelial cyst, and myoepithelioma. Sarcoma included chondrosarcoma, osteosarcoma, and fibrosarcoma. Other malignant tumors included basal cell carcinoma, squamous cell carcinoma, and melanoma.

Patient-level cross-validation and experimental design

A patient-level stratified five-fold cross-validation framework was used to evaluate benign–malignant classification. Patient-level partitioning was performed before model training to prevent WSIs or image patches originating from the same patient from being distributed across different data subsets. Accordingly, all WSIs from an individual patient were assigned exclusively to a single training, validation, or held-out test subset within each fold.

In each fold, approximately 80% of patients were allocated to model development and 20% were reserved as the held-out test set. The model-development subset was further divided into training and validation subsets at an approximately 8:2 ratio. The training subset was used for model training, while the validation subset was used for

hyperparameter tuning and checkpoint selection. The held-out test subset was reserved for performance evaluation. Across the five folds, the training set comprised 64 patients, the validation set comprised 16–17 patients, and the test set comprised 20–21 patients. Stratification was applied to maintain a similar benign-to-malignant ratio across subsets (Table 2). A fixed random seed (seed = 1) was used throughout the experiments. Four model configurations were evaluated to separately examine the effects of feature representation and MIL aggregation. CTransPath with CLAM-SB constituted the primary configuration. ResNet-50 with CLAM-SB was evaluated to compare feature extractors while retaining the same MIL architecture. Using CTransPath features, ABMIL and mean pooling were additionally evaluated to compare alternative slide-level aggregation strategies. The same patient-level cross-validation partitions were used across model configurations to facilitate paired comparisons.

Table 2. Case-level stratified five-fold cross-validation split details and class distribution

Fold	Training set ,n(B/M)	Validation set ,n(B/M)	Test set,n (B/M)	Proportion of total (%)
1	64(37/27)	17(10/7)	20(11/9)	19.80%
2	64(37/27)	17(10/7)	20(11/9)	19.80%
3	64(36/28)	17(10/7)	20(12/8)	19.80%
4	64(36/28)	17(10/7)	20(12/8)	19.80%
5	64(37/27)	16(9/7)	21(12/9)	20.79%

- Abbreviations: B, benign; M, malignant; n, number of cases.
- Stratified sampling was applied to preserve the benign-to-malignant ratio (approximately 1.3:1) consistent with the overall dataset across every fold, ensuring that the class distribution remained representative in all subsets.

Table 3. Experimental grouping.

Experimental group	Feature extractor	MIL model
Group 1 (Baseline)	CTransPath	CLAM-SB
Group 2 (Backbone comparator)	ResNet-50	CLAM-SB
Group 3 (MIL comparator 1)	CTransPath	ABMIL
Group 4 (MIL comparator 2)	CTransPath	Mean Pooling

Whole-Slide Image Acquisition and Preprocessing

All H&E-stained slides were digitized at 40× optical magnification using an SQS600P slide scanner (Shengqiang Technology, China) and initially stored in the proprietary SDPC format. The digitized slides were converted to pyramidal SVS files using the manufacturer's conversion tools with the lowest available compression level to preserve image quality as much as possible.

To reduce background and scanning-related artifacts, a multistage tissue-detection and quality-control pipeline was applied^[20]. At the thumbnail level, RGB images were transformed into hematoxylin–eosin–DAB (HED) optical-density space using the Ruifrok color-deconvolution method. Tissue regions were identified using grayscale

and saturation thresholds, followed by morphological opening and closing, connected-component filtering, and hole filling.

Additional rescue and exclusion rules were applied to accommodate heterogeneous staining and tissue composition. Regions with high nuclear optical density or strong eosinophilic staining were retained even when overall tissue coverage was relatively low, whereas predominantly pink, low-texture regions with minimal nuclear content were downgraded or excluded.

Tissue Segmentation and Patch Extraction

Candidate patches were extracted from the tissue mask at Level 0 using non-overlapping 256×256 -pixel windows. Patches with insufficient tissue coverage ($<25\%$) were discarded, and additional quality-control steps removed edge-adjacent, spatially fragmented, over-dark, or thin-edge patches. No manual tumor ROI was required for patch extraction. This procedure generated tens of thousands to hundreds of thousands of candidate patches per WSI, thereby retaining heterogeneous tissue information for subsequent weakly supervised MIL analysis.

Patch-Level Feature Extraction

Two feature extraction backbones were evaluated: the pathology-pretrained CTransPath model and the ImageNet-pretrained ResNet-50 model. CTransPath^[21, 22] was selected as a pathology-domain representation model, whereas ResNet-50^[23] served as a conventional natural-image-pretrained baseline.

Following quality control, each 256×256 -pixel patch was independently encoded by the corresponding pretrained backbone. The final classification layer was removed, and the resulting feature representation for each patch was retained. For each WSI, patch-level feature vectors were arranged as an $(N \times D)$ feature matrix, where (N) denotes the number of retained patches and (D) denotes the embedding dimension. The spatial coordinates of individual patches were retained to enable subsequent reconstruction of model attention maps. Both backbones were used solely as pretrained feature extractors, without joint fine-tuning with the downstream MIL aggregation models. The whole process is shown in Figure 1.

