Journal of Pathology
J Pathol 2026
Published online 5 August 2026 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.70105

ORIGINAL ARTICLE

EBV status prediction in gastric carcinoma biopsy images using multiple bag instance learning and foundation model in multicenter data

Haeyoun Kang[1†], Yujin Nam[2†], Dat Ngo[3], Min-Sun Cho[4], Wonae Lee[5], Hyeong-Chan Shin[6], Yujun Park[1], Hee Jung An[1], Hee-Cheol Kim[7], Nam Hoon Cho[8,9], Hyunki Kim[8*] and Baek Hwan Cho[2,10*]

[1] *Department of Pathology, CHA Bundang Medical Center, CHA University School of Medicine, Seongnam, Republic of Korea*
[2] *Medical AI Research Center, CHA Bundang Medical Center, CHA University School of Medicine, Seongnam, Republic of Korea*
[3] *Department of Computer Engineering, Korea National University of Transportation, Chungju, Republic of Korea*
[4] *Department of Pathology, College of Medicine, Ewha Womans University, Seoul, Republic of Korea*
[5] *Department of Pathology, Dankook University Hospital, Cheonan, Republic of Korea*
[6] *Department of Pathology, Keimyung University Dongsan Medical Center, Daegu, Republic of Korea*
[7] *Department of Computer Engineering/Institute of Digital Anti-Aging Healthcare/u-HARC, Inje University, Gimhae, Republic of Korea*
[8] *Department of Pathology, Yonsei University College of Medicine, Seoul, Republic of Korea*
[9] *MTS Company, Seoul, Republic of Korea*
[10] *Department of Biomedical Informatics, CHA University School of Medicine, Seongnam, Republic of Korea*

**Correspondence to: H Kim, Department of Pathology, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. E-mail: kimhyunki@yuhs.ac; BH Cho, 335 Pangyo-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, 13488, Republic of Korea. E-mail: baekhwan.cho@cha.ac.kr*

[†]These authors contributed equally to this work.

Abstract

Epstein–Barr virus associated gastric carcinoma (EBVaGC) represents a distinct molecular subtype of gastric cancer with prognostic and therapeutic significance. Current diagnostic practice relies on Epstein–Barr virus-encoded small RNA (EBER) *in situ* hybridization, which is rarely performed on initial biopsy specimens. We developed a deep learning-based diagnostic pipeline that predicts EBV status directly from H&E-stained gastric biopsy slides. Our method incorporates a multiple bag instance learning (MBIL) framework, which groups spatially adjacent tumor patches to address the inherent heterogeneity of biopsy samples and improve predictive performance. A pathology foundation model was employed to distinguish tumor from normal regions with minimal pathologist annotation, enabling robust tumor detection and contributing to downstream model accuracy. The model was trained on 744 biopsy cases from three general hospitals using 5-fold cross-validation and evaluated on an independent external cohort of 157 cases. MBIL-based models consistently outperformed single-bag approaches. The DSMIL-MB model achieved the best performance, with an area under the receiver operating characteristic curve (AUC) of 0.907 ± 0.044 in internal validation and 0.870 in external testing. TransMIL also showed improved performance with MBIL, with external AUC increasing from 0.701 to 0.853. This framework may serve as a practical prescreening tool for EBV status prediction in routine gastric biopsy specimens prior to EBER-ISH testing, supporting preoperative decision-making and facilitating broader implementation of EBV screening in gastric cancer diagnostics.

Keywords: digital pathology; gastric cancer; Epstein–Barr virus; deep learning; multi-instance learning; foundation model

Received 30 September 2025; Revised 9 April 2026; Accepted 15 July 2026

No conflicts of interest were declared.

Introduction

Epstein–Barr virus associated gastric carcinoma (EBVaGC) accounts for approximately 9% of all gastric cancer cases and is recognized as a distinct subtype with prominent lymphoid infiltration, a proximal tumor location, lower frequency of lymph node metastasis (LNM), and a relatively favorable prognosis compared with EBV-negative tumors [1–3]. Identifying EBV status carries important clinical implications, informing both prognostic evaluation and therapeutic decision-making, particularly regarding the appropriateness of endoscopic resection in early gastric cancer (EGC). Uemura *et al* described a distinctive lace-like histologic pattern of intramucosal EBV-associated EGC (EBVa EGC) [4]. The lace pattern lacks tubule formation and is often classified as poorly differentiated. These histologic features could be diagnostically subtle and challenging to

assess, but may be more objectively and reliably identified with the aid of artificial intelligence (AI). Because only a small subset of poorly differentiated or undifferentiated EGCs meet the criteria for curative endoscopic resection, patients with EBVa EGC may miss the opportunity for minimally invasive treatment, despite the very low risk of LNM. The current diagnostic gold standard, Epstein-Barr virus-encoded small RNA (EBER) *in situ* hybridization (EBER-ISH), which detects EBV-encoded small RNAs in tumor cells, offers high specificity but is constrained by technical complexity, cost, and limited accessibility in routine settings [5]. These limitations highlight the pressing need for alternative diagnostic strategies that are both accurate and more broadly deployable in routine daily practice.

Recent advances in AI have enabled deep learning models and foundation models to extract biologically meaningful features from histopathological images, allowing prediction of molecular subtypes and biomarkers directly from H&E slides [6–9]. While several studies have investigated AI-based EBV status prediction in gastric cancer, most have focused on resected specimens collected after surgery [10–12]. Although informative, these datasets have limited utility for pretreatment diagnosis and early clinical decision-making, since they are based on resection specimens [13]. In contrast, biopsy specimens routinely collected during early diagnostic workups, are minimally invasive, and better reflect real-world clinical settings [14,15]. Importantly, biopsy-based modeling may facilitate earlier stratification of patients and support treatment decisions prior to surgical intervention—especially in the context of EBVaGC, which exhibits favorable prognosis and distinct immune features that could justify less aggressive staging or nodal dissection [3]. However, few studies have systematically explored AI-based EBV prediction using gastric biopsy tissue, despite its potential utility [13,16].

Nonetheless, biopsy samples inherently contain smaller tissue volumes and are more susceptible to sampling variability, which can compromise the accuracy of EBV detection by both conventional techniques and AI-based approaches [17]. To address this, digital pathology models often adopt a multiple instance learning (MIL) approach, where image patches from a slide are grouped into a single bag for training [18–21]. Although effective for large resection specimens, this single-bag strategy may dilute localized morphological cues in small or histologically heterogeneous biopsies, thereby reducing sensitivity in detecting focal EBV-associated features [22].

