

A foundation-model-powered AI system with post-read alerts and concurrent assistance improves gastric biopsy diagnosis across 29 pathological categories

Chi-Chung Chen

`chenchc@aetherai.com`

aetherAI Co., Ltd. <https://orcid.org/0000-0003-0267-7704>

Shih-Chiang Huang

Chang Gung Memorial Hospital <https://orcid.org/0000-0002-5092-4434>

Tu-Yung Chang

<https://orcid.org/0009-0007-8589-7775>

Chen-Lin Chi

Chiayi Chang Gung Memorial Hospital

Huei-Chieh Chuang

Chiayi Chang Gung Memorial Hospital

Jui Lan

Department of Anatomic Pathology, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University, College of Medicine

Pei-Hua Chang

Linkou Chang Gung Memorial Hospital

Cheng-Hsi Lin

Linkou Chang Gung Memorial Hospital

Yi-Yin Hsieh

New Taipei Municipal Tu Cheng Hospital (Built and Operated by Chang Gung Medical Foundation)

Tzu-Chien Cheng

Keelung Chang Gung Memorial Hospital

Chao-Yuan Yeh

aetherAI <https://orcid.org/0000-0002-1947-1596>

Tse-Ching Chen

Chang Gung Memorial Hospital <https://orcid.org/0000-0003-2532-7328>

Article

Keywords:

Posted Date: July 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10103602/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. C.-Y.Y. is the Founder, Chairman, and Chief Executive Officer of aetherAI. C.-C.C. is the Director of Product and Machine Learning at aetherAI. The remaining authors declare no competing interests.

Abstract

Gastric biopsy interpretation requires simultaneous assessment of malignancies, precancerous lesions, infections, and benign conditions, yet existing artificial intelligence models address only two to three diagnostic categories and have been evaluated under a single mode of clinical integration. Here we develop a deep learning model that classifies gastric biopsies into 29 pathological categories spanning malignancy subtypes, dysplasia, metaplasia, *Helicobacter pylori*, and benign findings. The model leverages a vision foundation model pre-trained on 70,618 whole slide images and fine-tuned end-to-end on 10,462 gastric biopsies without pixel-level annotations. On internal validation ($n = 2,957$), the model achieves an area under the curve of 0.9945 for malignancy detection. We evaluate two paradigms for clinical integration: a post-read alert system and a concurrent assistance system. In retrospective simulation across four independent hospital cohorts ($n = 4,679$), the post-read system identifies 17 malignancies that were missed or miscoded in routine diagnostic records. In a prospective reader study with six pathologists, concurrent AI assistance significantly improves sensitivity for malignancy (92.2% to 94.8%, $P = .034$) and metaplasia (94.2% to 97.4%, $P = .013$). Furthermore, the foundation model fine-tuned on only 400 slides outperforms a conventional model trained on over 10,000 slides, demonstrating extreme data efficiency. In addition, the model produces pixel-resolution lesion localization across all 29 categories without any pixel-level annotations, providing the heatmaps used in the concurrent assistance system. These results establish a framework for dual-mode AI deployment in gastric biopsy pathology—concurrent assistance for real-time diagnostic support and post-read alerts for retrospective quality assurance.

Introduction

Gastric cancer ranks as the fifth most common malignancy worldwide, with approximately one million new cases annually and disproportionately high incidence in East Asia [1]. Histopathological detection at an early stage remains clinically paramount. A critical function of gastric biopsy interpretation involves the definitive differentiation between benign and malignant erosive or ulcerative lesions. Intestinal-type gastric adenocarcinoma usually develops through the well-characterized Correa precancerous cascade, a continuous progression from chronic gastritis through intestinal metaplasia and dysplasia to invasive carcinoma [2]—creating a window of opportunity in which early histopathological detection of precursor lesions can interrupt disease progression [2, 3]. Gastric biopsy remains the gold standard for this purpose, yet the task imposes a substantial workload on pathologists, particularly in high-prevalence regions where screening programs generate large volumes of specimens [4]. The diagnostic challenge is compounded by the diversity of findings that must be recognized—spanning malignant subtypes, precancerous lesions, *Helicobacter pylori* (*H. pylori*) infection, and numerous benign conditions—and by interobserver variability, especially for borderline entities such as low-grade dysplasia [5].

AI-assisted digital pathology has shown promise in gastric biopsy classification, but existing models are limited in scope and workflow integration. Most studies address only two to three diagnostic categories (e.g., malignant vs. non-malignant vs. adenoma) [6–9]. Reader studies have demonstrated that AI assistance can improve pathologists' sensitivity for gastric cancer detection and reduce review time [9, 10], yet these evaluations have examined only a single mode of AI integration—simultaneous assistance during slide

review. In other organ systems, prescreening and post-read paradigms have also been explored. Bilal et al. developed a prescreening tool for colorectal biopsies capable of filtering approximately 80% of normal cases at 95% sensitivity [11], and studies on prostate biopsy AI have compared prescreening, concurrent reading, and post-read workflows, finding distinct trade-offs between efficiency and accuracy for each paradigm [12]. However, no study in gastric biopsy pathology has empirically evaluated more than one workflow integration mode within the same framework.

On the technical front, training multi-class biopsy classifiers traditionally requires either extensive annotations (strongly supervised methods) or very large WSI collections (weakly supervised methods such as multiple instance learning) [13–16]. Vision Foundation Models (VFMs), pre-trained on diverse datasets through self-supervised learning, have recently demonstrated the potential to substantially reduce these data requirements when fine-tuned for downstream pathology tasks [17, 18], though this has not been quantified for gastric biopsy classification. Architectures such as ResNet [19] paired with self-supervised learning methods like Barlow Twins [20] have enabled such pre-training at scale. Equally important for clinical deployment is the ability to localize lesions within gigapixel images, which traditionally requires pixel- or region-level annotations that are prohibitively costly to obtain at scale.

In this study, we address these gaps through four contributions. First, we developed a deep learning model for multi-label classification of gastric biopsies into 29 pathological categories, leveraging a VFM pre-trained on over 70,000 WSIs and fine-tuned with the Enhanced Streaming CNN (ESCNN) [16] method, and validated it on an internal test set and four external hospital cohorts. Second, we designed and evaluated two paradigms for clinical AI integration: a post-read alert system that retrospectively flags potential malignancies missed on initial diagnosis, and a concurrent assistance system that provides real-time heatmaps during slide review. A prospective reader study with six pathologists and a three-week washout period was conducted to quantify the impact on diagnostic performance and efficiency. To our knowledge, this is the first study in gastric biopsy pathology to prospectively evaluate both workflow modes within the same framework. Third, we quantified the data efficiency of VFM fine-tuning through few-shot experiments, demonstrating that a VFM trained on 400 slides outperformed a conventional model trained on over 10,000 slides. Fourth, we demonstrated that pixel-resolution localization across all 29 categories can be obtained without any pixel-level supervision, by combining streaming convolution with ReLSE pooling [21] to produce Class Activation Maps (CAM) [22] directly from slide-level labels. Together, these contributions establish the foundational building blocks—data-efficient subtype classification and annotation-free localization—that future agentic AI systems will require to autonomously orchestrate pathology workflows.

Results

Gastric biopsy classification model for 29 classes

For the multi-label classification of gastric biopsies into 29 distinct pathological categories, a deep learning model was trained using a training dataset of 10,462 whole slide images (WSIs) from Linkou CGMH (2020–2023), with each image downsampled to 10× magnification. The model utilizes a ResNet50 [19] architecture, which was pre-trained on a large-scale dataset of 70,618 WSIs from multiple sites and specimen types using

Barlow Twins contrastive learning [20]. Subsequently, this VFM was fine-tuned for the 29-class task in an end-to-end, weakly supervised manner with the ESCNN method [16]. The classification labels encompassed a hierarchy of findings, from major types like malignancy and dysplasia to specific subtypes such as signet ring carcinoma and neuroendocrine tumor (Fig. 1a).

The model's performance was validated on an independent internal test set of 2,957 WSIs, with ground truth labels established from SNOMED codes within the institutional Laboratory Information System (LIS). In the critical task of identifying malignancy, the model achieved an outstanding area under the receiver operating characteristic curve (AUC) of 0.9945 (95% CI: 0.9878–1.0000). A comprehensive breakdown of AUCs for all 29 classes is detailed in Table 1. These categories—malignant, dysplasia, metaplasia, and *H. pylori*—were selected as they represent the most significant findings for patient management, spanning from critical malignant and precancerous lesions to treatable risk factors and key benign differentials. The model demonstrated excellent discriminatory performance across all these key classes, achieving high AUCs for Dysplasia (0.9707), Metaplasia (0.9489), and *H. pylori* (0.9210), in addition to the near-perfect score for malignancy. This consistent high performance highlights the model's potential to reliably assist pathologists in identifying a wide spectrum of clinically relevant gastric pathologies.

Table 1

Performance of the gastric biopsy classification model for all 29 pathological classes on the internal test set. Metrics include Area Under the Curve (AUC), sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). The 95% Confidence Intervals (CIs) for AUCs were calculated using DeLong's method, while CIs for all other metrics were calculated using Wilson's method. The indices in class names denote their position in the diagnostic hierarchy.