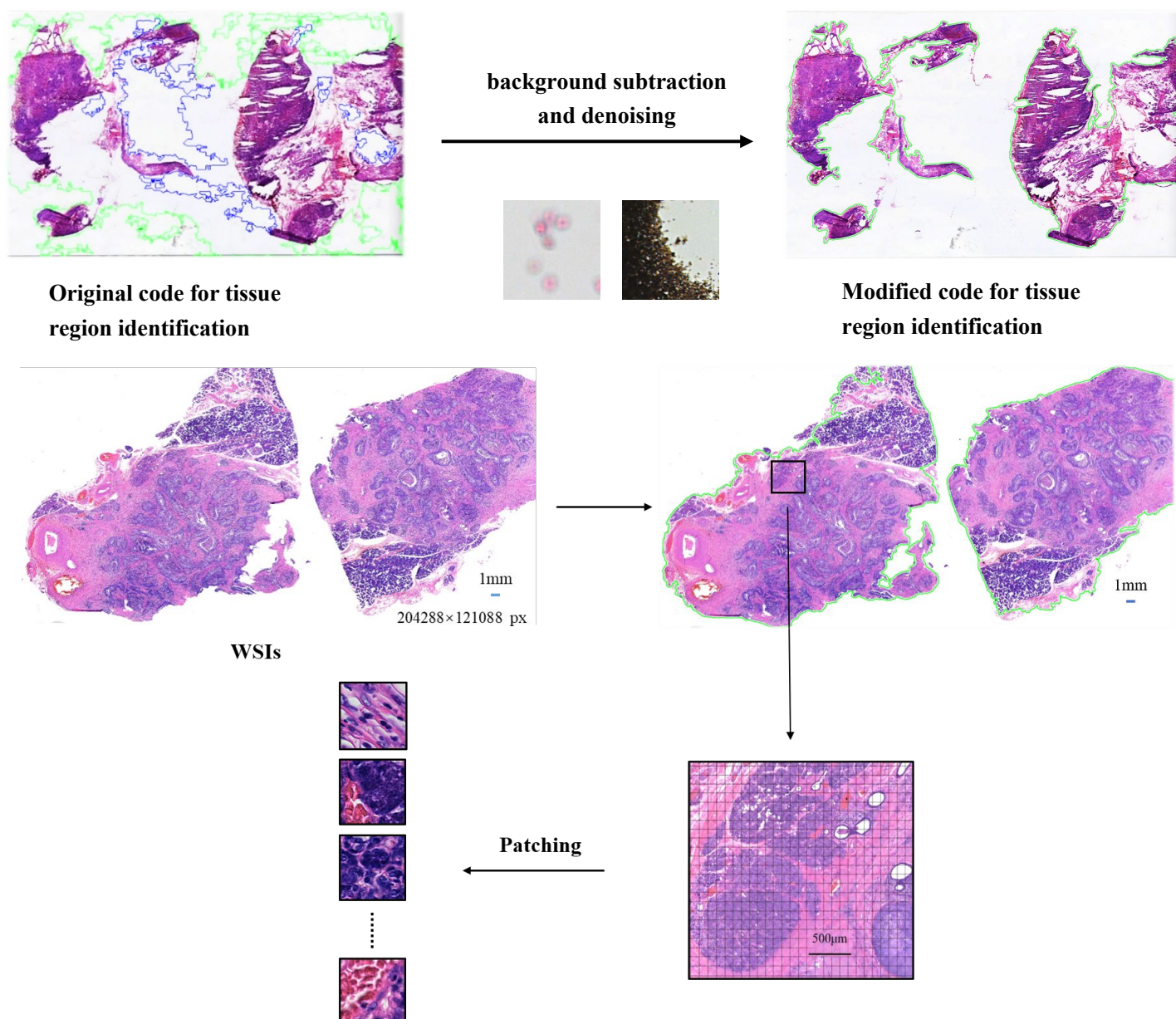


Figure 1. Whole-slide image preprocessing, segmentation, and patch extraction workflow.

MIL architectures and model training

Each WSI was represented as a bag of patch-level feature embeddings, with the corresponding histopathologic diagnosis serving as the weak supervisory label; no region- or pixel-level annotations were used. Three permutation-invariant aggregation strategies were evaluated: CLAM-SB, ABMIL, and mean pooling. CLAM-SB and ABMIL learned attention-weighted representations of patch embeddings, whereas mean pooling generated the slide-level representation by averaging the constituent patch embeddings.

Patch-level features were pre-extracted using pretrained CTransPath or ResNet-50 encoders. Both feature extractors were kept frozen during MIL training, so only the downstream MIL models were optimized for benign–malignant classification. The resulting feature embeddings and their spatial coordinates were retained for

downstream classification and attention-based spatial visualization.

For the baseline CLAM-SB configuration, the instance-sampling parameter B was set to 8, the initial learning rate to 2×10^{-4} , the dropout rate to 0.25, and the bag-level loss weight to 0.7^[24, 25]. Class-weighted cross-entropy was used for slide-level classification to account for class imbalance, and standard cross-entropy was used for the instance-level objective in CLAM-SB. Models were optimized with Adam (L2 weight decay = 1×10^{-4}), processing one WSI bag per iteration.

The same overall training and validation procedures were used across all evaluated configurations, while model-specific components were applied according to the corresponding MIL architecture.

Exploratory hyperparameter ablation

An exploratory hyperparameter ablation was performed on the CTransPath + CLAM-SB configuration to assess model sensitivity to individual training parameters. The baseline setting used B = 8, an initial learning rate of 2×10^{-4} , a dropout rate of 0.25, and a bag-level loss weight of 0.7.

Each parameter was varied individually while all other baseline settings remained fixed. Specifically, B was evaluated at 4, 8, 16, and 32; learning rates were examined around the baseline value; dropout rates of 0.25, 0.5, and 0.6 were tested; and bag-level loss weights of 0.7, 0.5, 0.3, and 0.2 were compared. A combined configuration incorporating the individually favorable settings was also explored.

Because these analyses were intended to characterize model sensitivity rather than to identify a universally optimal parameter set, the results were interpreted as exploratory. Full results for each individual setting are provided in Supplementary Table S1. In the main manuscript, the baseline configuration was compared with the alternative using a bag-level loss weight of 0.3, which yielded the most favorable mean ROC-AUC among the single-parameter ablations in this cohort.

Model evaluation

To systematically examine and compare model performance, we evaluated classification capability using accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (ROC-AUC), complemented by area under the precision-recall curve (PR-AUC) to assess discriminative power under class imbalance. Model probability calibration was additionally analyzed to verify reliability. Case-level performance was evaluated across a stratified 5-fold cross-validation scheme using both fold-averaged metrics (mean $\pm$ SD) for cross-fold stability and a pooled evaluation (aggregating predictions across all test folds) to mitigate sample-size-induced variance. Statistical comparisons between models were performed across test folds. Confusion matrix heatmaps were generated to visualize the distribution of classification errors across classes. In addition, attention heatmaps were used to localize spatial regions driving model decisions and to visualize decision confidence along tumor boundaries.

Result

Performance Comparison of Visual Feature Backbones: CTransPath vs. ResNet-50

To evaluate the impact of domain-specific feature representations, we compared the pathology-pretrained CTransPath backbone against the ImageNet-pretrained ResNet-50 counterpart within the CLAM-SB framework using a patient-level stratified 5-fold cross-validation scheme (Table 4).

Fold-averaged evaluation (mean $\pm$ SD): CTransPath achieved a higher mean AUC (0.83) compared with ResNet-50 (0.81) and demonstrated better specificity (0.80 vs. 0.75) with narrower variance. By contrast, ResNet-50 yielded slightly higher accuracy (0.82), sensitivity (0.86), and F1-score (0.82) under fold-averaged evaluation.

Pooled evaluation: When aggregating predictions across all five test folds, CTransPath outperformed ResNet-50 in pooled AUC (0.80 vs. 0.76), accuracy (0.76 vs. 0.71), specificity (0.85 vs. 0.74), and F1-score (0.73 vs. 0.71), with comparable sensitivity (0.66 vs. 0.68). Pooled evaluation provided a complementary assessment based on predictions from all held-out test folds and showed a consistent pattern of higher AUC and specificity for CTransPath. Overall, CTransPath demonstrated stronger discriminative performance and specificity, despite slightly higher sensitivity for ResNet-50, and was therefore selected as the feature backbone for subsequent MIL aggregation experiments.

Table 4. Performance Comparison of Feature Extractor Backbones (mean $\pm$ SD).