To address this limitation, we propose multiple bag instance learning (MBIL), a deep learning framework that preserves spatial histologic context through clustered multiple bag generation and improves EBV prediction in gastric biopsy slides. We developed and validated our framework using H&E-stained gastric biopsy samples curated from multiple institutions, capturing variability in patient demographics and tumor morphology. Our objectives were to evaluate the diagnostic accuracy and multicenter generalizability of this MBIL-based framework for EBV status prediction from routine H&E-stained gastric biopsy slides, and to assess its potential as a clinically deployable triage tool prior to confirmatory EBER-ISH testing. We trained the model on biopsy cohorts from three institutions and externally validated it on an independent cohort from a fourth institution, and provide explainable attention heatmaps to assist pathologists in understanding model decisions.

Materials and methods

Ethics statement and patient consent

H&E-stained whole-slide images (WSIs) were retrospectively collected from four medical institutions: Severance Hospital (SH) (Seoul, Republic of Korea); CHA Bundang Medical Center (CBMC) (Seongnam, Republic of Korea); Dankook University Hospital (DUH) (Cheonan, Republic of Korea) and Ewha Womans University Mokdong Hospital (EWUM) (Seoul, Republic of Korea). All data collection protocols were approved by the Institutional Review Boards (IRBs) of the respective institutions (IRB numbers: SEUMC 2021-07-016, EUMC 2021-08-013, DKUH 2022-03-041, KUDH 2021-08-112, CHABH 2021-09-021-011, YUMC 4-2021-0029), and the need for informed consent was waived due to the retrospective nature of the study.

Data acquisition, processing

The consecutive gastric biopsy specimens obtained between January 2000 and March 2024 were reviewed and only the cases with EBV-ISH test results available were included. One representative slide from each case was included in case of multiple-site biopsies. WSIs were scanned at 20× magnification using institution-specific scanners (Leica AT2, Leica Biosystems, Germany, Panoramic 1000DX, 3D Histech, Hungary) and anonymized. Images underwent H&E stain normalization during patch extraction to 512 × 512 pixels. We used 512 × 512 patches instead of the commonly used 224 × 224, as larger inputs have shown improved performance in recent histopathology studies [23,24]. The results of previous EBV-ISH testing were recorded from the pathology reports.

A summary of the datasets used for tumor and EBV classification tasks is provided in Table 1. The datasets were organized into a two-stage pipeline based on the availability of EBV labeling via EBER-ISH. In stage 1 (tumor classification), a total of 373 biopsy samples without EBV labels were collected from SH ($n = 163$), CBMC ($n = 19$), and Dankook University Hospital ($n = 191$). These samples were used to train a patch-level tumor detection model using 5-fold cross-validation at the patient level, yielding 18,107 training patches and 4,153 validation patches. In stage 2 (EBV classification), a total of 744 biopsy samples with confirmed EBER-ISH status were collected from SH ($n = 732$; 121 EBV-positive), CBMC ($n = 9$; 2 EBV-positive), and DUH ($n = 3$;

Table 1. Summary of datasets used for tumor and Epstein–Barr virus (EBV) classification tasks.

Dataset	Task	Institution	Total	EBV status (pos/neg)
Internal validation	Tumor classification	Severance	163	-
		CHA Bundang	19	-
		Dankook	191	-
	Internal EBV classification	Severance	732	121/611
		CHA Bundang	9	2/7
		Dankook	3	2/1
External validation	EBV classification	Ewha	157	25/132

EBV status was determined using EBER *in situ* hybridization (EBER-ISH). Tumor classification datasets were unlabeled for EBV and used solely for training the tumor detection model (stage 1). All datasets included only biopsy specimens and were used for 5-fold cross-validation. External dataset was reserved for independent validation and consisted of biopsy samples.

2 EBV-positive), resulting in an overall EBV-positive rate of approximately 17%. Among these, 696 patients were allocated for training and 173 for validation, yielding 10,189 and 2,331 tumor patches, respectively, under a 5-fold cross-validation scheme.

For external validation, completely independent datasets were obtained from EWUM ($n = 157$; 25 EBV-positive), corresponding to EBV-positive rates of 16%. These slides yielded a total of 13,503 patches. Patch-level preprocessing was performed using an entropy-based background filtering method, in which patches with entropy < 2 were excluded. To mitigate inter-institutional staining variation, all training patches were subjected to stain normalization prior to model development.

Pathologist annotations for reference standards

To delineate tumor and non-tumor regions within gastric biopsy WSIs, five board-certified gastrointestinal pathologists (HK, M-SC, WL, H-CS, HK) manually annotated one or two representative tissue sections per slide ($n = 373$ WSIs; SH = 163, CBMC = 19, DUH = 191). Representative tissue pieces were marked as areas of interest. Tumor annotations were performed by segmenting regions of interest at pixel level. Tiles lacking annotations within the area of interest were designated as non-tumor or normal tissue. These annotations served as the ground truth for developing a tumor detection model using the phase 1 internal dataset. EBV status for each case was determined based on corresponding pathology reports of EBER-ISH, which served as the reference standard.

Proposed deep learning algorithm

The proposed EBV prediction pipeline is illustrated in Figure 1 and overall architecture consists of two main components: (1) tumor region filtering via patch-level classification using a self-supervised foundation model, and (2) final EBV classification using MBIL, which was developed based on the MIL framework. [20,25].