Class	N (positive)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
High priority						
1. Malignant	2957 (58)	0.9945 (0.9878–1.0000)	0.9138 (0.8136–0.9626)	0.9976 (0.9950–0.9988)	0.8833 (0.7782–0.9423)	0.9983 (0.9960–0.9993)
1.1. Carcinoma	2957 (44)	0.9986 (0.9964–1.0000)	0.9091 (0.7884–0.9641)	0.9993 (0.9975–0.9998)	0.9524 (0.8421–0.9868)	0.9986 (0.9965–0.9995)
1.1.1. Adenocarcinoma	2956 (41)	0.9985 (0.9961–1.0000)	0.9024 (0.7745–0.9614)	0.9990 (0.9970–0.9996)	0.9250 (0.8014–0.9742)	0.9986 (0.9965–0.9995)
1.1.1.1. Signet ring cell carcinoma	2919 (4)	0.9873 (0.9626–1.0000)	1.0000 (0.5101–1.0000)	0.9496 (0.9410–0.9569)	0.0265 (0.0103–0.0661)	1.0000 (0.9986–1.0000)
1.1.2. Squamous cell carcinoma	2956 (2)	0.9998 (0.9994–1.0000)	1.0000 (0.3424–1.0000)	0.9997 (0.9981–0.9999)	0.6667 (0.2077–0.9385)	1.0000 (0.9987–1.0000)
1.1.3. Lymphoma	2957 (5)	0.9999 (0.9997–1.0000)	1.0000 (0.5655–1.0000)	0.9997 (0.9981–0.9999)	0.8333 (0.4365–0.9699)	1.0000 (0.9987–1.0000)
1.1.3.1. Diffuse large B-cell lymphoma	2956 (3)	0.9992 (0.9980–1.0000)	1.0000 (0.4385–1.0000)	0.9983 (0.9960–0.9993)	0.3750 (0.1368–0.6943)	1.0000 (0.9987–1.0000)
1.1.3.2. Marginal zone B-cell lymphoma	2956 (1)	0.9980 (N/A)	1.0000 (0.2065–1.0000)	0.9980 (0.9956–0.9991)	0.1429 (0.0257–0.5131)	1.0000 (0.9987–1.0000)
1.1.4. Neuroendocrine tumor	2957 (3)	0.9882 (0.9665–1.0000)	1.0000 (0.4385–1.0000)	0.9661 (0.9590–0.9721)	0.0291 (0.0100–0.0822)	1.0000 (0.9987–1.0000)
1.1.5. Sarcoma	2957 (6)	0.9992 (0.9977–1.0000)	1.0000 (0.6097–1.0000)	0.9956 (0.9925–0.9974)	0.3158 (0.1536–0.5399)	1.0000 (0.9987–1.0000)
1.1.5.1. Gastrointestinal stromal tumor	2957 (5)	0.9999 (0.9995–1.0000)	1.0000 (0.5655–1.0000)	0.9993 (0.9975–0.9998)	0.7143 (0.3589–0.9178)	1.0000 (0.9987–1.0000)
Medium priority						
2. Atypia	2900 (3)	0.7554 (0.5125–0.9982)	1.0000 (0.4385–1.0000)	0.6003 (0.5823–0.6180)	0.0026 (0.0009–0.0076)	1.0000 (0.9978–1.0000)

Class	N (positive)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
3. <i>H. pylori</i>	2899 (32)	0.9210 (0.8940–0.9480)	0.9063 (0.7578–0.9676)	0.8116 (0.7969–0.8255)	0.0510 (0.0357–0.0722)	0.9987 (0.9962–0.9996)
4. Benign tumor	2900 (6)	1.0000 (1.0000–1.0000)	1.0000 (0.6097–1.0000)	1.0000 (0.9987–1.0000)	1.0000 (0.6097–1.0000)	1.0000 (0.9987–1.0000)
5. Dysplasia	2899 (26)	0.9707 (0.9494–0.9920)	0.9231 (0.7586–0.9786)	0.8723 (0.8596–0.8840)	0.0614 (0.0416–0.0897)	0.9992 (0.9971–0.9998)
5.1. Adenoma	2893 (2)	0.9983 (0.9968–0.9998)	1.0000 (0.3424–1.0000)	0.9983 (0.9960–0.9993)	0.2857 (0.0822–0.6411)	1.0000 (0.9987–1.0000)
5.2. Mild dysplasia	2893 (11)	0.9571 (0.9281–0.9860)	0.9091 (0.6226–0.9838)	0.8869 (0.8748–0.8979)	0.0298 (0.0162–0.0539)	0.9996 (0.9978–0.9999)
5.3. Severe dysplasia	2893 (6)	0.9916 (0.9833–0.9999)	1.0000 (0.6097–1.0000)	0.9761 (0.9699–0.9811)	0.0800 (0.0372–0.1637)	1.0000 (0.9986–1.0000)
6. Infection	2899 (2)	0.9377 (0.8814–0.9940)	1.0000 (0.3424–1.0000)	0.9092 (0.8982–0.9191)	0.0075 (0.0021–0.0271)	1.0000 (0.9985–1.0000)
7. Metaplasia	2900 (358)	0.9489 (0.9404–0.9574)	0.9050 (0.8702–0.9312)	0.9017 (0.8895–0.9126)	0.5645 (0.5236–0.6045)	0.9854 (0.9796–0.9895)
7.1. Intestinal metaplasia	2900 (355)	0.9481 (0.9395–0.9568)	0.9099 (0.8755–0.9354)	0.9006 (0.8884–0.9116)	0.5608 (0.5200–0.6008)	0.9862 (0.9806–0.9902)
Low priority						
8. Ectopia	2899 (2)	0.8583 (0.7330–0.9836)	1.0000 (0.3424–1.0000)	0.7946 (0.7795–0.8089)	0.0034 (0.0009–0.0121)	1.0000 (0.9983–1.0000)
9. Erosion	2899 (275)	0.7794 (0.7552–0.8037)	0.9018 (0.8609–0.9316)	0.4928 (0.4737–0.5119)	0.1571 (0.1399–0.1758)	0.9795 (0.9704–0.9859)
10. Fibrosis	2899 (1)	0.5580 (N/A)	1.0000 (0.2065–1.0000)	0.5580 (0.5398–0.5760)	0.0008 (0.0001–0.0044)	1.0000 (0.9976–1.0000)
11. Inflammation	2900 (1083)	0.8257 (0.8111–0.8402)	0.9003 (0.8810–0.9167)	0.5850 (0.5622–0.6075)	0.5639 (0.5404–0.5871)	0.9078 (0.8898–0.9230)
12. Necrosis	2899 (2)	0.9981 (0.9959–1.0000)	1.0000 (0.3424–1.0000)	0.9972 (0.9946–0.9986)	0.2000 (0.0567–0.5098)	1.0000 (0.9987–1.0000)

Class	N (positive)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
13. Polyp	2899 (954)	0.9758 (0.9708– 0.9809)	0.9015 (0.8809– 0.9188)	0.9440 (0.9328– 0.9533)	0.8875 (0.8661– 0.9059)	0.9513 (0.9408– 0.9600)
14. Ulcer	2899 (346)	0.9181 (0.9007– 0.9355)	0.9017 (0.8658– 0.9288)	0.7521 (0.7349– 0.7684)	0.3302 (0.3009– 0.3608)	0.9826 (0.9758– 0.9875)
15. Xanthoma	2899 (6)	0.8636 (0.6831– 1.0000)	1.0000 (0.6097– 1.0000)	0.4234 (0.4055– 0.4415)	0.0036 (0.0016– 0.0078)	1.0000 (0.9969– 1.0000)

To enable deployment as a clinical alert system, we calibrated binary operating points on the internal test set using a two-step optimization: first ensuring a minimum sensitivity of 90% for each class, then maximizing specificity at that sensitivity threshold. At the calibrated operating point for malignancy, the model achieved a sensitivity of 0.9138 (95% CI: 0.8136–0.9626), specificity of 0.9976 (95% CI: 0.9950–0.9988), positive predictive value (PPV) of 0.8833 (95% CI: 0.7782–0.9423), and negative predictive value (NPV) of 0.9983 (95% CI: 0.9960–0.9993). The complete set of calibrated metrics for all 29 classes is detailed in Table 1. These operating points were subsequently fixed and applied without modification to all external validation cohorts described below.

Annotation-free lesion localization across 29 categories

Beyond slide-level classification, the model also produces pixel-resolution localization heatmaps for each of the 29 categories without any pixel- or region-level annotations during training. Because every class is supervised only by the slide-level diagnostic label routinely captured in the LIS, extending localization to additional categories or new institutions requires no additional annotation effort—a key prerequisite for scaling AI-assisted pathology to diverse diagnostic settings.

Although localization maps are generated for all 29 classes, only four are surfaced to the pathologist during slide review: malignancy (red), dysplasia (yellow), metaplasia (green), and *H. pylori* (blue contours). This restriction reflects two design considerations: rendering 29 simultaneous overlays would impose substantial cognitive load, and the four selected categories together capture the apex of the diagnostic hierarchy—the only malignant or premalignant findings, plus the most clinically actionable infectious agent. *H. pylori* is rendered as a tissue-fragment outline rather than a heatmap because the underlying activation reflects bacterial-induced gastritis rather than the organisms themselves; outlining the entire affected fragment instead directs the pathologist to perform high-magnification review across the relevant tissue area.

Figure 2 illustrates the resulting localizations across a representative panel of cases. The model produces tightly aligned overlays for adenocarcinoma (Fig. 2a), and resolves subtle malignant foci in signet ring cell carcinoma (Fig. 2b)—a notoriously difficult subtype where scattered tumor cells are easily overlooked. Histologically distinct entities such as gastrointestinal stromal tumor (GIST, Fig. 2c) are also accurately localized despite morphologies that bear no resemblance to epithelial neoplasia, demonstrating that the model has learned features specific to each tumor class rather than a generic "abnormality" signal. The

multi-label formulation further enables simultaneous visualization of co-existing lesions: in a biopsy harboring malignancy, intestinal metaplasia, and *H. pylori* infection, the three pathological territories are correctly delineated within a single tissue fragment (Fig. 2d), illustrating that the system handles the diagnostic complexity routinely encountered in clinical practice. Localization examples for additional categories—diffuse large B-cell lymphoma, neuroendocrine tumor, low-grade dysplasia, and intestinal metaplasia—are provided in Supplementary Fig. 1, including a partially-localized neuroendocrine tumor case where the limited training data ($n = 15$ in the training set) constrains full segmentation, yet the model still highlights diagnostically useful regions. Localization across non-neoplastic categories (inflammation, polyp, and ulcer) is further illustrated in Supplementary Fig. 2.

Performance as a Post-Read Alert System

To evaluate the model's utility as a post-read quality assurance tool (Fig. 1b), we conducted a retrospective simulation across four independent hospital cohorts (Linkou, Keelung, Kaohsiung, and Tucheng CGMH). In this simulation, the existing LIS diagnoses served as a proxy for the pathologist's initial, unassisted interpretation. The model's malignancy predictions were then compared against these LIS labels using the pre-calibrated operating points established on the internal test set, and all discordant cases—where the model flagged malignancy but the LIS record indicated a benign diagnosis, or vice versa—were identified. A senior pathologist, blinded to the model output, subsequently reviewed each discordant case to render an adjudicated diagnosis. If this adjudication resulted in a revised diagnosis, it constituted direct evidence that the post-read alert would have corrected a clinically significant error in routine practice. Cases determined to be invalid (e.g., non-gastric biopsies) were excluded.

Initial evaluation against the original LIS labels (Table 2, 'Ground Truth = LIS') demonstrated high specificity across all sites but variable sensitivity: Linkou CGMH (sensitivity 0.8947; specificity 0.9920), Keelung CGMH (0.8000; 0.9867), Kaohsiung CGMH (0.7941; 0.9554), and Tucheng CGMH (1.0000; 0.9893). The adjudication process revealed that a substantial proportion of the model's apparent false positives were in fact true malignancies absent from the original LIS records. For example, in the Keelung CGMH cohort, six cases initially classified as false positives were reclassified as true positives upon expert review, increasing the confirmed malignant case count from 15 to 21. Similar corrections were observed at Linkou CGMH (5 false positives reclassified), Kaohsiung CGMH (3 reclassified), and Tucheng CGMH (3 reclassified). Conversely, expert adjudication also reclassified 5 cases originally labeled as malignant in the LIS but flagged as benign by the model—1 at Linkou CGMH and 4 at Kaohsiung CGMH—as true negatives upon review, indicating that the model also surfaced miscoded benign cases that had been incorrectly recorded as malignant in the LIS. These bidirectional corrections highlight the prevalence of LIS coding errors in routine workflows.