Metric	CTransPath	ResNet-50
AUC	0.83 $\pm$ 0.07	0.81 $\pm$ 0.06
Accuracy	0.81 $\pm$ 0.04	0.82 $\pm$ 0.04
Sensitivity	0.82 $\pm$ 0.14	0.86 $\pm$ 0.09
Specificity	0.80 $\pm$ 0.09	0.75 $\pm$ 0.16
F1-score	0.80 $\pm$ 0.04	0.82 $\pm$ 0.05

Performance Comparison of Weakly Supervised MIL Aggregators

Holding the CTransPath feature extractor constant, we evaluated three MIL aggregation strategies: CLAM-SB, ABMIL, and Mean Pooling, with performance metrics summarized in Table 5.

In the fold-averaged evaluation, CLAM-SB achieved the highest mean AUC (0.83), accuracy (0.81), F1-score (0.80), and specificity (0.80), with the lowest cross-fold variance. Conversely, Mean Pooling (0.89) and ABMIL (0.84) yielded higher sensitivity but substantially reduced specificity (0.65 and 0.73, respectively), indicating a diagnostic bias toward over-calling malignancy (Figure 2-1). These results reveal a clear trade-off between sensitivity and specificity across aggregation strategies: CLAM-SB achieved the lowest false-positive rate, whereas Mean Pooling attained the highest sensitivity at the cost of increased false positives. In the pooled evaluation, CLAM-SB maintained consistent superiority over the other two methods in AUC, accuracy, specificity, and F1-score (Figure 2-2a,b).

We further assessed the reliability of model output probabilities using calibration

analysis (Figure 2-2c). CLAM-SB achieved a Brier score of 0.19, with relatively good calibration in the intermediate probability range, although some overconfidence was observed at higher predicted probabilities. Mean Pooling achieved the lowest Brier score (0.188) but exhibited the largest cross-fold variance and noticeable calibration fluctuations in the intermediate range. ResNet50-CLAM yielded a higher Brier score (0.22) with pronounced overconfidence at high probabilities. ABMIL yielded a Brier score of 0.21, with the smallest cross-fold variance among the four models but marked underconfidence at low probabilities and overconfidence at higher probabilities. Overall, Brier score and discriminative performance showed divergent patterns across the four models: Mean Pooling achieved the lowest Brier score but the lowest AUC, whereas CLAM-SB achieved the highest AUC with an intermediate Brier score. To assess hyperparameter sensitivity, ablation experiments were conducted under the CTransPath + CLAM-SB framework. Relative to the standard configuration (B=8, dropout=0.25, learning rate= 2×10^{-4} , bag_loss_weight=0.7), the optimized model (bag_loss_weight=0.3) yielded consistent gains across most metrics, with AUC improving from 0.83 to 0.86, accuracy from 0.81 to 0.83, sensitivity from 0.82 to 0.87, and F1-score from 0.80 to 0.83, accompanied by a slight reduction in specificity (0.80 to 0.76) (Table 6). Cross-fold variance decreased for most metrics, indicating improved stability. Detailed results for each parameter setting are provided in Supplementary Table S1.

Table 5. Performance Comparison of MIL Aggregation Strategies (mean $\pm$ SD).

Metric	CLAM-SB	ABMIL	Mean Pooling
AUC	0.83 $\pm$ 0.07	0.80 $\pm$ 0.08	0.77 $\pm$ 0.12
Accuracy	0.81 $\pm$ 0.04	0.78 $\pm$ 0.07	0.78 $\pm$ 0.07
Sensitivity	0.82 $\pm$ 0.14	0.84 $\pm$ 0.11	0.89 $\pm$ 0.09
Specificity	0.80 $\pm$ 0.09	0.73 $\pm$ 0.09	0.65 $\pm$ 0.14
F1-score	0.80 $\pm$ 0.04	0.79 $\pm$ 0.08	0.80 $\pm$ 0.06