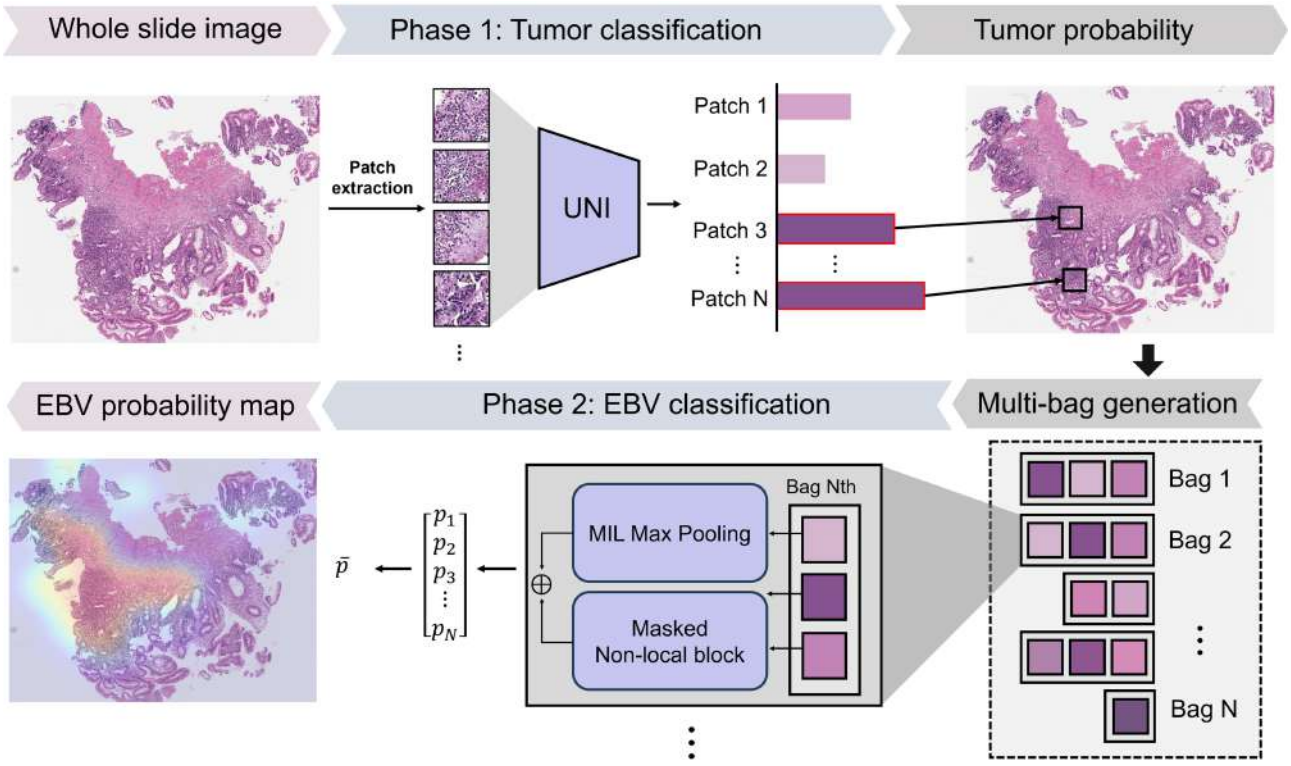

Figure 1. Overview of the proposed Epstein–Barr virus (EBV) prediction framework. In phase 1, whole-slide images (WSIs) are divided into patches and passed through a pretrained UNI model fine-tuned for patch-wise tumor classification. Only patches predicted as tumor, based on a predefined probability threshold, are retained for downstream analysis. In phase 2, the tumor patches are spatially clustered to form multiple tumor-centric bags. Each bag is processed through a multiple instance learning (MIL) model to generate instance-level predictions. The final EBV probability is computed by averaging the logits across all bags. A heatmap is then generated to visualize EBV-associated regions within the WSI.

Whole-slide biopsy images were divided into non-overlapping patches at 20× magnification, and the spatial coordinates of each patch were retained to enable spatial-aware modeling in subsequent steps.

Tumor classification using foundation model

To localize EBV prediction to diagnostically relevant regions, a tumor classifier based on a self-supervised pathology foundation model (UNI) [25] was trained to distinguish tumor from non-tumor patches. The training dataset consisted of manually annotated biopsy slides, in which gastrointestinal pathologists delineated tumor regions as ground truth. Only patches predicted as tumor above a predefined probability threshold were retained for downstream EBV classification. The threshold was set at 0.6, with its impact on EBV prediction performance examined in supplementary materials, Table S1. This filtering step was introduced to minimize the influence of histologically irrelevant areas and enhance the model's ability to capture EBV-related morphological features.

Multiple bag generation

Unlike traditional MIL methods that treat the entire slide as a single bag, MBIL clusters tumor patches using DBSCAN based on spatial proximity, forming multiple region-specific bags per slide [26]. Each bag is processed independently, allowing localized EBV patterns to be captured while minimizing information dilution. Patch features are extracted using the tumor classifier from the previous step.

Multiple bag instance learning framework

To effectively learn from spatially clustered biopsy patches, we developed the MBIL approach by leveraging the MIL framework to perform EBV prediction. In this framework, deep features are extracted for each spatially defined bag and processed through an MIL-based classifier. Instance-level predictions were independently computed within each bag, and a bag-level logit $\hat{y}_i$ was derived following the standard MIL mechanism. To obtain the final slide-level prediction, we averaged the outputs across all N bags and applied a sigmoid activation to produce the final EBV prediction:

$$\hat{y}_{\text{slide}} = \sigma\left(\frac{1}{N}\sum_{i=1}^{N}\hat{y}_i\right) \quad (1)$$

The binary cross-entropy (BCE) loss was then calculated using the ground truth label y and the aggregated prediction:

$$L_{\text{slide}} = \text{BCE}(y, \hat{y}_{\text{slide}}) \quad (2)$$

This bag-wise aggregation strategy enables the model to effectively capture heterogeneous EBV-related histologic patterns distributed across different biopsy regions, while mitigating over-reliance on any single dominant patch.

Training details

In phase 1, data augmentation for tumor classification was performed using a combination of random horizontal and vertical flips, Gaussian noise, and spatial transformations such as shifting, scaling, and rotation. To mitigate overfitting and address class imbalance in the dataset, we adopted a 5-fold cross-validation strategy using the internal dataset. All images were normalized and converted into tensors prior to training. The model was trained using the Adam optimizer with a learning rate of 5e-5 [27].

In phase 2, training on the EBV-labeled internal dataset was performed to ensure robustness and generalizability, using the same cross-validation scheme as in phase 1. The model was trained using the Adam optimizer with a learning rate of 1e-5 and a weight decay of 1e-4. Binary cross-entropy was used as the loss function, and training was performed with a batch size of 1 to accommodate instance-level predictions. All training procedures for both phases were conducted on an NVIDIA Quadro RTX 8000 GPU.

Statistical evaluation

To assess the diagnostic performance of the proposed EBV prediction model, we calculated standard classification metrics, including sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). Internal evaluation employed 5-fold cross-validation at the patient level, with results reported as mean ± standard deviation (SD) across folds. Receiver operating characteristic (ROC) curve analysis was performed to quantify the discriminative ability of each model, with the area under the ROC curve (AUC) used as the principal summary metric. For external validation, model outputs were generated by averaging the predictions across the five cross-validated models (ensemble averaging), and AUC comparisons between competing models were performed using the non-parametric DeLong test to determine statistical significance [28]. The resulting p values were reported, with $p < 0.05$ considered statistically significant. In addition, we compared the UNI-based model with other pretrained backbones (Swin Transformer [29], and GigaPath [30]). We evaluated the efficacy of the MBIL approach by comparing the proposed multi-bag (MB) strategy against the conventional single-bag (SB) MIL strategy.