Table 2

External validation performance of the malignancy classification model on four independent, out-of-sample test cohorts. Performance is compared using two ground truth standards: the original LIS-based labels and an adjudicated set of labels. The adjudicated set was established after a senior pathologist reviewed all discordant (false positive and false negative) predictions, correcting misclassifications (detailed in footnotes a–d) and excluding invalid cases (e.g., non-gastric biopsies). This accounts for differences in the total (N) and positive case numbers. 95% CIs for AUCs were calculated using DeLong's method; CIs for all other metrics were calculated using Wilson's method.

Test set	Ground Truth	N (positive)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Linkou CGMH external test set	LIS	3020 (19)	0.9901 (0.9853–0.9949)	0.8947 (0.6861–0.9706)	0.9920 (0.9881–0.9946)	0.4146 (0.2776–0.5663)	0.9993 (0.9976–0.9998)
Linkou CGMH external test set	Adjudicated	3014 (23)	0.9988 (0.9945–1.0000)	0.9565 (0.7901–0.9923)	0.9957 (0.9926–0.9975)	0.6286 (0.4634–0.7683)	0.9997 (0.9981–0.9999)
Keelung CGMH external test set	LIS	690 (15)	0.9347 (0.9233–0.9461)	0.8000 (0.5481–0.9295)	0.9867 (0.9749–0.9930)	0.5714 (0.3655–0.7553)	0.9955 (0.9869–0.9985)
Keelung CGMH external test set	Adjudicated	690 (21)	0.9541 (0.9447–0.9635)	0.8571 (0.6536–0.9502)	0.9955 (0.9869–0.9985)	0.8571 (0.6536–0.9502)	0.9955 (0.9869–0.9985)
Kaohsiung CGMH external test set	LIS	684 (34)	0.9527 (0.9450–0.9603)	0.7941 (0.6320–0.8965)	0.9554 (0.9367–0.9688)	0.4821 (0.3567–0.6099)	0.9889 (0.9772–0.9946)
Kaohsiung CGMH external test set	Adjudicated	684 (33)	0.9814 (0.9738–0.9891)	0.9091 (0.7643–0.9686)	0.9601 (0.9421–0.9726)	0.5357 (0.4070–0.6598)	0.9952 (0.9861–0.9984)
Tucheng CGMH external test set	LIS	285 (4)	0.9973 (0.9677–1.0000)	1.0000 (0.5101–1.0000)	0.9893 (0.9691–0.9964)	0.5714 (0.2505–0.8418)	1.0000 (0.9864–1.0000)
Tucheng CGMH external test set	Adjudicated	285 (7)	1.0000 (0.9736–1.0000)	1.0000 (0.6457–1.0000)	1.0000 (0.9864–1.0000)	1.0000 (0.6457–1.0000)	1.0000 (0.9864–1.0000)
a Linkou CGMH: 5 false positives (FP) corrected to true positives (TP); 1 false negative (FN) corrected to true negative (TN). 6 invalid cases were excluded.							
b Keelung CGMH: 6 FPs corrected to TPs.							
c Kaohsiung CGMH: 3 FPs corrected to TPs; 4 FNs corrected to TNs.							
d Tucheng CGMH: 3 FPs corrected to TPs.							

When performance was re-evaluated against the adjudicated ground truth (Table 2, 'Ground Truth = Adjudicated'), metrics improved consistently: Linkou CGMH sensitivity rose to 0.9565 (specificity 0.9957), Keelung CGMH to 0.8571 (0.9955), Kaohsiung CGMH to 0.9091 (0.9601), and Tucheng CGMH achieved perfect sensitivity and specificity (1.0000; 1.0000). The systematic upward revision of both sensitivity and the number of confirmed positives across all four cohorts provides strong evidence that the model, deployed as a post-read alert system, can identify clinically significant malignancies that were missed or miscoded in routine diagnostic workflows.

Performance as a Concurrent Assistance System in a Simulated Clinical Workflow

To quantify the impact of a concurrent assistance system on diagnostic workflow (Fig. 1c), we conducted a clinical reader study. Seven participants, including four attending pathologists (C.-L.C., H.-C.C., J.L., and P.-H.C.) and three resident pathologists (C.-H.L., Y.-Y.H., and T.-C. Cheng), were recruited. One participant's session data was lost due to a technical issue during submission and was excluded from analysis. Participants reviewed a set of 120 whole slide images under two conditions: (1) unaided and (2) with AI assistance. The AI assistance provided heatmaps highlighting areas suspicious for malignancy, dysplasia, metaplasia, and *H. pylori*. A three-week washout period was implemented between the two reading sessions to minimize recall bias. The results of this experiment are summarized in Fig. 3.

The AI-assisted workflow demonstrated a slight but statistically significant improvement in efficiency, with the median review time per slide decreasing from 43.8 seconds to 41.1 seconds ($n = 720$, $P = .026$). More importantly, diagnostic sensitivity for malignancy was significantly enhanced. With AI assistance, pathologists' sensitivity for malignancy increased from 92.2% to 94.8% ($n = 306$, $P = .034$), while specificity remained statistically unchanged (97.3% vs. 97.6%, $n = 414$, $P = .667$).

For secondary findings (excluding malignant cases), AI assistance also significantly improved the sensitivity for metaplasia, which rose from 94.2% to 97.4% ($n = 342$, $P = .013$); the corresponding change in specificity was not statistically significant ($P = .09$). No significant differences in sensitivity or specificity were observed for the detection of dysplasia ($P = .687$ and $P = .78$, respectively) or *H. pylori* ($P = .382$ and $P = 1.0$), respectively. These results suggest the primary benefits of the AI-assisted workflow are a slight reduction in review time and an increased detection sensitivity for both malignant and metaplastic lesions.

Enhanced Performance and Data Efficiency via VFM Fine-tuning

To assess the value of initializing our deep learning workflow with a VFM, we compared the performance of our fine-tuned VFM against a model trained from scratch (w/o VFM) using identical architectures and datasets. The evaluation was conducted across two distinct regimes: a "many-shot" scenario using the full training dataset and a "few-shot" scenario using limited subsets of data.

In the "many-shot" experiment utilizing the complete training set of 10,462 WSIs, the VFM-fine-tuned model demonstrated statistically significant superiority over the model trained from scratch across the majority of

pathological classes (Table 3). For the primary Malignant class, the VFM approach achieved an AUC of 0.9945 compared to 0.9755 for the scratch model ($P < .001$). This performance gap was consistent across critical malignancy subtypes, including Carcinoma (0.9986 vs. 0.9777, $P < .001$), Adenocarcinoma (0.9985 vs. 0.9755, $P < .001$), and Sarcoma (0.9992 vs. 0.9295, $P < .001$). Notably, distinguishing difficult subtypes such as Signet ring cell carcinoma, where scattered tumor cells with subtle cytological atypia are easily overlooked on routine review, was significantly improved by the VFM, raising the AUC from 0.8241 to 0.9873 ($P = .003$). The magnitude of this gain suggests that VFM pre-training captures fine-grained cytological features that scratch training fails to extract from the same data.

Table 3

Comparison of classification performance between models with and without VFM pre-training. AUC values (95% CI) are reported for the internal test set ($n = 2,957$). The many-shot experiment uses the full 10,462-WSI training set; few-shot experiments use balanced subsets of 200, 400, and 800 slides. 95% CI were calculated using DeLong's method and P values were calculated using DeLong's test. CI, confidence interval; VFM, Vision Foundation Model.

Many-shot (the 10,462-WSI training set; 488 malignant, 9974 benign)				
Class	n (positive)	w/ VFM	w/o VFM	P
1. Malignant	2957 (58)	0.9945 (0.9878–1.0000)	0.9755 (0.9507–1.0000)	< .001
1.1. Carcinoma	2957 (44)	0.9986 (0.9964–1.0000)	0.9777 (0.9465–1.0000)	< .001
1.1.1. Adenocarcinoma	2956 (41)	0.9985 (0.9961–1.0000)	0.9755 (0.9428–1.0000)	< .001
1.1.1.1. Signet ring cell carcinoma	2919 (4)	0.9873 (0.9626–1.0000)	0.8241 (0.4800–1.0000)	.003
1.1.2. Squamous cell carcinoma	2956 (2)	0.9998 (0.9994–1.0000)	0.9292 (0.8367–1.0000)	.138
1.1.3. Lymphoma	2957 (5)	0.9999 (0.9997–1.0000)	0.9988 (0.9972–1.0000)	.156
1.1.3.1. Diffuse large B-cell lymphoma	2956 (3)	0.9992 (0.9980–1.0000)	0.9900 (0.9718–1.0000)	.037
1.1.3.2. Marginal zone B-cell lymphoma	2956 (1)	0.9980 (N/A)	0.9956 (N/A)	N/A
1.1.4. Neuroendocrine tumor	2957 (3)	0.9882 (0.9665–1.0000)	0.9227 (0.7877–1.0000)	.078
1.1.5. Sarcoma	2957 (6)	0.9992 (0.9977–1.0000)	0.9295 (0.7968–1.0000)	< .001
1.1.5.1. Gastrointestinal stromal tumor	2957 (5)	0.9999 (0.9995–1.0000)	0.9969 (0.9930–1.0000)	.001
2. Atypia	2900 (3)	0.7554 (0.5125–0.9982)	0.7226 (0.4472–0.9980)	< .001
3. <i>H. pylori</i>	2899 (32)	0.9210 (0.8940–0.9480)	0.9157 (0.8906–0.9408)	< .001
4. Benign tumor	2900 (6)	1.0000 (1.0000–1.0000)	0.9995 (0.9986–1.0000)	.458
5. Dysplasia	2899 (26)	0.9707 (0.9494–0.9920)	0.9590 (0.9289–0.9890)	< .001
5.1. Adenoma	2893 (2)	0.9983 (0.9968–0.9998)	0.9983 (0.9967–0.9999)	1.000
5.2. Mild dysplasia	2893 (11)	0.9571 (0.9281–0.9860)	0.9480 (0.9054–0.9907)	< .001