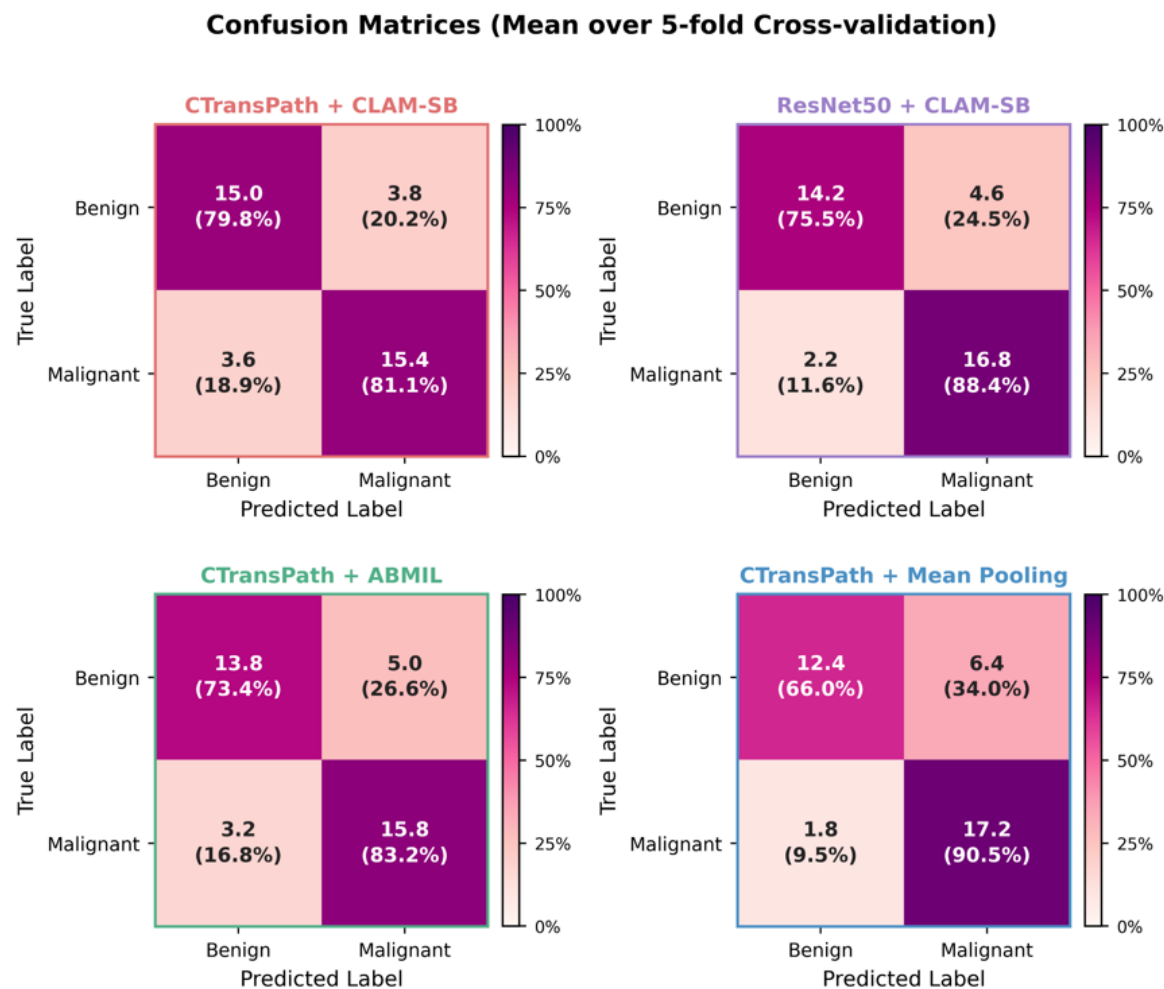


Figure 2-1. Average confusion matrix heatmaps of four MIL models under 5-fold cross-validation. Average confusion matrix heatmaps of four MIL models under 5-fold cross-validation. Each cell shows the 5-fold average count (upper) and row-normalized proportion (lower, in parentheses). Rows: true labels; columns: predicted labels. Color intensity corresponds to normalized proportion (0%–100%). Among the four groups, CTransPath + CLAM-SB achieved the most balanced performance (specificity 79.8%, sensitivity 81.1%), with the lowest false-positive rate (20.2%) and the highest false-negative rate (18.9%), indicating the most conservative strategy for benign case identification. ResNet50 + CLAM-SB achieved a sensitivity of 88.4% and a benign misclassification rate of 24.5%. CTransPath + ABMIL showed a specificity of 73.4% and a sensitivity of 83.2%. CTransPath + Mean Pooling achieved the highest sensitivity (90.5%), but also the highest benign misclassification rate (34.0%), indicating a tendency toward malignant-class predictions.

Table 6. Performance comparison between the optimized and Baseline (mean $\pm$ SD).

Metric	Baseline (bag=0.7)	Optimized model(bag=0.3)
AUC	0.83 $\pm$ 0.07	0.86 $\pm$ 0.04
Accuracy	0.81 $\pm$ 0.04	0.83 $\pm$ 0.02
Sensitivity	0.82 $\pm$ 0.14	0.87 $\pm$ 0.11
Specificity	0.80 $\pm$ 0.09	0.76 $\pm$ 0.11
F1-score	0.80 $\pm$ 0.04	0.83 $\pm$ 0.05

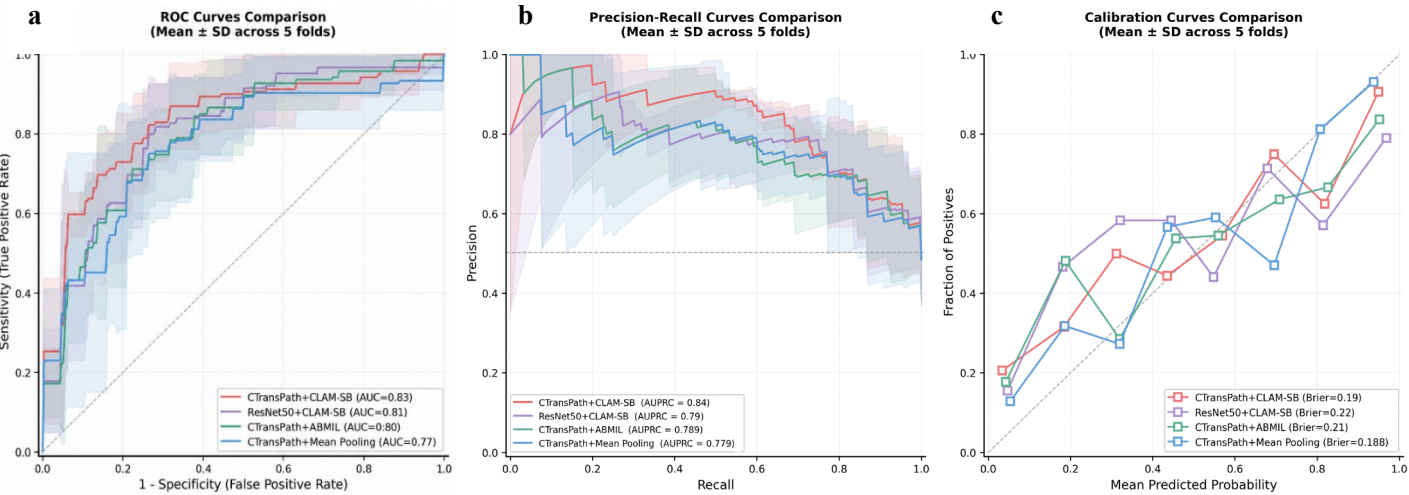


Figure 2-2. Comparative performance and calibration analysis of the evaluated models. (a) ROC curves and corresponding AUC values; (b) precision–recall curves and corresponding AUPRC values; (c) calibration curves (reliability diagrams) with corresponding Brier scores; and (d) summary of AUC, accuracy, sensitivity, specificity, and F1-score. Error bars represent the standard deviation across cross-validation folds.

A representative mucoepidermoid carcinoma case (Figure 2-3) showed substantial spatial overlap with pathologist-identified tumor regions, suggesting that the model