Results

Prediction results of tumor classification task

Table 2 presents the results of three pretrained models – Swin Transformer, GigaPath, and UNI – used

Table 2. Performance comparison of pretrained models (Swin transformer, UNI, and GigaPath models) for tumor classification in Epstein–Barr virus (EBV) prediction.

Model	Accuracy	Sensitivity	Specificity	AUC
Swin transformer	0.915 ± 0.020	0.901 ± 0.022	0.926 ± 0.019	0.867 ± 0.019
GigaPath	0.972 ± 0.052	0.984 ± 0.025	0.962 ± 0.075	0.985 ± 0.031
UNI	0.978 ± 0.039	0.983 ± 0.031	0.974 ± 0.045	0.990 ± 0.020

AUC, area under the receiver operating characteristic curve. All values are reported as mean ± SD based on 5-fold cross-validation.

for patch-level tumor classification in the context of EBV prediction. Among the models, the UNI model achieved the highest overall performance, with an accuracy of 0.978 ± 0.039, sensitivity of 0.983 ± 0.031, specificity of 0.974 ± 0.045, and an AUC of 0.990 ± 0.020. It also exhibited the most balanced performance across both PPV (0.970 ± 0.051) and NPV (0.986 ± 0.027). The GigaPath model also demonstrated strong classification capability, particularly in sensitivity (0.984 ± 0.025) and AUC (0.985 ± 0.031), though it showed slightly greater variability in specificity (0.962 ± 0.075) and PPV (0.960 ± 0.078). In contrast, the Swin Transformer yielded comparatively lower performance across most metrics, with an AUC of 0.867 ± 0.019, despite maintaining relatively high specificity (0.926 ± 0.019) and NPV (0.926 ± 0.022).

Prediction results of EBV classification task with internal dataset

Based on Table 3, internal validation demonstrated that the MB strategy generally improved or maintained performance across all evaluated MIL-based models such as CLAM, DSMIL, and TransMIL [19–21]. In particular, DSMIL-MB achieved the highest AUC (0.907 ± 0.044), along with the best sensitivity (0.806 ± 0.038) and specificity (0.849 ± 0.071), confirming its robust discriminative performance in detecting EBV-positive cases. Compared with its SB counterpart, TransMIL-MB showed a substantial improvement in AUC (0.844 ± 0.055 versus 0.745 ± 0.090) and sensitivity (0.704 ± 0.070 versus 0.557 ± 0.253), underscoring the benefit of multiple bag generation for models that initially performed less effectively. While CLAM-MB and CLAM-SB demonstrated similar AUCs (0.840 ± 0.050 versus 0.845 ± 0.017), the MB strategy enhanced sensitivity (0.740 ± 0.029 versus 0.672 ± 0.072), suggesting that MB input may improve detection power without compromising specificity. Overall, MBIL strategy using MB led to more balanced sensitivity and specificity profiles and higher AUC values, supporting the effectiveness of spatial subdivision in preserving local morphological signals during MIL-based EBV prediction from gastric biopsy slides. For reference, comparative performance of previously published EBV prediction models is summarized in Table 4. Representative attention-highlighted patches from EBV prediction are shown in Figure 2. High-probability regions (Figure 2A,B) displayed dense intratumoral lymphocytic infiltration and lace-like patterns consistent with previously described EBV-associated morphology, whereas low-probability regions (Figure 2C,D) corresponded to well-differentiated carcinoma with sparse lymphoid stroma.

Prediction results of EBV classification task with external dataset

External validation on EWUM demonstrated that the MB strategy provided more balanced performance across all MIL-based models compared with their SB counterparts (Figure 3). Among the three architectures, the DSMIL-MB model achieved the highest AUC (0.870), followed by CLAM-MB (0.865) and TransMIL-MB (0.853). All MB models showed improved or at least comparable sensitivity relative to their SB versions, with DSMIL-MB reaching the highest sensitivity (0.800), thereby enhancing the detection of EBV-positive cases. In contrast, CLAM-SB and DSMIL-SB exhibited identical sensitivities (0.720), but DSMIL-SB yielded a slightly lower AUC (0.856 versus 0.865), indicating improved discriminative performance with MB aggregation. Notably, TransMIL-SB displayed the weakest baseline performance (sensitivity: 0.560, AUC: 0.701), but its performance markedly improved with MB (sensitivity: 0.680, AUC: 0.853), highlighting

Table 3. Comparison of Epstein–Barr virus (EBV) prediction performance using single bag and multiple bag approaches across TransMIL, CLAM, and DSMIL models with internal validation.

Model	Accuracy	Sensitivity	Specificity	AUC
TransMIL-SB	0.750 ± 0.055	0.557 ± 0.253	0.787 ± 0.104	0.745 ± 0.090
CLAM-SB	0.792 ± 0.040	0.672 ± 0.072	0.815 ± 0.062	**0.845 ± 0.017**
DSMIL-SB	0.799 ± 0.035	0.804 ± 0.063	0.797 ± 0.035	0.882 ± 0.037
TransMIL-MB	0.761 ± 0.081	0.704 ± 0.070	0.770 ± 0.108	**0.844 ± 0.055**
CLAM-MB	0.769 ± 0.049	0.740 ± 0.029	0.774 ± 0.063	0.840 ± 0.050
DSMIL-MB	0.842 ± 0.056	0.806 ± 0.038	0.849 ± 0.071	**0.907 ± 0.044**

AUC, area under the receiver operating characteristic curve; MB, multiple bag; SB, single bag. All values are reported as mean ± SD based on 5-fold cross-validation. Bold values indicate the higher AUC between the single-bag and multiple-bag versions of each model.

Table 4. Comparison of previously published Epstein–Barr virus (EBV) prediction models based on resection specimens.