Many-shot (the 10,462-WSI training set; 488 malignant, 9974 benign)				
Class	<i>n</i> (positive)	w/ VFM	w/o VFM	<i>P</i>
5.3. Severe dysplasia	2893 (6)	0.9916 (0.9833–0.9999)	0.9368 (0.8464–1.0000)	< .001
6. Infection	2899 (2)	0.9377 (0.8814–0.9940)	0.8290 (0.6071–1.0000)	.199
7. Metaplasia	2900 (358)	0.9489 (0.9404–0.9574)	0.9514 (0.9437–0.9591)	1.000
7.1. Intestinal metaplasia	2900 (355)	0.9481 (0.9395–0.9568)	0.9511 (0.9434–0.9588)	1.000
8. Ectopia	2899 (2)	0.8583 (0.7330–0.9836)	0.7755 (0.3637–1.0000)	.571
9. Erosion	2899 (275)	0.7794 (0.7552–0.8037)	0.7718 (0.7468–0.7969)	1.000
10. Fibrosis	2899 (1)	0.5580 (N/A)	0.6480 (N/A)	N/A
11. Inflammation	2900 (1083)	0.8257 (0.8111–0.8402)	0.8188 (0.8039–0.8337)	1.000
12. Necrosis	2899 (2)	0.9981 (0.9959–1.0000)	0.9986 (0.9973–1.0000)	.724
13. Polyp	2899 (954)	0.9758 (0.9708–0.9809)	0.9731 (0.9676–0.9786)	1.000
14. Ulcer	2899 (346)	0.9181 (0.9007–0.9355)	0.9045 (0.8866–0.9224)	1.000
15. Xanthoma	2899 (6)	0.8636 (0.6831–1.0000)	0.8033 (0.5639–1.0000)	< .001
Few-shot (a 200-WSI balanced subset; 100 malignant, 100 benign)				
Class	<i>n</i> (positive)	w/ VFM	w/o VFM	<i>P</i>
Malignant	2957 (58)	0.9628 (0.9442–0.9814)	0.9360 (0.9158–0.9562)	< .001
Few-shot (a 400-WSI balanced subset; 200 malignant, 200 benign)				
Class	<i>n</i> (positive)	w/ VFM	w/o VFM	<i>P</i>
Malignant	2957 (58)	0.9906 (0.9840–0.9972)	0.9660 (0.9489–0.9831)	< .001
Few-shot (a 800-WSI balanced subset; 400 malignant, 400 benign)				
Class	<i>n</i> (positive)	w/ VFM	w/o VFM	<i>P</i>
Malignant	2957 (58)	0.9935 (0.9883–0.9987)	0.9157 (0.8792–0.9522)	< .001

The benefits extended beyond malignancies to pre-cancerous and benign conditions. The VFM achieved significantly higher AUCs for Dysplasia (0.9707 vs. 0.9590, $P < .001$), Mild dysplasia (0.9571 vs. 0.9480, $P < .001$), and Severe dysplasia (0.9916 vs. 0.9368, $P < .001$). It also showed superior performance in identifying *H. pylori* (0.9210 vs. 0.9157, $P < .001$) and Xanthoma (0.8636 vs. 0.8033, $P < .001$), underscoring the foundation model's robust feature extraction capabilities across a diverse range of gastric pathologies.

The advantages of the VFM were most pronounced in data-constrained environments. We conducted "few-shot" experiments using balanced subsets of malignant and benign slides ($n = 100, 200, \text{ and } 400$ per group). In every subset, the VFM significantly outperformed the model trained from scratch ($P < .001$ for all comparisons).

Most remarkably, the VFM demonstrated an extreme reduction in data requirements. A VFM fine-tuned on a mere 400 slides (200 malignant, 200 benign) achieved a Malignant classification AUC of 0.9906. This performance not only surpassed the scratch model trained on the same subset (AUC 0.9660), but surprisingly, it also outperformed the model trained from scratch on the entire 10,462-slide dataset (AUC 0.9755). This result indicates that leveraging a foundation model can reduce the requirement for annotated training data by orders of magnitude while maintaining, or even exceeding, the diagnostic accuracy of traditional large-scale training paradigms.

Discussion

This study presents a deep learning system for gastric biopsy classification spanning 29 pathological categories and evaluates its clinical utility through two complementary workflow integration modes—a post-read alert system and a concurrent assistance system—across multiple independent hospital cohorts. Three principal findings emerge: the model achieves high discriminatory performance across a clinically comprehensive diagnostic spectrum; the post-read alert system identifies malignancies missed in routine practice; and concurrent AI assistance significantly improves pathologists' sensitivity for malignancy and metaplasia.

A defining feature of our model is its 29-class classification scope, which substantially exceeds the two- to three-category frameworks employed in all prior gastric biopsy AI studies [6–10]. This breadth reflects the clinical reality of gastric biopsy interpretation, where pathologists navigate a continuous pathological spectrum rather than rendering a binary cancer-versus-non-cancer judgment, simultaneously evaluating each specimen for malignancy across multiple histological subtypes, precancerous findings along the Correa cascade (chronic gastritis, intestinal metaplasia, and varying grades of dysplasia), treatable infections (*H. pylori*), and a range of benign conditions. Existing AI systems limited to binary or three-tier classification [9, 10] address only the apex of this diagnostic hierarchy and provide no assistance for the precancerous and inflammatory findings that constitute the majority of clinical workload and are essential for surveillance decisions under current guidelines [3]. Our model is the first to simultaneously flag malignancy, grade dysplasia, detect metaplasia, and identify *H. pylori* within a single inference pass, mirroring the full scope of pathologist decision-making. Notably, this comprehensive coverage was achieved on the highest-priority diagnostic target: AUCs exceeded 0.9707 for dysplasia, 0.9489 for metaplasia, and 0.9210 for *H. pylori*, while malignancy detection reached 0.9945.

To our knowledge, this is the first study in gastric biopsy pathology to prospectively evaluate two distinct AI workflow paradigms within a single framework, and the most clinically comprehensive validation reported to date. Concurrent AI assistance yielded significant gains in malignancy sensitivity (92.2% to 94.8%, $P = .034$) and metaplasia sensitivity (94.2% to 97.4%, $P = .013$), with a modest 6% reduction in median review time. The sensitivity gains are consistent in direction with Ba et al. [10], but our study extends prior work in three important ways. First, our reader study employed a rigorous two-round crossover design with a three-week washout period to control for recall bias and order effects, whereas most previous evaluations relied on single-session designs. Second, our concurrent assistance simultaneously supports four diagnostic targets rather than the binary or three-tier evaluations used by Park et al. [9] and Ba et al. [10], more closely mirroring the cognitive demands of real-world sign-out. The deliberate restriction of heatmap display to four high-yield categories—rather than all 29—reflects a key design principle: AI overlays should direct pathologist attention to the most consequential findings without overwhelming the visual field, preserving the pathologist's diagnostic agency over the full hierarchy of findings. The more modest review-time reduction compared with Park et al.'s 47% savings [9] should be interpreted in this context: *H. pylori* detection alone requires careful scanning for small, often inconspicuous organisms scattered across multiple tissue fragments and is inherently more time-consuming than binary malignancy assessment. Third, no prior gastric biopsy AI study has empirically evaluated more than one workflow integration mode, leaving open the question of which paradigm best suits which clinical setting—a gap our work directly addresses. One observation that warrants attention is a downward trend in metaplasia specificity under AI assistance ($P = .09$); although this did not reach statistical significance and was substantially outweighed by the corresponding sensitivity gain (94.2% to 97.4%), it suggests that heatmap overlays may, in some borderline cases, encourage overcalling. This trade-off—favoring sensitivity for findings central to surveillance decisions—is consistent with our calibration philosophy of prioritizing sensitivity for screening-relevant categories, but warrants monitoring in prospective deployments.

The post-read alert system addresses a fundamentally different need, functioning as a retrospective quality assurance mechanism that compares AI predictions against finalized LIS diagnoses and flags discordant cases. Across four independent hospital cohorts ($n = 4,679$), 17 cases initially classified as model false positives were confirmed upon expert adjudication to be true malignancies absent from the original LIS records. These represent clinically significant missed or miscoded diagnoses that a deployed post-read alert would have intercepted in real time. While Abe et al. [23] previously proposed a discordance-based review concept, our work provides the first direct empirical evidence of diagnostic corrections at scale, validated across four geographically distinct institutions. This finding also carries a methodological implication: algorithm evaluation against imperfect ground truth systematically underestimates true clinical utility, and the magnitude of this underestimation can be substantial.

These two modes present complementary trade-offs suited to different deployment scenarios. Concurrent assistance requires integration into a digital pathology viewer but offers real-time gains most valuable in high-throughput screening settings. The post-read alert system operates independently of the primary workflow and can run as a batch process on already-reported cases, making it implementable even without real-time digital infrastructure. Studies on prostate biopsy AI have reported analogous trade-offs across deployment modes [12, 24], and clinical validation studies have similarly demonstrated AI-augmented

diagnostic gains in prostate pathology [25]; our results extend these observations to gastric pathology and demonstrate a unique advantage of the post-read paradigm—the ability to uncover latent diagnostic errors in historical case archives, a capability not previously demonstrated for gastric biopsy. Beyond the two paradigms evaluated here, further efficiency gains may be achievable through deeper workflow integration. A future direction is to position the validated VFM as a callable, agent-ready tool that an LLM-driven orchestrator can invoke to populate structured pathology reports—integrating risk stratification, heatmap-derived lesion characteristics, and fine-grained subtype queries into a single standardized output (Fig. 4). Such an agentic configuration [26] would extend the data-efficient VFM foundation demonstrated here from a passive classifier into an active reporting collaborator.

A second key contribution is the systematic quantification of data efficiency conferred by VFM pre-training in gastric biopsy classification—an analysis not previously reported for this organ system. A VFM fine-tuned on merely 400 slides achieved a malignancy AUC of 0.9906, surpassing a model trained from scratch on the full 10,462-slide dataset (AUC 0.9755). This represents a roughly 26-fold reduction in annotated data requirement while exceeding the performance of traditional large-scale training, with profound practical implications. Institutions wishing to deploy clinical-grade AI for rare diseases, organ systems with limited annotated cohorts, or regional histological variants can now do so with a fraction of the curation effort previously assumed necessary. Although recent reports have demonstrated VFM data efficiency in other contexts [17, 18], our work is the first to extend this finding to multi-class gastric biopsy classification and to directly compare VFM and from-scratch training under matched architectures and protocols. Looking forward, this data efficiency lays the foundation for agentic AI systems in pathology [26], in which large language model-driven agents autonomously orchestrate VFM-based tools—selecting, fine-tuning, and composing classifiers on demand for novel diagnostic tasks—without the prohibitive data curation cycles that have historically constrained clinical AI deployment. Such agentic workflows could enable rapid adaptation to emerging entities, rare diseases, and institution-specific histological patterns, transforming AI from a collection of static, single-purpose models into a flexible diagnostic collaborator.