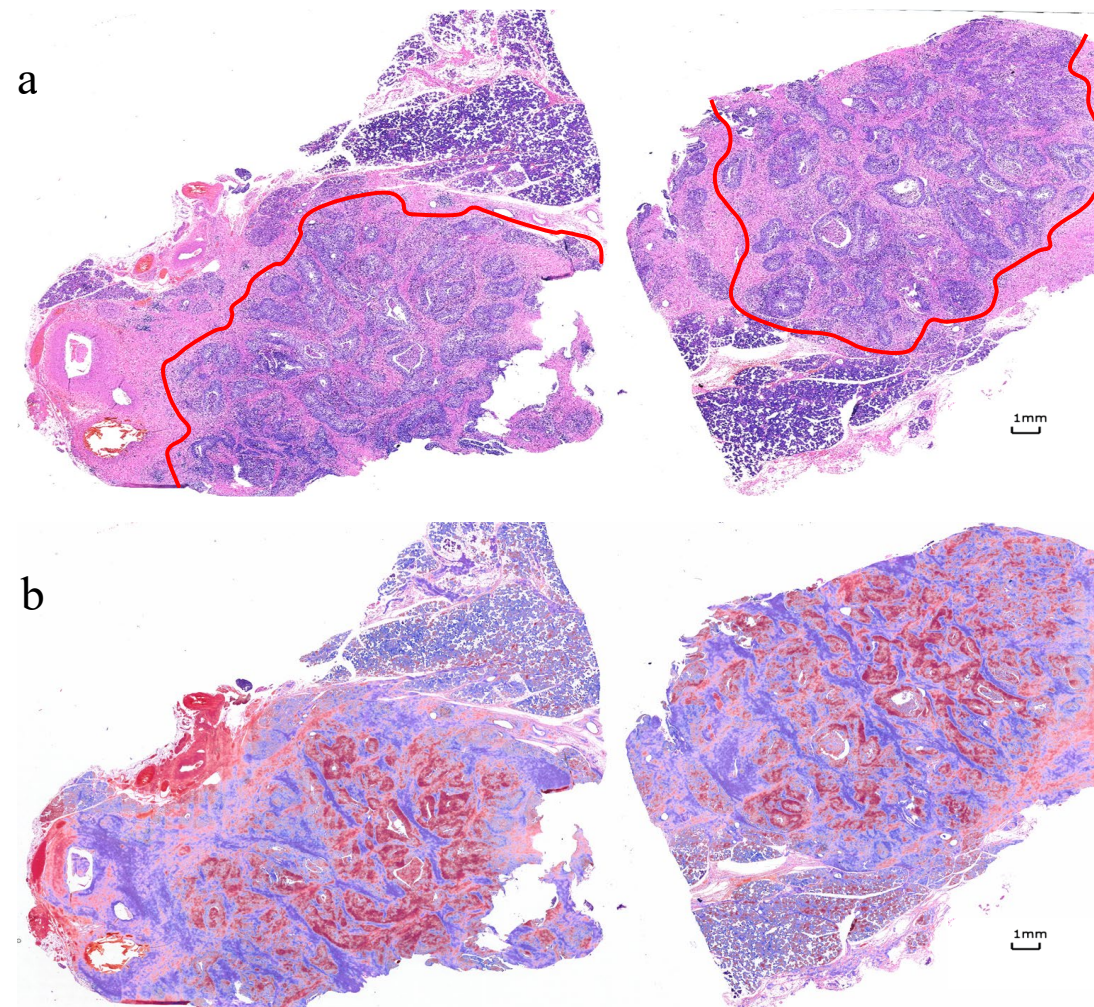


Figure 2-3. CLAM-SB attention heatmap visualization.

(a) Original H&E-stained WSI with tumor regions outlined in red by pathologists. (b) Corresponding attention heatmap: deep red denotes high-attention regions, and cool blue denotes low-attention regions. The heatmap is intended for qualitative visualization of model attention and does not represent validated tumor segmentation.

attention was concentrated in diagnostically relevant tissue areas. High-attention regions were observed in three types of histomorphologic areas: solid/glandular tumor structures, hypercellular and hyperchromatic areas, and tumor-stromal interfaces, suggesting that the model may have captured information related to cellular composition, nuclear atypia, and invasive growth. Notably, hemorrhagic areas also received high attention in this case, possibly reflecting either spurious attention due to histomorphologic similarity or the potential use of microenvironmental information as an auxiliary cue. This observation is consistent with previous reports^[11], suggesting that the model may utilize not only tumor-cell morphology but also microenvironmental features as auxiliary discriminative cues^[26].

Discussion

In this study, we addressed the currently limited application of artificial intelligence in the histopathological diagnosis of PGTs by systematically evaluating weakly supervised MIL frameworks on WSIs^[27-29]. Unlike prior studies that relied on manually annotated ROIs or preoperative radiomics^[30-32], our approach required only slide-level diagnostic labels, demonstrating that clinically meaningful benign-malignant differentiation is achievable even under small-sample constraints. Specifically, the combination of a pathology-pretrained CTransPath backbone and the CLAM-SB aggregation architecture yielded the best overall discriminative performance, with CLAM-SB effectively balancing sensitivity and specificity while minimizing false positives. Although statistical comparisons did not reach significance ($p > 0.05$), likely due to limited sample size, the consistent superiority of CTransPath and CLAM-SB across multiple evaluation metrics supports the robustness of our primary conclusions. The choice of feature extractor serves as the cornerstone for digital pathology models^[13, 33], directly dictating the quality of instance representations fed into downstream MIL aggregators. In our comparative analysis, CTransPath demonstrated superior overall discrimination, F1-score, and cross-fold stability compared to the ImageNet-pretrained ResNet-50 counterpart. Rather than fine-tuning backbones end-to-end—a regime prone to severe overfitting and prohibitive computational overhead in small datasets—we leveraged self-supervised representations pretrained on large-scale pathology slide repositories to obtain a more robust and reproducible feature space. The advantage of Transformer-based architectures over CNN counterparts in capturing global contextual information has been systematically validated in pathological image analysis^[22]. Specifically, the self-attention mechanism enables direct modeling of relationships between spatially distant tissue regions, which is particularly advantageous for tumors with complex microenvironments where local morphological features alone are insufficient for accurate classification.

Alsanie et al.^[11] constructed machine learning classifiers using 320 H&E-stained WSIs, achieving F1-scores of 0.95 for both benign versus malignant differentiation and malignant subtype classification, while maintaining an F1-score of 0.87 on an independent external validation cohort. Their study used a larger sample size (320 WSIs) with conventional machine learning classifiers, whereas our weakly supervised deep learning MIL framework achieved clinically acceptable diagnostic performance under small-sample conditions, suggesting that MIL approaches may offer viable diagnostic potential on small-scale pathology datasets where sample acquisition is challenging. Notably, Alsanie also found in their small-sample experiment (36 cases, 91 WSIs) that detailed annotations significantly improved classification performance, which is consistent with our observation that weakly supervised MIL can still achieve clinically acceptable results under small-sample constraints, collectively suggesting that both high-quality annotation strategies and appropriate MIL framework design are important in data-limited scenarios.

The superiority of CTransPath has been independently validated across multiple diseases. Bourgade et al.^[34] combined CTransPath with attention-based MIL for giant cell arteritis diagnosis, achieving a mean AUROC of 0.994 on an external test cohort, with attention successfully localizing medial inflammation as the key predictive feature.

Although that task differs substantially from PGT discrimination in lesion nature, its excellent performance on external multi-center data suggests that CTransPath possesses robust cross-disease, cross-distribution feature extraction capability, providing methodological confidence for its application in our small-sample PGT cohort. Hezil et al.^[22] employed CTransPath as the histological encoder in a multimodal framework for breast cancer subtype classification, achieving 94.93% accuracy on TCGA-BRCA (924 patients), significantly outperforming unimodal baselines. Although their study involved a multimodal fusion scenario, the fact that CTransPath alone as the histological encoder provided sufficiently discriminative features aligns with our finding that it performs well even as a single-modality feature extractor under small-sample conditions, further supporting its practical value as a general-purpose pathological feature backbone.