Study	Specimen	Number of patients (training set)	Training dataset	External validation	Evaluation metrics
Jeong *et al* (Scientific Reports, 2022)	Resection	427	TCGA+ISH	Hanyang University Guri Hospital	Internal: ACC 0.99 External: AUC 0.88
Muti *et al* (Lancet Digital Health, 2021)	Resection	2,685	Berlin + LMU Munich	6-center cohort	Internal: AUC 0.86 External: AUC 0.672–0.897
Zhang *et al* (Cancers, 2021)	Resection	427	TCGA	None	Internal: AUC 0.85
Le Vuong, TT *et al* (JAMA Netw Open, 2022)	Resection	Image patches: 137,184 TMA cores: 708 WSI: 24	Kangbuk Samsung Hospital (WSI + TMA)	Kangbuk Samsung Hospital (biopsy)	Internal: ACC 0.94 External: AUC 0.87
Zheng *et al* (Nat Commun, 2022)	Resection	727	1-center cohort	1. MultiCenter-STAD cohort 2. TCGA-STAD	Internal: AUC 0.96 External 1: AUC 0.94 External 2: AUC 0.89
Present study	Biopsy	744	3-center cohort	1-center cohort	Internal: AUC 0.91 External: AUC 0.87

ACC, accuracy; AUC, area under the receiver operating characteristic curve. ACC was calculated at the patch level, not the patient or slide level, while AUC was computed at the patient level. EBV status labels in TCGA datasets were derived from RNA-seq-based classification, whereas labels denoted as ISH were based on EBER *in situ* hybridization (EBER-ISH), the diagnostic gold standard.

the particular advantage of MB for weaker base models. Statistical comparison using the DeLong test confirmed that the AUC improvement of TransMIL from 0.701 (SB) to 0.853 (MB) reached significance ($p = 0.05$), while the differences between SB and MB for CLAM ($p = 0.633$) and DSMIL ($p = 0.687$) were not statistically significant. Collectively, these findings demonstrate that MB consistently enhanced both sensitivity and AUC, supporting its effectiveness for biopsy-level EBV prediction in the external validation setting.

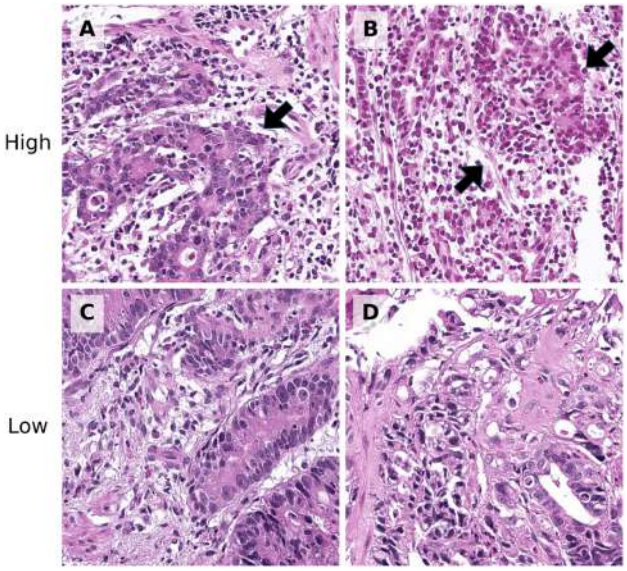

Figure 2. Representative patches from the cases with high and low probability of Epstein–Barr virus (EBV)-positivity. (A, B) are patches from cases with high probability of EBV-positivity; (A) probability 0.790, (B) probability 0.867. (A, B) show lace-like features with intratumoral lymphocytic infiltration (arrows), (C, D) are patches from low probability of EBV-positivity; (C) probability 0.130, (D) probability 0.066. (C, D) show well to moderately differentiated carcinomas with less lymphocytic infiltration.

Qualitative comparison of attention maps across models

The heatmap visualizations in Figure 4 illustrate the spatial attention maps generated by three different MIL-based models – CLAM, DSMIL, and the proposed MBIL with DSMIL – across three representative EBV-positive gastric biopsy cases (case A, B, and C). Each row corresponds to the model used, and each column shows a different case. The top row presents the original H&E-stained ROI images for reference. These visualizations support the interpretability advantage of the proposed MB strategy. By preserving spatial context through clustered bag generation, the model can better associate focal histopathologic patterns with EBV status. This provides both quantitative improvements and qualitative interpretability, which is essential for clinical adoption of AI-assisted diagnostic tools in pathology.

Discussion

In this study, we explored the feasibility of a MBIL approach, which integrates an MB strategy with pathology foundation models for EBV status prediction in gastric cancer biopsies. This methodological refinement seeks to address the practical constraints of conventional SB MIL by prioritizing spatial context preservation through clustered bag generation. To enhance the generalizability of our findings, we employed a multi-institutional training dataset comprising 744 cases from three centers, which reflects real-world variability in patient populations, staining protocols, and scanning systems. By leveraging foundation models pre-trained on large-scale pathology datasets, our framework learns transferable feature representations that help mitigate the inherent class imbalance (EBV-positive rate of approximately 17%). Our systematic comparison of multiple MIL architectures – including TransMIL, CLAM, and