A third contribution is the demonstration that pixel-resolution lesion localization across 29 categories can be obtained without any pixel-level supervision. By combining streaming convolution with ReLSE pooling, the model produces CAM for every class directly from slide-level labels, eliminating the need for the dense annotations that have historically constrained multi-class WSI models. This annotation-free localization is what makes the concurrent assistance heatmaps practical to scale: extending the system to additional categories or deploying it at new institutions requires only slide-level diagnostic labels—routinely available from LIS/SNOMED records—rather than the costly pixel-level annotations that pathologists are unable to provide at scale. Combined with the data efficiency of VFM pre-training, this annotation-free localization capability lowers two of the principal barriers to clinical AI deployment in pathology.

Several limitations warrant discussion. Performance on rare entities such as neuroendocrine tumor was constrained by limited training data, and these classes would benefit from data augmentation strategies or targeted curation in future versions. The post-read alert evaluation was a retrospective simulation; a prospective deployment study would be needed to confirm the observed benefits in routine workflow. Although external validation across four hospital cohorts demonstrated robust generalizability within the

Chang Gung Healthcare System, validation on geographically and ethnically more diverse populations remains necessary. Finally, the reader study involved six effective participants; while the crossover design enhances statistical power per reader, larger multi-site reader studies would further strengthen the generalizability of our findings.

In conclusion, we present a foundation-model–powered AI system that delivers comprehensive 29-class gastric biopsy classification with annotation-free localization, evaluated under two complementary clinical integration paradigms—concurrent assistance for real-time sensitivity gains and post-read alerts for retrospective error recovery. By combining clinical impact with extreme data efficiency, our framework provides a practical blueprint for scalable AI deployment across diverse pathology practices.

Methods

Datasets

Training and internal test sets. Gastric biopsy whole slide images (WSIs) were retrospectively collected from Linkou Chang Gung Memorial Hospital (CGMH) between 2020 and 2023. Inclusion criteria were: (1) formalin-fixed, paraffin-embedded (FFPE) tissue sections stained with hematoxylin and eosin (H&E); (2) gastric origin confirmed by specimen label; and (3) one representative slide selected per case to simplify the analytic unit. The total collection comprised 13,419 WSIs, which were randomly split into a training set of 10,462 WSIs and an independent internal test set of 2,957 WSIs.

Diagnostic labels were derived from SNOMED (Systematized Nomenclature of Medicine) codes recorded in the institutional Laboratory Information System (LIS) at the time of original clinical sign-out. The initial extraction yielded 99 unique SNOMED morphology codes. These were consolidated into 29 clinically meaningful categories through a two-step process: first, closely related codes were grouped under a common parent class (e.g., multiple adenocarcinoma variants mapped to "Adenocarcinoma"); second, categories with fewer than 10 positive cases in the training set were excluded to ensure sufficient representation for model training. The resulting 29-class hierarchy spans three priority tiers: high-priority classes (malignancy and its subtypes), medium-priority classes (atypia, *H. pylori*, dysplasia, metaplasia, benign tumors, infection), and low-priority classes (ectopia, erosion, fibrosis, inflammation, necrosis, polyp, ulcer, xanthoma). A multi-label formulation was adopted, as a single biopsy may carry more than one diagnosis (e.g., adenocarcinoma coexisting with intestinal metaplasia and *H. pylori*).

External validation cohorts for the post-read alert system. To evaluate model generalizability, four independent cohorts were assembled from hospitals within the Chang Gung Healthcare System but outside the training institution: Linkou CGMH external set ($n = 3,020$; 2024), Keelung CGMH ($n = 690$; 2024), Kaohsiung CGMH ($n = 684$; 2024), and Tucheng CGMH ($n = 285$; 2024). All cohorts followed the same inclusion criteria as the training set.

Concurrent assistance reader study set. A separate set of 120 WSIs was curated from the Linkou CGMH 2024 cohort (i.e., the external validation set described above) for use in the reader study. Slides were eligible for inclusion if the AI model predicted at least one of the following four categories as positive: malignancy,

dysplasia, metaplasia, or *H. pylori*. From this eligible pool, 120 WSIs were randomly selected. Ground truth labels for the four target categories were established by a senior gastrointestinal pathologist (S.-C.H.) who reviewed all 120 WSIs in a blinded fashion. Of the 120 WSIs, 51 were positive for malignancy. The malignancy sensitivity and specificity analyses used these 51 malignant cases and 69 non-malignant cases, respectively. To enable evaluation of secondary findings independently of malignancy, the dysplasia, metaplasia, and *H. pylori* analyses were restricted to the 69 non-malignant cases; among these, 11 were positive for dysplasia, 57 for metaplasia, and 16 for *H. pylori*. Because of the multi-label formulation, individual cases could be positive for more than one secondary category simultaneously.

Image Acquisition and Digitization. All H&E-stained formalin-fixed paraffin-embedded (FFPE) tissue sections across the training and all external validation cohorts were digitized using the same model of high-throughput whole-slide scanner (NanoZoomer S360, Hamamatsu Photonics, Japan). Slides were scanned at 40× magnification with a consistent resolution of 0.23 μm/pixel.

Vision Foundation Model pre-training data. The VFM backbone was pre-trained on a large-scale, multi-source dataset of 70,618 H&E-stained WSIs aggregated from 11 publicly available repositories spanning diverse organ systems, disease types, and staining protocols. The largest contributors were The Cancer Genome Atlas (TCGA; 29,863 WSIs across 33 cancer types) [35], the Genotype-Tissue Expression project (GTEx; 25,430 WSIs across 54 non-diseased tissue sites) [36], and the Clinical Proteomic Tumor Analysis Consortium (CPTAC; 7,505 WSIs) [37]. Additional sources included the Cancer Genome Characterization Initiative (CGCI; 2,823 WSIs) [38], the NADT-Prostate cohort (Nascent prostate cancer heterogeneity following neoadjuvant androgen deprivation therapy; 1,392 WSIs) [39], the National Lung Screening Trial (NLST; 1,225 WSIs) [40], the Camelyon17 challenge (1,000 WSIs) [41], the Cancer Moonshot Biobank (CMB; 590 WSIs) [42], the Proteogenomic Translational Research Centers High-Grade Serous Ovarian Carcinoma cohort (PTRC-HGSOC; 348 WSIs) [43], the Human Cancer Models Initiative (HCMI; 312 WSIs) [44], and a sentinel lymph node breast cancer dataset (SLN-Breast; 130 WSIs) [45].

Vision Foundation Model pre-training

A ResNet-50 [19] encoder was pre-trained on image tiles extracted from the 70,618 WSIs described above using Barlow Twins [20], a self-supervised contrastive learning framework that maximizes agreement between augmented views of the same tile while minimizing redundancy among embedding dimensions. The encoder was initialized from publicly available DINO [27] self-supervised weights pre-trained on ImageNet, rather than from random initialization, to accelerate convergence.

Pre-training was implemented using the MMPretrain framework [28]. Tiles of 512 × 512 pixels were randomly cropped from foreground tissue regions of each WSI, with a saturation-based validity check to exclude background and low-content patches. Tiles were sampled uniformly across available magnification levels to expose the model to both architectural and cytological features. For each sampled tile, two augmented views of 224 × 224 pixels were generated through random resized cropping, random horizontal and vertical flips, 90° rotation, HED color perturbation [29], color jitter (brightness, contrast, saturation, and hue), random grayscale conversion, and Gaussian blur—following standard multi-view contrastive learning protocols. Input images were normalized using ImageNet mean and standard deviation values.

Training was conducted in FP32 precision with a total batch size of 1,536 (384 per node $\times$ 4 NVIDIA H100 GPUs) using the LARS optimizer [30] with a base learning rate of 1.2 (linearly scaled as $0.2/256 \times 1,536$), momentum of 0.9, and weight decay of 1×10^{-6} . Batch normalization and bias parameters were excluded from weight decay and LARS adaptation. The learning rate schedule comprised a linear warmup over the first 34,000 iterations followed by cosine annealing over the remaining iterations, for a total of 680,000 training iterations. The resulting encoder serves as a general-purpose histopathological feature extractor.

Gastric Biopsy Classification Model Training

The pre-trained ResNet-50 encoder was subsequently fine-tuned end-to-end for the 29-class gastric biopsy classification task using the Enhanced Streaming CNN (ESCNN) method [16], which enables whole-slide-level training without tile-level annotations by streaming the entire WSI through the network in a memory-efficient manner. The global average pooling layer of the ResNet-50 was replaced with a Rectified Log-Sum-Exp (ReLSE) pooling layer [21], which provides a smooth, differentiable approximation to max pooling that is better suited for weakly supervised localization of focal lesions within large tissue areas. The classification head comprised a fully connected layer with 29 sigmoid outputs, one per class, trained with multi-label binary cross-entropy loss. All other hyperparameters (learning rate, optimizer, augmentation strategy) followed the defaults specified in the original ESCNN publication [16]. WSIs were processed at 10 $\times$ magnification. Training was performed on 4 NVIDIA H100 GPUs with early stopping based on validation loss.

For each of the 29 classes, a binary decision threshold was calibrated on the internal test set via a two-step optimization: (1) restricting candidate thresholds to those yielding sensitivity ≥ 0.90 , and (2) selecting the threshold that maximized specificity within this constrained set. The 29 calibrated thresholds were fixed prior to external validation and reader study evaluation, and applied without modification to all downstream cohorts.

To quantify the data efficiency conferred by VFM pre-training, we trained models on balanced subsets of the training data containing 100, 200, and 400 slides per group (malignant vs. benign), for a total of 200, 400, and 800 slides respectively. For each subset size, models were trained both with VFM initialization and from ImageNet initialization, using identical architectures and hyperparameters. Performance was evaluated on the full internal test set of 2,957 WSIs.

Lesion Localization

During inference, pixel-resolution localization heatmaps were generated for each of the 29 classes by applying CAM [22] to the final convolutional feature maps prior to ReLSE pooling. The combination of streaming convolution—which preserves spatial resolution across the gigapixel WSI—and ReLSE pooling—which provides a smooth approximation to max pooling and concentrates gradient flow on discriminative regions—yields CAMs directly comparable to expert-annotated lesion regions, despite training using only slide-level multi-label supervision.

Although CAMs were generated for all 29 categories, only four parent categories were surfaced to the pathologist during slide review: malignancy (red), dysplasia (yellow), metaplasia (green), and *H. pylori* (blue contours; see below). For each parent category, a pixel-wise score was computed as the maximum across

all of its child subclasses—for example, the malignancy score at each pixel was the maximum of the scores for malignancy, carcinoma, adenocarcinoma, signet ring cell carcinoma, lymphoma, neuroendocrine tumor, sarcoma, and their subclasses. At each pixel, the four parent-category scores were compared against a fixed threshold of 0.55; pixels exceeding the threshold for at least one category were rendered in the color of the highest-scoring category, while pixels not exceeding the threshold for any category were left unannotated. This winner-takes-all rendering ensured that each pixel carried at most a single color overlay even when the underlying tissue was simultaneously suspicious for multiple findings, preserving visual clarity at the cost of suppressing lower-confidence co-existing predictions.