In the salivary gland tumor domain, Nakamura et al.^[35] compared eight CNN architectures on 107 WSIs (83 patients) and found ResNet-50 to be the top performer across 7 out of 9 metrics, with DenseNet121 showing comparable specificity (94% vs 93%) and AUC (0.97). Although that study employed a supervised CNN framework for distinguishing pleomorphic adenoma from carcinoma ex pleomorphic adenoma, the strong performance of ResNet-50 in that dataset is consistent with its role as a robust baseline in our study, while our finding that CTransPath further surpassed ResNet-50 suggests that domain-specific pretrained Transformer architectures may provide additional gains. Together, these two studies indicate that even in closely related salivary gland pathology tasks, Transformer architectures may offer advantages over conventional CNNs. More recently, Li et al.^[36] applied a Vision Transformer (ViT) to classify 3,046 WSIs across nine salivary gland tumor subtypes, achieving 0.9966 accuracy and 0.9848 F1-score, surpassing benchmark models and demonstrating diagnostic performance comparable to a chief pathologist on a 100-WSI subset with substantially reduced diagnosis time. Their findings confirm the strong potential of Transformer architectures in salivary gland tumor pathology, aligning with our use of CTransPath. Notably, while Li et al. addressed a large-sample multi-class task, our observation of similar Transformer superiority (CTransPath over ResNet-50) in a small-sample binary task suggests that Transformer architectures may offer advantages across different sample sizes and classification granularities, supporting their good scalability and adaptability in salivary gland/head and neck pathology tasks.

Notably, the two backbones exhibited distinct diagnostic tendencies: ResNet-50 achieved higher sensitivity (0.86) while CTransPath prioritized specificity (0.80). ResNet-50's local receptive fields excel at capturing granular cytological features, minimizing false negatives, but often amplify non-tumoral reactive changes, causing false positives. Conversely, CTransPath's self-attention mechanisms model long-range spatial dependencies, filtering out non-specific stromal noise by contextualizing local variations within the broader tissue architecture^[37]. Thus, high-sensitivity backbones suit frontline screening to avoid missed diagnoses, while high-specificity models are better for confirmatory consultations to prevent overtreatment. Both tendencies were clearly observed in our small-sample cohort, suggesting that backbone selection should be guided by the specific clinical context.

With the CTransPath feature extractor held constant, CLAM-SB achieved the highest mean AUC, accuracy, F1-score, and specificity, whereas mean pooling attained the highest sensitivity at the cost of substantially reduced specificity. This pattern suggests that the mean pooling strategy, which treats all patches equally, tends to introduce stromal noise in highly heterogeneous tissues such as the parotid gland, whereas CLAM-SB's clustering-constrained attention mechanism can more effectively focus on diagnostically relevant regions^[22, 38]. Under small-sample conditions, this mechanistic advantage of CLAM-SB is particularly pronounced, as its clustering constraint actively suppresses background noise and thereby reduces overfitting to non-specific features with limited training data.

Lysukhin et al.^[39] further explored the impact of sample size on model performance. Using patient-level weak labels from 1,102 patients (5,104 WSIs), their MIL model achieved a mean test F1-score of 0.85; their small-sample experiment (36 cases, 91 WSIs) additionally demonstrated that detailed annotations significantly improved classification performance. Notably, compared with that large-sample study, our study achieved comparable diagnostic performance (AUC 0.83) with only 101 patients, suggesting that in patient-level weakly labeled MIL frameworks, a few hundred cases may approach the saturation point of performance gain. Butt et al.^[40], in a large-scale weakly supervised MIL study on prostate cancer using the PANDA dataset (10,616 WSIs), systematically compared six MIL architectures and three feature encoders, and similarly found that CTransPath significantly outperformed ResNet-50, consistent with our findings. Together, these large-sample studies and our small-sample validation suggest that the CTransPath-based weakly supervised MIL framework performs robustly across multiple tumor types, and that its effectiveness under small-sample conditions is unlikely to be coincidental. Moreover, CLAM-SB demonstrated stable performance across multiple experimental configurations, validating its effectiveness. However, under small-sample conditions, its clustering constraint mechanism is sensitive to hyperparameters, and training time is substantially longer than that of mean pooling and ABMIL; when class imbalance is severe or slide quality is highly inconsistent, the clustering branches may overfit to the majority class.

Our systematic ablation experiments further revealed the hyperparameter sensitivity of CLAM-SB on small-scale datasets. Among the multiple factors affecting model performance, adjustment of the bag-level loss weight (`bag_loss_weight`) produced the most pronounced effect: reducing it from 0.7 to 0.3 led to stable performance improvement with consistently low cross-fold variance. This finding suggests that moderately relaxing strict adherence to slide-level labels allows the attention mechanism to more effectively focus on locally discriminative regions in weakly supervised MIL frameworks. However, combining all single-factor optimal settings into a multi-parameter joint model did not yield further improvement, instead showing signs of cross-fold distribution shift. Vabalas et al.^[41] demonstrated that when training data are limited, K-fold cross-validation itself may produce optimistically biased estimates of complex hyperparameter tuning, making the true gains of multi-parameter joint optimization difficult to reliably detect. This observation suggests that in sample-limited exploratory studies, a streamlined single-factor optimization strategy may be

more reliable than complex multi-parameter joint tuning. Our ablation results in the 101-patient small-sample cohort are highly consistent with these findings, further validating the practical value of single-factor optimization under small-scale data conditions.

Early pre-experiments in this study further revealed a noteworthy issue: unaugmented CTransPath embeddings produced near-perfect classification scores. Subsequent label-permutation and instance-reduction tests ruled out data leakage, indicating that the Vision Transformer had overfit to non-pathological, dataset-specific confounding features, including staining batch, scanner color temperature, and background hue shifts, rather than learning true histomorphological representations. This finding highlights an important vulnerability of deep learning models in pathological image analysis: when training samples are limited, models tend to exploit low-level color shortcuts related to staining and scanning, rather than learning genuine histomorphological features. Tellez et al.^[42], using a heterogeneous H&E dataset spanning four organs and nine pathology laboratories, systematically demonstrated that stain color variation is a critical factor affecting cross-center generalization of CNNs, and identified color augmentation and stain normalization as important strategies to mitigate this issue. Franchet et al.^[43] further quantified the impact of staining-related bias, showing that AI model decisions are often based on dataset-specific bias rather than biological features. To address this bias, we incorporated aggressive online color jittering during feature extraction, and successfully recovered unbiased benchmarks through resampling and retraining on color-perturbed features. This experience suggests that color/stain normalization or augmentation strategies should be considered a standard preprocessing step in digital pathology model development, particularly when training data originate from different scanners or staining batches. In our 101-patient small-sample PGT cohort, this strategy proved to be a critical step in restoring model generalizability, highlighting the particular importance of vigilance against non-biological shortcuts in small-sample settings.

Building on this, dataset expansion experiments further demonstrated the value of multi-WSI modeling. Expanding the dataset from a single slide per case to additional slides from the same patients led to stable performance improvement and a marked reduction in cross-fold standard deviation. Parotid neoplasms harbor substantial intratumoral spatial heterogeneity; multi-slide sampling captures diverse cellular architectural patterns across different tumor regions, effectively mitigating sampling bias and enhancing model stability. Under the small-sample condition of only 101 patients, increasing the number of slides per case (from 101 to 189 WSIs) yielded significant performance improvement, suggesting that fully leveraging available slide resources from existing cases is an effective strategy for enhancing model performance in small-sample settings.