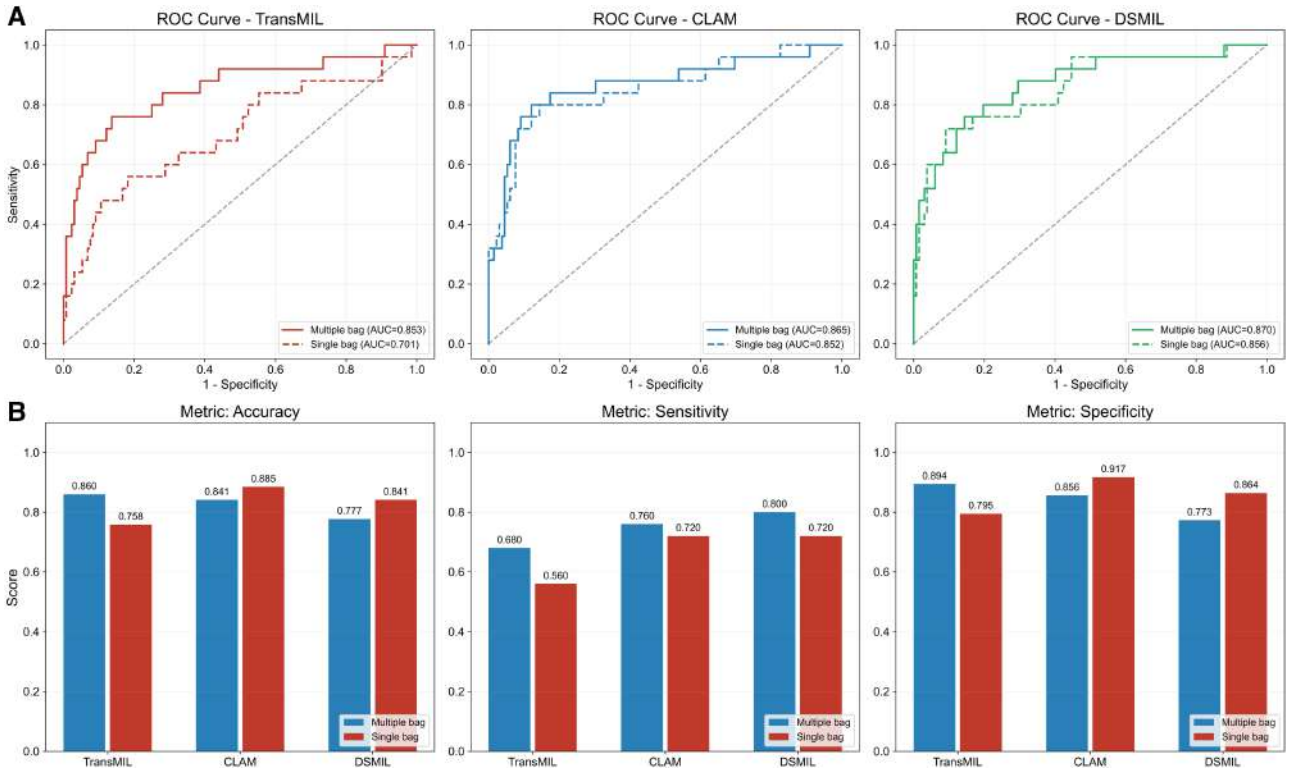

Figure 3. Comparison of single-bag (SB) and multiple-bag (MB) frameworks for Epstein–Barr virus (EBV) status prediction in gastric cancer using multiple-instance learning (MIL) models. (A) Receiver operating characteristic (ROC) curves of TransMIL, CLAM, and DSMIL. Solid lines represent MB models and dashed lines represent SB models. (B) Bar plots comparing accuracy, sensitivity, and specificity across SB (red) and MB (blue) frameworks for each model. Improvements with the MB approach are model-dependent and vary across metrics. AUC, area under the receiver operating characteristic curve.

DSMIL – supports the consistency and robustness of the MBIL approach across multi-institutional validation cohorts. These results suggest that even models with initially lower performance can achieve a more balanced specificity profile by effectively capturing focal, EBV-related histologic details that might otherwise be attenuated in standard MIL applications.

Recent studies have validated the feasibility of using deep learning to predict EBV status in gastric cancer from histopathologic images. Muti *et al* developed ShuffleNet-based classifiers with transfer learning on pooled multicentre cohorts (2,685 patients), showing variable but generally strong performance across institutions [13]. Jeong *et al* proposed EBVNet, a two-stage model combining tumor classification and EBV prediction on 319 TCGA whole-slide images, achieving high accuracy on external validation [12]. Zheng *et al* used an ensemble of five ResNet50 classifiers with human–machine fusion strategies across three cohorts, achieving top-tier performance on both internal and external validation [10]. Zhang *et al* applied a ResNet18-based model on 400 TCGA diagnostic slides, highlighting that model predictions could be influenced by lymphocyte infiltration patterns rather than direct viral detection [11]. However, these studies were primarily conducted on gastrectomy specimens, which differ substantially from small biopsy tissues encountered in routine clinical practice. The performance of our framework is consistent with previously reported studies, demonstrating its feasibility for the specific challenges associated with clinically relevant biopsy specimens. While some resection-based studies report higher AUCs [10,13], these approaches are limited to post-surgical cases and may not reflect the diagnostic challenges encountered in routine clinical practice. Our biopsy-based approach seeks to address a clinical need for preoperative molecular stratification. Compared with other biopsy-based studies, such as the work by Le Vuong *et al*, our framework demonstrates competitive performance while prioritizing the preservation of spatial context. Furthermore, the attention heatmaps generated by our system provide interpretability that may assist pathologists in identifying regions of interest for further examination. Although further studies are needed to confirm correlations with specific histologic features, these patterns show promise for highlighting diagnostically relevant areas. To better understand the model's failure modes and address concerns regarding biological specificity, we reviewed representative false positive, false negative, and true negative cases with prominent lymphocytic infiltration (supplementary material, Figure S1). The false positive case – an EBV-negative tumor with dense lymphocytic infiltration – directly illustrates the confounding effect of lymphoid-rich stroma on model predictions. High-attention regions in this case co-localized with areas of immune infiltration