Direct heatmap overlay for *H. pylori* highlights regions of bacterial-induced chronic gastritis rather than the organisms themselves and provides limited assistance for the pathologist's task of locating bacteria at high magnification. We therefore introduced a tissue-fragment-level aggregation step. Tissue contours were delineated on the WSI thumbnail using the morphological Chan–Vese active contour algorithm [31], initialized from a foreground mask obtained by Otsu thresholding on the saturation channel of the HSV-converted image. For each detected contour, the maximum *H. pylori* heatmap activation within the contour was computed; contours exceeding the calibrated threshold were rendered as blue outlines in the digital viewer, directing the pathologist to perform high-magnification review of the entire flagged fragment.

Post-read Alert System Evaluation

The post-read alert system was evaluated through a retrospective simulation. For each of the four external validation cohorts, the trained model generated malignancy predictions for all WSIs using the operating points calibrated on the internal test set (see Results). These predictions were compared against the original LIS diagnoses. All discordant cases—defined as instances where the model prediction and LIS label disagreed on the presence or absence of malignancy—were compiled into a review list. A senior gastrointestinal pathologist (S.-C.H.), blinded to the model output, independently reviewed each discordant WSI and rendered an adjudicated diagnosis. Cases identified as invalid specimens (e.g., non-gastric tissue inadvertently included in the gastric biopsy cohort) were excluded from analysis. Performance metrics were computed against both the original LIS labels and the adjudicated ground truth.

Concurrent Assistance Reader Study

Seven pathologists from the Chang Gung Healthcare System participated in the reader study: four attending pathologists (C.-L.C., H.-C.C., J.L., and P.-H.C.) and three resident pathologists (C.-H.L., Y.-Y.H., and T.-C. Cheng). All participants had experience with digital slide review. One participant was excluded from the final analysis because the diagnostic labels were not successfully recorded in the study database. This was due to a technical oversight where the participant completed the review but failed to trigger the final submission command for the session, resulting in a non-recoverable loss of the session data. Thus, six readers contributed to the final analysis.

The 120 WSIs were divided into four slide groups of 30 WSIs each, stratified to ensure a comparable distribution of malignant, dysplastic, metaplastic, *H. pylori*-positive, and benign cases across groups. Pathologists were assigned to one of two pathologist groups (group α : three pathologists; group β : four pathologists). The study employed a two-round crossover design with a three-week washout period,

ensuring that every pathologist evaluated every slide under both conditions (unaided and AI-assisted) while controlling for order and carryover effects. In the first round, pathologist group α reviewed slide groups 1 and 4 with AI assistance and slide groups 2 and 3 without; pathologist group β received the opposite assignment (slide groups 2 and 3 with AI, slide groups 1 and 4 without). In the second round, conducted after the washout period, the AI-condition assignments were reversed: pathologist group α reviewed slide groups 2 and 3 with AI assistance and slide groups 1 and 4 without, while pathologist group β reviewed slide groups 1 and 4 with AI and slide groups 2 and 3 without. This crossover ensured that each slide was evaluated by each pathologist under both the unaided and AI-assisted conditions exactly once, and that the allocation of AI assistance was balanced across slide groups and reading sessions. The excluded participant belonged to group α (leaving group α : two, group β : four). This does not affect the analysis: every slide was still read by every effective reader under both conditions, each phase retained a balanced AI/non-AI split, and the mixed-effects models (with pathologist and slide as crossed random intercepts) explicitly absorbed between-reader variability.

In the AI-assisted condition, heatmaps highlighting regions suspicious for malignancy (red), dysplasia (yellow), and metaplasia (green) were overlaid on the WSI in the digital viewer (aetherSlide, aetherAI), with blue contours indicating regions suspicious for *H. pylori*. The AI overlay could be toggled on and off at the pathologist's discretion. For each WSI, readers recorded binary diagnostic judgments for each of the four categories (malignancy, dysplasia, metaplasia, and *H. pylori*); review time per slide was automatically logged by the platform. Slides within each group were reviewed sequentially, and pathologists were instructed to complete two slide groups per session (groups 1–2 and groups 3–4) to minimize intra-day variability.

Statistical Analysis

Model performance. Receiver operating characteristic (ROC) analysis was performed using the pROC package [32] in R. Area under the ROC curve (AUC) values and their 95% confidence intervals (CIs) were calculated using DeLong's method [33]. Binary classification metrics (sensitivity, specificity, positive predictive value, negative predictive value) and their 95% CIs were computed using Wilson's score interval. For comparisons between VFM-initialized and scratch-trained models, statistical significance of AUC differences was assessed using DeLong's test for multiple comparisons across the 29 classes.

Reader study. The effect of AI assistance on review time was assessed using a linear mixed-effects model with AI condition (unaided vs. assisted) as a fixed effect and pathologist and slide as crossed random intercepts. Review times were log-transformed to approximate normality. The effect of AI assistance on diagnostic sensitivity and specificity was assessed using generalized linear mixed-effects models (GLMMs) with a binomial family and logit link function, with AI condition as a fixed effect and pathologist and slide as crossed random intercepts. Statistical significance was determined by likelihood ratio tests comparing models with and without the AI condition term. All reader study analyses were conducted in R using the lme4 package [34]. A two-sided P value of $< .05$ was considered statistically significant. No adjustment for multiple comparisons was applied to the reader study analyses, as the primary endpoint (malignancy sensitivity) was pre-specified and secondary endpoints were regarded as exploratory.

Declarations

The study was approved by the Institutional Review Board of Chang Gung Memorial Hospital (IRB No. 202400780B0). Given the retrospective use of de-identified archival specimens and data, the requirement for informed consent was waived. For the prospective reader study, informed consent was obtained from the participating pathologists.

Data and Code Availability

Data availability

The gastric biopsy whole-slide images and associated clinical metadata used in this study are not publicly available due to institutional data privacy and security policies of the Chang Gung Memorial Hospital system. Access to de-identified data for non-commercial research purposes may be granted upon reasonable request to the corresponding authors, subject to approval by the Institutional Review Board and a formal data transfer agreement. Source data for the main figures and tables are provided with this paper.

Code availability

The source code for model training and the pre-trained weights of the Vision Foundation Model (VFM) used in this study are open-source and can be accessed at the following GitHub repository:

<https://github.com/aetherAI/hms2>.

Acknowledgments

This work was supported by the National Science and Technology Council, Taiwan (grant nos. 109-2320-B-182A-011 and 114-2320-B-182A-030) and the Ministry of Health and Welfare, Taiwan (grant no. MOHW114-TDU-B-222-1440). The results published here are in whole or part based upon data generated by the TCGA Research Network:<https://www.cancer.gov/tcga>. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the U.S. National Institutes of Health, and by the National Cancer Institute, National Human Genome Research Institute, National Heart, Lung, and Blood Institute, National Institute on Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke; we gratefully acknowledge the donors and their families for the tissue donations that made GTEx possible. Data used in this publication were also generated by the National Cancer Institute Cancer Moonshot Biobank.

Author Contribution

S.-C.H. and C.-C.C. conceptualized and designed the study. S.-C.H. and T.-Y.C. performed data collection. S.-C.H. established the ground truth through expert slide review, and conducted qualitative analysis. C.-C.C. developed the AI methodology, performed model training and quantitative data analysis, and drafted the manuscript. C.-L.C., H.-C.C., J.L., P.-H.C., C.-H.L., Y.-Y.H., and T.-C. Cheng participated in the reader study and provided diagnostic interpretations. C.-Y.Y. and T.-C. Chen provided administrative support, supervised the research, and critically revised the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

Competing Interests

C.-Y.Y. is the Founder, Chairman, and Chief Executive Officer of aetherAI. C.-C.C. is the Director of Product and Machine Learning at aetherAI. The remaining authors declare no competing interests.

References

1. Thrift, A. P., Wenker, T. N. & El-Serag, H. B. Global burden of gastric cancer: epidemiological trends, risk factors, screening and prevention. *Nat. Rev. Clin. Oncol.* **20**, 338–349 (2023).
2. Correa, P. & Piazuelo, M. B. The gastric precancerous cascade. *J. Dig. Dis.* **13**, 2–9 (2012).
3. Pimentel-Nunes, P. et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II). *Endoscopy* **51**, 365–388 (2019).
4. Mabe, K., Inoue, K., Kamada, T., Kato, K. & Kato, M. Endoscopic screening for gastric cancer in Japan: current status and future perspectives. *Dig. Endosc.* **34**, 412–419 (2022).
5. Yilmaz, F. et al. Advancing artificial intelligence integration into the pathology workflow: exploring opportunities in gastrointestinal tract biopsies. *Lab. Invest.* **104**, 102043 (2024).
6. Yoshida, H. et al. Automated histological classification of whole-slide images of gastric biopsy specimens. *Gastric Cancer* **21**, 249–257 (2018).
7. Iizuka, O. et al. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci. Rep.* **10**, 1504 (2020).
8. Song, Z. et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat. Commun.* **11**, 4294 (2020).
9. Park, J. et al. A prospective validation and observer performance study of a deep learning algorithm for pathologic diagnosis of gastric tumors in endoscopic biopsies. *Clin. Cancer Res.* **27**, 719–728 (2021).
10. Ba, W. et al. Assessment of deep learning assistance for the pathological diagnosis of gastric cancer. *Mod. Pathol.* **35**, 1262–1268 (2022).
11. Bilal, M. et al. Development and validation of artificial intelligence-based prescreening of large-bowel biopsies taken in the UK and Portugal: a retrospective cohort study. *Lancet Digit. Health* **5**, e786–e797 (2023).
12. Horton, M. R., Parke, A., Gulturk, E. et al. The reading paradigm: how the sequence and presentation of AI results to pathologists influences endpoints and outcomes. Poster presented at the *European Congress on Digital Pathology*, Budapest, Hungary (2023).
13. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
14. Pinckaers, H., van Ginneken, B. & Litjens, G. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1581–1590 (2022).
15. Chen, C.-L., Chen, C.-C., Yu, W.-H. et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat. Commun.* **12**, 1193 (2021).
16. Huang, S.-C., Chen, C.-C., Lan, J. et al. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. *Nat. Commun.* **13**, 3347 (2022).