Beyond discriminative performance, we further assessed the reliability of model predictions through calibration analysis. Among the four configurations, CLAM-SB achieved a Brier score of 0.19, with the most stable calibration in the mid-probability range, supporting its clinical utility as a risk estimation tool. Notably, reducing the bag-level loss weight from 0.7 to 0.3 further improved the mean AUC from 0.83 to 0.86

and sensitivity from 0.82 to 0.87, while slightly decreasing specificity from 0.80 to 0.76. This observation aligns with the principle that moderately relaxing slide-level supervision allows the attention mechanism to better capture locally discriminative patterns, which appears particularly beneficial under small-sample conditions. The optimized model also exhibited reduced cross-fold variance across most metrics, indicating enhanced stability. Collectively, these findings suggest that the CTransPath-CLAM combination, with appropriate hyperparameter tuning, offers a practical balance between sensitivity, specificity, and calibration reliability for PGT diagnosis in resource-limited settings.

Conclusion

In this study, we explored the feasibility of weakly supervised MIL for distinguishing benign from malignant parotid gland tumors on WSIs. The CTransPath + CLAM-SB combination delivered clinically acceptable performance in a 101-patient cohort, with consistent advantages across multiple metrics. External evidence further supported CTransPath as a robust cross-tissue feature extractor, and highlighted the value of CLAM-SB's clustering-constrained attention in histologically heterogeneous settings. Calibration analysis confirmed the reliability of CLAM-SB predictions, and tuning the bag-level loss weight to 0.3 further raised the mean AUC to 0.86 with improved stability. Ablation experiments underscored the challenges of hyperparameter tuning in small samples, while color augmentation helped reduce the model's dependence on staining- and scanner-related shortcuts. Overall, our findings suggest that weakly supervised MIL holds promise for PGT pathology, and warrant further validation in larger, multi-center cohorts.

References

- [1] Lee DH, Jung EK, Lee JK, Lim SC: Comparative analysis of benign and malignant parotid gland tumors: A retrospective study of 992 patients. *Am J Otolaryngol* 2023, 44(2):103690. DOI:10.1016/j.amjoto.2022.103690
- [2] Faur AC, Buzaş R, Lăzărescu AE, Ghenciu LA: Current Developments in Diagnosis of Salivary Gland Tumors: From Structure to Artificial Intelligence. *Life (Basel)* 2024, 14(6). DOI:10.3390/life14060727
- [3] Johnson F, Bozzato A, Mansour N, Mantsopoulos K, Psychogios G, Zengel P, Hofauer B: Sonography of Salivary Gland Tumors and Disorders. *Ultraschall Med* 2025, 46(4):318-344. DOI:10.1055/a-2481-7248
- [4] Huang AT, Tang C, Bell D, Yener M, Izquierdo L, Frank SJ, El-Naggar AK, Hanna EY, Weber RS, Kupferman ME: Prognostic factors in adenocarcinoma of the salivary glands. *Oral Oncol* 2015, 51(6):610-615. DOI:10.1016/j.oraloncology.2015.03.005
- [5] Muller S, Tilakaratne WM: Update from the 5th Edition of the World Health Organization Classification of Head and Neck Tumors: Tumours of the Oral Cavity and Mobile Tongue. *Head Neck Pathol* 2022, 16(1):54-62. DOI:10.1007/s12105-021-01402-9
- [6] Locati LD, Ferrarotto R, Licitra L, Benazzo M, Preda L, Farina D, Gatta G,

- Lombardi D, Nicolai P, Vander Poorten V *et al*: Current management and future challenges in salivary glands cancer. *Front Oncol* 2023, 13:1264287. DOI:10.3389/fonc.2023.1264287
- [7] Zoccali F, Cialente F, Colizza A, Ralli M, Greco A, de Vincentiis M: Clinico-histopathological review of 255 patients who underwent parotidectomy for pleomorphic adenoma: a 10-year retrospective study-a proposal for an optimal diagnostic and therapeutic algorithm for patients with recurrent pleomorphic adenoma. *Eur Arch Otorhinolaryngol* 2023, 280(7):3329-3335. DOI:10.1007/s00405-023-07897-y
- [8] Cheng PC, Chiang HK: Diagnosis of Salivary Gland Tumors Using Transfer Learning with Fine-Tuning and Gradual Unfreezing. *Diagnostics (Basel)* 2023, 13(21). DOI:10.3390/diagnostics13213333
- [9] Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A: Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019, 16(11):703-715. DOI:10.1038/s41571-019-0252-y
- [10] Rossi ED, Faquin WC, Baloch Z, Barkan GA, Foschini MP, Pusztaszeri M, Vielh P, Kurtycz DFI: The Milan System for Reporting Salivary Gland Cytopathology: Analysis and suggestions of initial survey. *Cancer Cytopathol* 2017, 125(10):757-766. DOI:10.1002/cncy.21898
- [11] Alsanie I, Shephard A, Azarmehr N, Vargas P, Pring M, Rajpoot NM, Khurram SA: Exploring the feasibility of AI-based analysis of histopathological variability in salivary gland tumours. *Sci Rep* 2025, 15(1):29171. DOI:10.1038/s41598-025-15249-5
- [12] Zbären P, Nuyens M, Loosli H, Stauffer E: Diagnostic accuracy of fine-needle aspiration cytology and frozen section in primary parotid carcinoma. *Cancer* 2004, 100(9):1876-1883. DOI:10.1002/cncr.20186
- [13] Madabhushi A, Lee G: Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med Image Anal* 2016, 33:170-175. DOI:10.1016/j.media.2016.06.037
- [14] Wilson CM, Ospina OE, Townsend MK, Nguyen J, Moran Segura C, Schildkraut JM, Tworoger SS, Peres LC, Fridley BL: Challenges and Opportunities in the Statistical Analysis of Multiplex Immunofluorescence Data. *Cancers (Basel)* 2021, 13(12). DOI:10.3390/cancers13123031
- [15] Classe M, Lerousseau M, Scoazec JY, Deutsch E: Perspectives in pathomics in head and neck cancer. *Curr Opin Oncol* 2021, 33(3):175-183. DOI:10.1097/cco.0000000000000731
- [16] Wang S, Yang DM, Rong R, Zhan X, Fujimoto J, Liu H, Minna J, Wistuba, II, Xie Y, Xiao G: Artificial Intelligence in Lung Cancer Pathology Image Analysis. *Cancers (Basel)* 2019, 11(11). DOI:10.3390/cancers11111673
- [17] Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, West RB, van de Rijn M, Koller D: Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* 2011, 3(108):108ra113. DOI:10.1126/scitranslmed.3002564