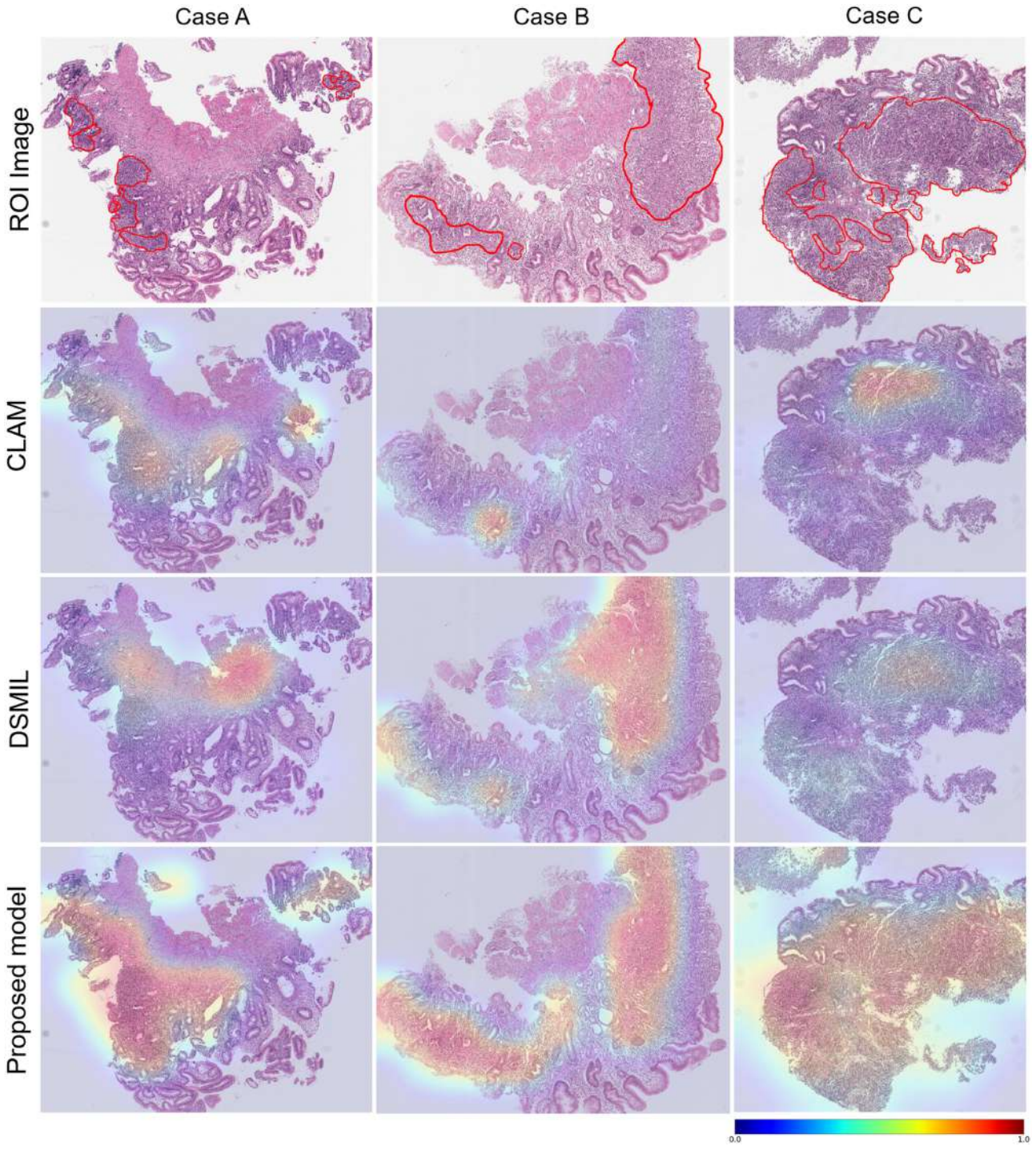

Figure 4. Visual comparison of attention heatmaps generated by three multiple-instance learning (MIL) models – CLAM, DSMIL, and proposed multiple bag with DSMIL – for three representative biopsy cases (A–C). The top row shows the original H&E-stained region of interest (ROI) images. Annotated areas are the tumor areas. Subsequent rows visualize instance-level attention scores overlaid as heatmaps, highlighting regions considered important for Epstein–Barr virus (EBV) prediction.

morphologically similar to EBV-associated gastric cancers, suggesting that the model partially leverages immune microenvironment features as a surrogate for EBV status. The false negative case displayed relatively well to moderately differentiated adenocarcinoma morphology with limited lymphoid stroma, reflecting an absence of the immune-related histologic features the model relies upon, which led to an incorrect EBV-negative prediction despite true positivity. Importantly, the true negative case with prominent lymphocytic infiltration – including areas with poorly differentiated carcinoma component not compatible with typical EBVaGC – demonstrated focally elevated patch-level attention scores in lymphoid-rich regions, yet was correctly classified as EBV-negative at the slide level (predicted probability: 0.238). The direct comparison

between the false positive and this true negative case, both featuring lymphoid-rich morphology but differing in slide-level outcome, suggests that the model's slide-level approach captures spatially extensive immune infiltration patterns that extend beyond focal lymphocytic aggregates alone. Collectively, these observations indicate that the model's predictions are substantially influenced by immune-related histologic patterns – a limitation inherent to H&E-based prediction in the absence of a direct viral signal. However, the surrogate features exploited by the model are biologically plausible, as prominent lymphocytic infiltration is a hallmark of EBV-positive gastric cancer. Future work incorporating spatial immune profiling or ISH-guided training labels may help further disentangle EBV-specific morphology from immune-mediated surrogate features.

We acknowledge several limitations, most notably the relatively modest size of our external validation cohort ($n = 157$). While this cohort provided an independent assessment, its size may limit the statistical power to capture the full spectrum of morphological variations. Furthermore, the inherent dataset imbalance, with an EBV-positive rate of approximately 17%, presents a challenge for real-world deployment. In clinical settings with lower prevalence, this imbalance could potentially lead to a higher rate of false positives, necessitating careful calibration of decision thresholds. Additionally, although we utilized multi-institutional data to capture institutional variability in staining and scanning, differences in laboratory protocols across a broader range of clinical sites could still affect the model's generalizability. Future prospective studies with larger, more diverse cohorts are required to further validate the robust performance of our system in varied real-world environments. The two-stage pipeline design, while computationally efficient, could benefit from end-to-end optimization. Future iterations may explore integrated approaches that simultaneously perform tumor detection and EBV prediction, potentially improving overall system performance and reducing computational complexity.

Our approach is designed to complement rather than replace existing diagnostic methods. We envision a phased implementation where our system is first applied as a screening tool to triage cases for confirmatory EBER-ISH testing, improving workflow efficiency while preserving diagnostic accuracy through gold-standard confirmation. To operationalize this workflow, we evaluated a clinically oriented decision threshold on the external validation cohort. At a probability threshold of 0.29, optimized to achieve sensitivity ≥ 0.88, the model achieved a sensitivity of 0.88, specificity of 0.70, and NPV of 0.97. In a proposed two-step triage workflow, model-negative cases (score < 0.29) could safely forego EBER-ISH testing, given the high NPV suggesting fewer than 3.2% of such cases would be true EBV-positive. Model-positive cases would proceed to confirmatory EBER-ISH, preserving diagnostic accuracy. Although the false-positive rate at this threshold is approximately 30%, this may be clinically acceptable in the context of a screening paradigm – false positives incur only the cost of an additional EBER-ISH test, while false negatives risk missing an actionable EBV-positive diagnosis. Integration into routine workflows is feasible given the increasing adoption of digital pathology platforms and the reliance on routine H&E staining without additional reagents. As validation data accumulate and institutional protocols mature, the threshold and its associated operating characteristics can be prospectively calibrated to local prevalence and resource constraints. Furthermore, quality assurance can be adapted from existing laboratory standards, with our multi-institutional validation providing a foundation for clinical performance benchmarks. This work demonstrates the potential for spatial context preservation in MIL applications, with implications extending beyond EBV prediction to other biomarker assessment tasks in computational pathology. The MBIL framework could be readily adapted to predict other molecular subtypes or therapeutic targets from routine histology, potentially extending diagnostic accessibility in diverse clinical settings. Our use of foundation models underscores the value of large-scale pretraining for medical AI applications, particularly in domains with limited labeled data. This approach provides a pathway for developing robust diagnostic tools even for relatively rare molecular subtypes.