17. Lu, M. Y. et al. A visual–language foundation model for computational pathology. *Nat. Med.***30**, 863–874 (2024).
18. Roth, B., Koch, V., Wagner, S. J., Schnabel, J. A., Marr, C. & Peng, T. Low-resource finetuning of foundation models beats state-of-the-art in histopathology. In *2024 IEEE 21st International Symposium on Biomedical Imaging (ISBI)* 1–5 (IEEE, 2024).
19. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Pattern Recog.* 770–778 (IEEE, 2016).
20. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow Twins: self-supervised learning via redundancy reduction. In *Proc. Int. Conf. Mach. Learn.* 12310–12320 (PMLR, 2021).
21. Chen, C.-C., Yeh, Y.-C., Lin, M. M. & Yeh, C.-Y. Pixel-precise lesion localization in WSIs via weakly supervised streaming convolution with ReLSE and adaptive self-training. *bioRxiv* 2024.10.03.616445 (2025).
22. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. *Proc. IEEE Conf. Comp. Vis. Pattern Recog.* 2921–2929 (2016).
23. Abe, H. et al. Development and multi-institutional validation of an artificial intelligence-based diagnostic system for gastric biopsy. *Cancer Sci.***113**, 3608–3617 (2022).
24. Perincheri, S. et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod. Pathol.***34**, 1588–1595 (2021).
25. Raciti, P. et al. Clinical validation of artificial intelligence–augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. *Arch. Pathol. Lab. Med.***147**, 1178–1185 (2023).
26. Qiu, J. et al. LLM-based agentic systems in medicine and healthcare. *Nat. Mach. Intell.***6**, 1418–1420 (2024).
27. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proc. IEEE/CVF Int. Conf. Comp. Vis.* 9650–9660 (IEEE, 2021).
28. Contributors, M. OpenMMLab's pre-training toolbox and benchmark. <https://github.com/open-mmlab/mmpretrain> (2023).
29. Chang, J.-R. et al. Stain Mix-Up: Unsupervised Domain Generalization for Histopathology Images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 117–126 (Springer, 2021).
30. You, Y., Gitman, I. & Ginsburg, B. Large batch training of convolutional networks. Preprint at <https://arxiv.org/abs/1708.03888> (2017).
31. Chan, T. F. & Vese, L. A. Active contours without edges. *IEEE Trans. Image Process.***10**, 266–277 (2001).
32. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics***12**, 77 (2011).
33. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics***44**, 837–845 (1988).
34. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.***67**, 1–48 (2015).

35. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature***513**, 202–209 (2014).
36. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science***369**, 1318–1330 (2020).
37. Edwards, N. J. et al. The CPTAC data portal: a resource for cancer proteomics research. *J. Proteome Res.***14**, 2707–2713 (2015).
38. National Cancer Institute. Cancer Genome Characterization Initiative (CGCI).<https://www.cancer.gov/ccg/research/genome-sequencing/cgci>
39. Wilkinson, S. et al. Nascent prostate cancer heterogeneity drives evolution and resistance to intense hormonal therapy. *The Cancer Imaging Archive*<https://doi.org/10.7937/TCIA.JHQD-FR46> (2020).
40. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.***365**, 395–409 (2011).
41. Bándi, P. et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Trans. Med. Imaging***38**, 550–560 (2019).
42. Cancer Moonshot Biobank. Cancer Moonshot Biobank [dataset]. *The Cancer Imaging Archive* (2023).
43. Chowdhury, S. et al. Proteogenomic analysis of chemo-refractory high grade serous ovarian cancer (PTRC-HGSOC) [dataset]. *The Cancer Imaging Archive*<https://doi.org/10.7937/6RDA-P940> (2023).
44. Meltzer, P. S. et al. The Human Cancer Models Initiative (HCMI): a new resource for cancer research. *Cancer Res.***79** (suppl.), 1292 (2019).
45. Campanella, G., Hanna, M. G., Brogi, E. & Fuchs, T. J. Breast Metastases to Axillary Lymph Nodes [dataset]. *The Cancer Imaging Archive*<https://doi.org/10.7937/tcia.2019.3xbn2jcc> (2019).

Figures

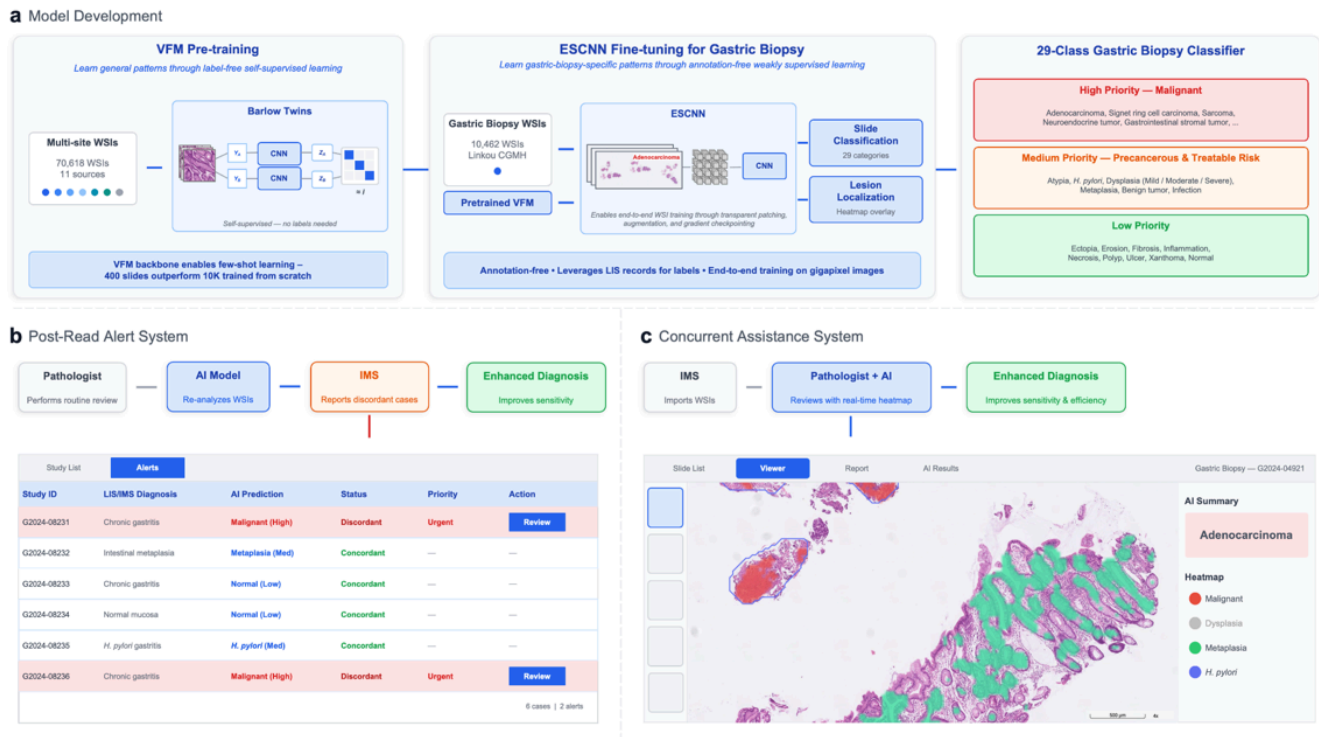


Figure 1

Overview of the AI-assisted gastric biopsy diagnostic system. **a**, Model development pipeline. A ResNet-50 encoder was first pre-trained on 70,618 whole slide images (WSIs) from 11 sources using Barlow Twins self-supervised learning to build a Vision Foundation Model (VFM). The VFM was then fine-tuned on 10,462 gastric biopsy WSIs from Linkou CGMH using the Enhanced Streaming CNN (ESCNN) method, which enables end-to-end training on gigapixel images using only slide-level labels without pixel-level annotations. The model outputs both a 29-class slide classification and lesion-localization heatmaps. **b**, Post-read alert workflow. The pathologist performs routine review and records a diagnosis in the LIS/IMS. The AI model subsequently re-analyzes the WSI and compares its prediction with the recorded diagnosis; discordant cases are flagged for secondary review. The dashboard illustrates an example alert interface showing concordant and discordant cases with their respective priority levels. **c**, Concurrent assistance workflow. During primary slide review, the AI model provides real-time heatmap overlays within the IMS viewer, color-coded for malignancy (red), dysplasia (yellow), metaplasia (green), and *H. pylori* (blue contours), which the pathologist can toggle on or off.