- [18] Wetstein SC, de Jong VMT, Stathonikos N, Opdam M, Dackus G, Pluim JPW, van Diest PJ, Veta M: Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Sci Rep* 2022, 12(1):15102. DOI:10.1038/s41598-022-19112-9
- [19] Försch S, Klauschen F, Hufnagl P, Roth W: Artificial Intelligence in Pathology. *Dtsch Arztebl Int* 2021, 118(12):194-204. DOI:10.3238/arztebl.m2021.0011
- [20] Roy M, Wang F, Teodoro G, Bhattarai S, Bhargava M, Rekha TS, Aneja R, Kong J: Deep learning based registration of serial whole-slide histopathology images in different stains. *J Pathol Inform* 2023, 14:100311. DOI:10.1016/j.jpi.2023.100311
- [21] Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, Huang J, Han X: Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal* 2022, 81:102559. DOI:10.1016/j.media.2022.102559
- [22] Xu H, Wang M, Shi D, Qin H, Zhang Y, Liu Z, Madabhushi A, Gao P, Cong F, Lu C: When multiple instance learning meets foundation models: Advancing histological whole slide image analysis. *Med Image Anal* 2025, 101:103456. DOI:10.1016/j.media.2025.103456
- [23] Viswanathan K, Maleki Z, Pantanowitz L, Cantley R, Faquin WC: Cytomorphology and diagnostic pitfalls of sebaceous and nonsebaceous salivary gland lymphadenoma: A multi-institutional study. *Diagn Cytopathol* 2021, 49(1):83-95. DOI:10.1002/dc.24602
- [24] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021, 5(6):555-570. DOI:10.1038/s41551-020-00682-w
- [25] Ilse M, Tomczak JM, Welling M: Attention-based deep multiple instance learning. Proceedings of the 35th International Conference on Machine Learning (ICML) 2018, 2127-2136. DOI:10.48550/arXiv.1802.04712
- [26] Weber A, Schuster D, Heyer J, Becker C, Burkhardt V, Werner M, Spörlein A, Bronsert P, Schulz T: Deep learning-based histopathological classification and subclassification of benign and malignant salivary gland tumors. *Eur Arch Otorhinolaryngol* 2026, 283(6):3905-3915. DOI:10.1007/s00405-026-10082-6
- [27] Li K, Qian Z, Han Y, Chang EI, Wei B, Lai M, Liao J, Fan Y, Xu Y: Weakly supervised histopathology image segmentation with self-attention. *Med Image Anal* 2023, 86:102791. DOI:10.1016/j.media.2023.102791
- [28] Lu Y, Song W, Arachie C, Huang B: Weakly supervised label learning flows. *Neural Netw* 2025, 182:106892. DOI:10.1016/j.neunet.2024.106892
- [29] Neidlinger P, El Nahhas OSM, Muti HS, Lenz T, Hoffmeister M, Brenner H, van Treeck M, Langer R, Dislich B, Behrens HM *et al*: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nat Biomed Eng* 2026, 10(6):1113-1123. DOI:10.1038/s41551-025-01516-3
- [30] Wang Y, Gao J, Yin Z, Wen Y, Sun M, Han R: Differentiation of benign and

- malignant parotid gland tumors based on the fusion of radiomics and deep learning features on ultrasound images. *Front Oncol* 2024, 14:1384105. DOI:10.3389/fonc.2024.1384105
- [31] Wang Y, Dai A, Wen Y, Sun M, Gao J, Yin Z, Han R: Prediction of High-risk Capsule Characteristics for Recurrence of Pleomorphic Adenoma in the Parotid Gland Based on Habitat Imaging and Peritumoral Radiomics: A Two-center Study. *Acad Radiol* 2025, 32(7):4134-4145. DOI:10.1016/j.acra.2025.01.030
- [32] Guo J, Feng J, Huang Y, Li X, Hu Z, Zhou Q, Xu H: Diagnostic performance of MRI-based radiomics models using machine learning approaches for the triple classification of parotid tumors. *Heliyon* 2024, 10(17):e36601. DOI:10.1016/j.heliyon.2024.e36601
- [33] Li MY, Pan Y, Lv Y, Ma H, Sun PL, Gao HW: Digital pathology and artificial intelligence in renal cell carcinoma focusing on feature extraction: a literature review. *Front Oncol* 2025, 15:1516264. DOI:10.3389/fonc.2025.1516264
- [34] Bourgade R, Elhannani M, Loussouarn D, Guédon A, Péteri R, Allix-Béguet C, Berthier M, Toquet C, Espitia O, Roncato C *et al*: Deep learning for giant cell arteritis diagnosis on temporal artery biopsy. *Comput Biol Med* 2025, 196(Pt A):110707. DOI:10.1016/j.compbimed.2025.110707
- [35] Hezil N, Bouridane A, Hamoudi R, Al-Maadeed S, Akbari Y, Abdullakutty F: Selective multimodal deep learning for reliable breast cancer subtype classification from histopathology and genomic data. *Med Eng Phys* 2026, 147(3). DOI:10.1088/1873-4030/ae449b
- [36] Li M, Shen ZL, Xian HC, Zheng ZJ, Yu ZW, Liang XH, Gao R, Tang YL, Zhang Z: A Recognition System for Diagnosing Salivary Gland Neoplasms Based on Vision Transformer. *Am J Pathol* 2025, 195(2):221-231. DOI:10.1016/j.ajpath.2024.09.010
- [37] Zhao L, Zhang Z: A improved pooling method for convolutional neural networks. *Sci Rep* 2024, 14(1):1589. DOI:10.1038/s41598-024-51258-6
- [38] Zhou Y, Jin F, Suo G, Yang J: ResViT-GANNet: a deep learning framework for classifying breast cancer histopathology images using multimodal attention and GAN-based augmentation. *BMC Med Imaging* 2025, 25(1):401. DOI:10.1186/s12880-025-01940-6
- [39] Lysukhin D, Varlamov A, Yakimov B, Porubayeva E, Pachuashvili N, Kovaleva E, Vanushko V, Platonova N, Shirshin E, Mokrysheva N *et al*: Multiple-Instance Learning for thyroid gland disease classification: A hands-on experience. *Comput Biol Med* 2025, 184:109424. DOI:10.1016/j.compbimed.2024.109424
- [40] Butt NA, Sarwat D, Noya ID, Tutusaus K, Samee NA, Ashraf I: Benchmarking multiple instance learning architectures from patches to pathology for prostate cancer detection and grading using attention-based weak supervision. *Sci Rep* 2026, 16(1). DOI:10.1038/s41598-026-39196-x
- [41] Vabalas A, Gowen E, Poliakoff E, Casson AJ: Machine learning algorithm validation with a limited sample size. *PLoS One* 2019, 14(11):e0224365. DOI:10.1371/journal.pone.0224365

- [42] Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, van der Laak J: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal* 2019, 58:101544. DOI:10.1016/j.media.2019.101544
- [43] Franchet C, Schwob R, Bataillon G, Syrykh C, Péricart S, Frenois FX, Penault-Llorca F, Lacroix-Triki M, Arnould L, Lemonnier J *et al*: Bias reduction using combined stain normalization and augmentation for AI-based classification of histological images. *Comput Biol Med* 2024, 171:108130. DOI:10.1016/j.combiomed.2024.108130