In conclusion, we have demonstrated the feasibility of predicting EBV status directly from gastric biopsy specimens by employing a MB strategy alongside a pathology foundation model. This methodological refinement addresses the constraints of conventional MIL frameworks by effectively preserving spatial context, resulting in competitive performance across multi-institutional, real-world data. While the relatively modest size of our external cohort necessitates further validation in larger prospective settings, our results provide supportive evidence for the implementation of this approach as a preoperative triage tool. By potentially reducing the reliance on specialized molecular testing, this framework offers a scalable basis for enhancing diagnostic accessibility and advancing treatment personalization in gastric cancer.

Acknowledgements

This study was funded by The Korea Health Industry Development Institute (KHIDI) (Grant No. HI21C0977). We would like to express our sincere gratitude to Professor Faisal Mahmood's lab at the Brigham and Women's hospital for generously providing the UNI algorithm employed in this study. We also wish to thank Mr Damin Moon of JLK Artificial Intelligence R&D Center, Seoul, Korea for his valuable contribution to data management and technical assistance throughout the study.

Author contributions statement

HaK, HyK and BHC conceptualized the study. BHC, HaK, and HyK developed the methodology. YN, BHC and DN contributed to software development. Validation and formal analysis were performed by YN, BHC, HaK and HyK. Investigation was conducted by WAL, MSC, HCS, YJP, HaK, HyK, YN, BHC and HCK. WAL, MSC, HCS, YJP, HaK, HyK, HJA and NHC provided resources. Data curation was carried out by WAL, MSC, HCS, YJP, HaK, and HyK. YN, HaK, HyK and BHC wrote the original draft. Visualization, supervision, and project administration were managed by YN, HaK, HyK, BHC and NHC. NHC acquired funding for the study. All authors reviewed, edited, and approved the final version of the manuscript.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Park JH, Kim EK, Kim YH, *et al*. Epstein–Barr virus positivity, not mismatch repair-deficiency, is a favorable risk factor for lymph node metastasis in submucosa-invasive early gastric cancer. *Gastric Cancer* 2016; **19:** 1041–1051.
2. Lee JH, Kim SH, Han SH, *et al*. Clinicopathological and molecular characteristics of Epstein–Barr virus-associated gastric carcinoma: a meta-analysis. *J Gastroenterol Hepatol* 2009; **24:** 354–365.
3. Camargo MC, Kim W-H, Chiaravalli AM, *et al*. Improved survival of gastric cancer with tumour Epstein–Barr virus positivity: an international pooled analysis. *Gut* 2014; **63:** 236–243.
4. Uemura Y, Tokunaga M, Arikawa J, *et al*. A unique morphology of Epstein-Barr virus-related early gastric carcinoma. *Cancer Epidemiol Biomarkers Prev* 1994; **3:** 607–611.
5. Weiss LM, Chen Y-Y. EBER in situ hybridization for Epstein–Barr virus. *Hematological Malignancies* 2013; **999:** 9223–9230.
6. Barragán-Montero A, Javaid U, Valdés G, *et al*. Artificial intelligence and machine learning for medical imaging: a technology review. *Phys Med* 2021; **83:** 242–256.
7. Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; **42:** 60–88.
8. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform* 2016; **7:** 29.
9. Coudray N, Ocampo PS, Sakellaropoulos T, *et al*. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24:** 1559–1567.
10. Zheng X, Wang R, Zhang X, *et al*. A deep learning model and human-machine fusion for prediction of EBV-associated gastric cancer from histopathology. *Nat Commun* 2022; **13:** 2790.
11. Zhang B, Yao K, Xu M, *et al*. Deep learning predicts EBV status in gastric cancer based on spatial patterns of lymphocyte infiltration. *Cancers (Basel)* 2021; **13:** 6002.
12. Jeong Y, Cho CE, Kim J-E, *et al*. Deep learning model to predict Epstein–Barr virus associated gastric cancer in histology. *Sci Rep* 2022; **12:** 18466.
13. Muti HS, Heij LR, Keller G, *et al*. Development and validation of deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric cancer: a retrospective multicentre cohort study. *Lancet Digit Health* 2021; **3:** e654–e664.
14. Kettritz U. Minimally invasive biopsy methods–diagnostics or therapy? Personal opinion and review of the literature. *Breast Care* 2011; **6:** 94–97.
15. Vargas HI, Masood S. Implementation of a minimally invasive breast biopsy program in countries with limited resources. *Breast J* 2003; **9:** S81–S85.
16. Le Vuong TT, Song B, Kwak JT, *et al*. Prediction of Epstein-Barr virus status in gastric cancer biopsy specimens using a deep learning algorithm. *JAMA Netw Open* 2022; **5:** e2236408-e2236408.
17. Abusalah MAH, Gan SH, Al-Hatamleh MA, *et al*. Recent advances in diagnostic approaches for epstein–barr virus. *Pathogens* 2020; **9:** 226.
18. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*. PMLR: Stockholm, 2018; 2127–2136.
19. Lu MY, Williamson DF, Chen TY, *et al*. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021; **5:** 555–570.
20. Li B, Li Y, Eliceiri KW. 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*. IEEE: Piscataway, 2021; 14318–14328.
21. Shao Z, Bian H, Chen Y, *et al*. Transmil: transformer based correlated multiple instance learning for whole slide image classification. *Adv Neural Inf Process Syst* 2021; **34:** 2136–2147.
22. Hense J, Jamshidi Idaji M, Eberle O, *et al*. XMIL: insightful explanations for multiple instance learning in histopathology. *Adv Neural Inf Process Syst* 2024; **37:** 8300–8328.
23. Vong CK, Wang A, Dragunow M, *et al*. Brain tumour histopathology through the lens of deep learning: a systematic review. *Comput Biol Med* 2025; **186:** 109642.
24. Shaban M, Awan R, Fraz MM, *et al*. Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Trans Med Imaging* 2020; **39:** 2395–2405.
25. Chen RJ, Ding T, Lu MY, *et al*. Towards a general-purpose foundation model for computational pathology. *Nat Med* 2024; **30:** 850–862.
26. Ester M, Kriegel H-P, Sander J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press: Menlo Park, 1996; 226–231.
27. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv* preprint arXiv:14126980; 2014.
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44:** 837–845.
29. Liu Z, Lin Y, Cao Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE: Piscataway, 2021; 10012–10022.
30. Xu H, Usuyama N, Bagga J, *et al*. A whole-slide foundation model for digital pathology from real-world data. *Nature* 2024; **630:** 181–188.

SUPPLEMENTARY MATERIAL ONLINE

Figure S1. Representative cases illustrating model attention patterns across different prediction outcomes

Table S1. Impact of bag-level tumor probability thresholds on the performance of EBV prediction across various MIL models