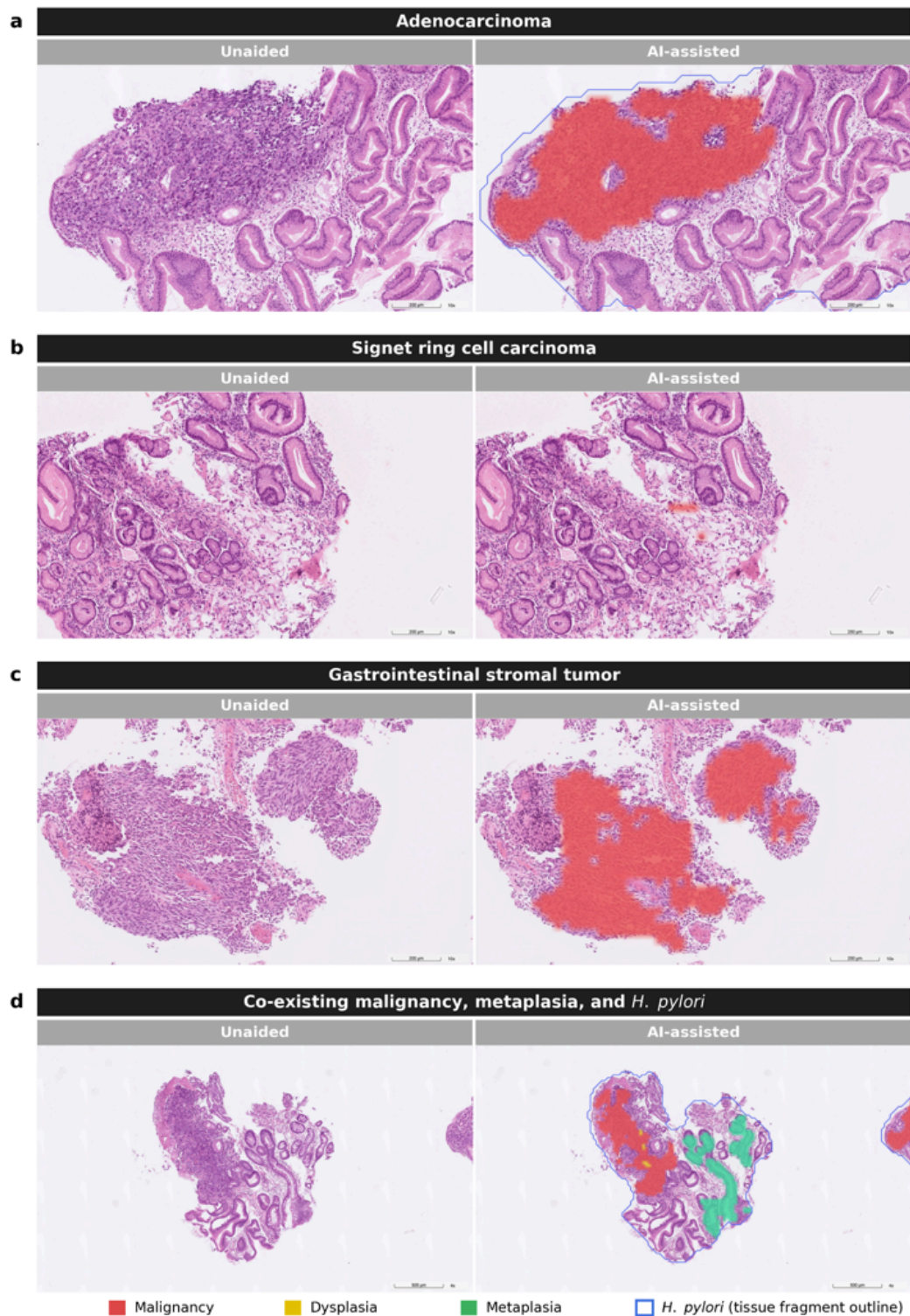


Figure 2

Qualitative lesion localization across diverse gastric pathologies. Representative biopsies from the internal test set are shown under unaided review (left) and with the AI-assisted overlay (right). The model produces pixel-resolution localization for **a**, adenocarcinoma; **b**, signet ring cell carcinoma; **c**, gastrointestinal stromal tumor (GIST); and **d**, a biopsy with co-existing malignancy, intestinal metaplasia, and *H. pylori* infection. Heatmaps are color-coded by lesion category (red, malignancy; yellow, dysplasia; green, metaplasia). For *H.*

pylori, tissue fragments containing positive heatmap signal are outlined in blue rather than rendered as a heatmap, prompting high-magnification review of the entire fragment to locate the organisms (visible in panels **a** and **d**). All overlays are produced by CAM from the same network trained with slide-level multi-label supervision; no pixel- or region-level annotations were used during training. Original H&E images are shown at 10× magnification with embedded scale bars (200 μm) except panel d at 4× (500 μm). Localization examples for additional pathological categories are shown in Supplementary Fig. 1.

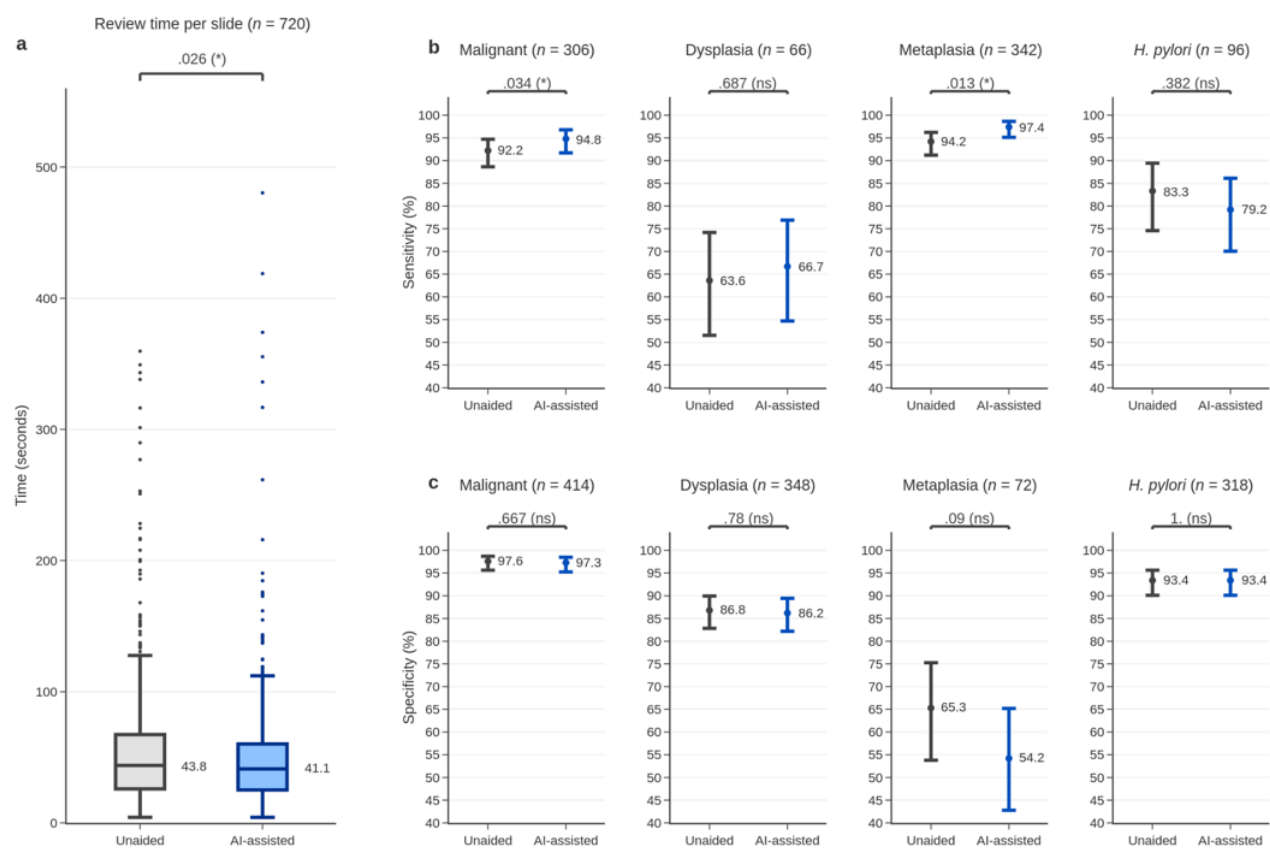


Figure 3

Quantitative results of the concurrent AI-assistance system in a simulated clinical workflow. Six pathologists reviewed 120 whole-slide images under unaided and AI-assisted conditions in a crossover design with a three-week washout period. **a**, Box-and-whisker plots of review time per slide (seconds) under unaided and AI-assisted conditions. Boxes indicate the interquartile range; center lines represent the median (annotated); whiskers extend to 1.5× IQR; outliers shown as circles. **b**, Diagnostic sensitivity for malignant, dysplasia, metaplasia, and *H. pylori* detection. Dysplasia, metaplasia, and *H. pylori* analyses were performed after excluding malignant cases. **c**, Diagnostic specificity for the corresponding categories. In **b** and **c**, dots represent point estimates; error bars indicate 95% confidence intervals calculated using Wilson's score method. *P* values were derived from generalized linear mixed-effects models (sensitivity and specificity) or linear mixed-effects models (review time).

Toward agentic AI pathology: orchestrating risk stratification, lesion localization, and structured reporting

Conceptual future workflow — an LLM-driven agent integrates model outputs with clinical context to produce a structured report

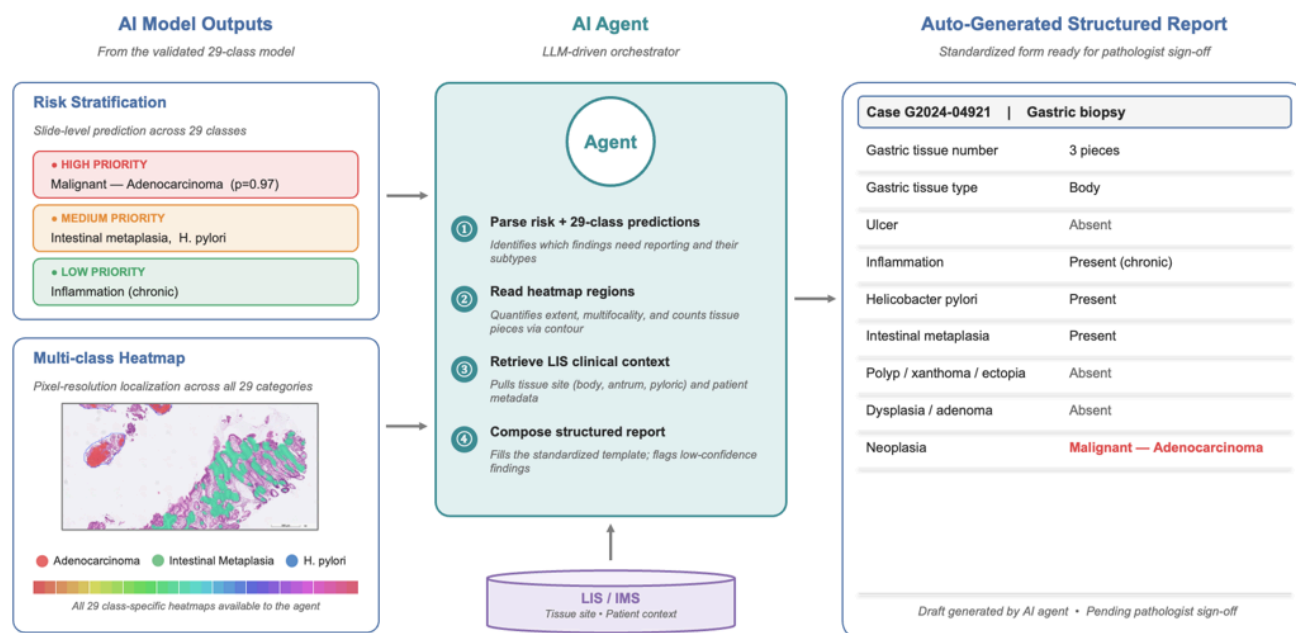


Figure 4

Conceptual workflow for an AI-agent–assisted structured reporting pipeline. The future vision builds on the two validated components of this study—post-read risk stratification (Fig. 1b) and pixel-resolution lesion localization across all 29 categories (Fig. 1c)—and adds an LLM-driven agent that also retrieves clinical context from the laboratory information system (LIS) to populate a standardized pathology report. The agent (i) parses the slide-level risk tiers together with the 29-class predictions to identify reportable findings and resolve subtypes (e.g., adenocarcinoma vs. signet ring cell carcinoma); (ii) reads the class-specific heatmaps to quantify lesion extent, multifocality, and the number of tissue pieces via the same contour-detection step used for *H. pylori* tissue-fragment outlining (Methods); (iii) retrieves the gastric tissue site (body, antrum, pyloric) and patient metadata from the LIS; and (iv) composes a structured report following the institutional template, flagging low-confidence fields for pathologist review. The example shown corresponds to a case harboring adenocarcinoma with coexisting intestinal metaplasia and *H. pylori* infection, analogous to Fig. 2d. The agent and auto-report generation are not part of the current evaluation; only the VFM-based classification and heatmap outputs feeding into the agent are validated in this study.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.docx](#)
- [rawdata20260605.xlsx](#)