Computer Science and Information Systems 23(2):801–826 https://doi.org/10.2298/CSIS251130024L

Development and Validation of a Few-Shot Rapid Screening Model for Gastrointestinal Cancers Using AGI Large Vision Models

Lijue Liu[1], Fangjie Yin[1], Genjian Yang[2], Qi Li[3], Siya Li[4], Teng Pan[5], Ting Liu[6], Jin Tang[1,7], Ruijie Ming[8], Yu Song[9], Xue Feng[10], Dan Wang[11], Xingang Zhou[6], Wenbai Chen[2], and Jinhai Deng[11,12]

[1] School of Automation, Central South University
410083 Changsha, China
{ljliu, tjin}@csu.edu.cn, yinfangjie2023@126.com
[2] School of Automation, Beijing Information Technology Science and University
102206 Beijing, China
457706420@qq.com, chenwb@bistu.edu.cn (corresponding author)
[3] Department of Pathology, Beijing Integrated Traditional Chinese and Western Medicine Hospital
100039 Beijing, China
15201232918@163.com
[4] CAS Blue Bay Cloud Technology (Guangdong) Co., Ltd.
518001 Guangzhou, China
1004297233@qq.com
[5] Longgang District Maternity & Child Healthcare Hospital of Shenzhen City, Longgang Maternity and Child Institute of Shantou University Medical College
518172 Shenzhen, China
2570758402@qq.com
[6] Department of Pathology, Beijing Ditan Hospital, Capital Medical University
100015 Beijing, China
liuting1981_2005@126.com, zhouxg1980@126.com (corresponding author)
[7] Xiangjiang Laboratory
410205 Changsha, China
tjin@csu.edu.cn
[8] Department of Oncology, Chongqing University Three Gorges Hospital
404010 Chongqing, China
ming_ruijie@cqu.edu.cn
[9] Department of Otolaryngology, Head & Neck Surgery, Peking University First Hospital
100034 Beijing, China
syandf@163.com
[10] Department of Respiratory and Critical Care Medicine, Tianjin Chest Hospital
300222 Tianjin, China
fengxuenku@163.com
[11] Richard Dimbleby Laboratory of Cancer Research, Randall Division and Division of Cancer and Pharmaceutical Sciences, King's College London
SE1 1UL London, UK
dan.7.wang@kcl.ac.uk, jinhaideng_kcl@163.com
[12] Guangzhou Baiyunshan Pharmaceutical Holding Co., Ltd. Baiyunshan Pharmaceutical General Factory/Guangdong Province Key Laboratory for Core Technology of Chemical Raw Materials and Pharmaceutical Formulations
510515 Guangzhou, China
jinhaideng_kcl@163.com (corresponding author)

Abstract. Existing deep learning models in digital pathology typically require extensive labeled data and show limited generalization across organs. In contrast, large vision models exhibit effective feature extraction capabilities, enabling pathological image analysis for gastrointestinal cancer with relatively small sample sizes. In this study, we developed a screening framework leveraging a large vision model for coarse-grained classification of gastric and colorectal tissues. The model was evaluated on multicenter cohorts and under limited-data conditions. Using labeled tiles from only 76 whole-slide images, the model achieved class-averaged sensitivity and precision of 0.9816 and 0.9808 on the internal test set, and 0.9161 and 0.9179 on the external test set. When trained with only 200 tiles per class from 20 whole-slide images, the model maintained comparable performance, achieving sensitivity and precision of 0.9548 and 0.9518. These findings suggest that the model has reliable performance across multicenter cohorts and potential applicability in clinical pathology workflows.

Keywords: Deep Learning, Gastrointestinal Cancers, Histopathology, Unified Screening.

1. Introduction

Gastrointestinal cancers, primarily including esophageal, gastric, and colorectal malignancies, are among the most prevalent human cancers [53]. Currently, histopathological analysis remains the gold standard for diagnosis [30], providing reliable tumor typing and informing treatment decisions through the examination of tissue or cell morphology. However, traditional pathological diagnosis requires pathologists to carefully examine entire slides by pathologists, which is labor-intensive, time-consuming, and susceptible to inter-observer variability [55, 9]. These challenges, combined with the global shortage of pathology specialists [47], highlight the growing need for developing automated computational approaches to assist pathologists in histopathological diagnosis.

Convolutional neural networks (CNNs) have fueled the explosive interest in applying deep learning to histopathology, owing to their ability to learn features directly from raw data [40]. For instance, Wang et al. [48] developed an Inception-V3-based method, achieving an area under the receiver operating characteristic curve (AUC) of 0.9880 in distinguishing colorectal cancer from normal tissue. Similarly, Song et al. [39] proposed a clinical pathology diagnostic system that reached nearly 100% sensitivity and an average specificity of 80.6% in a real-world cohort. In addition, Fu et al. [14] introduced the StoHisNet, a multiscale model that attained over 94% accuracy in classifying gastric pathological images, including normal tissue and adenocarcinoma subtypes.

Nevertheless, these studies largely focused on single-organ analysis. In practical clinical settings, however, pathologists frequently encounter diagnostic tasks involving multiple organs. Images from various organs show variation in staining, tissue architecture, resolution, and imaging modality[45]. These differences create a domain shift that can cause a classifier trained on one organ to perform poorly on another despite the same underlying malignant morphology [7].

Despite this, several studies have explored multi-organ diagnostic frameworks for gastrointestinal pathology. Iizuka et al. [20] utilized the same network architecture to classify adenocarcinoma, adenoma, and non-tumor tissues from gastric and intestinal pathological images, demonstrating the potential of CNNs for unified, coarse-grained classification

across the gastrointestinal tract. Likewise, Oh et al. [32] developed a two-stage gastric cancer model that generalized well to intestinal datasets, highlighting the feasibility of joint classification. Although these approaches demonstrate the possibility of cross-domain generalization, they, along with other CNN methods, require substantial annotated data for training, ranging from 85 to 7164 whole slide images (WSIs) [48, 39, 14, 20, 32, 52, 44, 1, 51, 18, 42, 19, 43, 23, 22, 15, 2]. In practice, obtaining sufficiently annotated pathological data remains costly and challenging, making model training under low-resource conditions an important problem.

Concurrently, the rise of large vision models (also known as visual foundation models) trained with self-supervision on large-scale datasets has emerged as a promising trend in pathological image analysis, enabling broad adaptability to diverse downstream tasks [11]. Previous studies have used datasets of 32,000 to 3,100,000 WSIs to successfully train pathology large vision models with parameter scales ranging from 28 million to 11 billion [54, 6, 46, 37, 50, 56]. Among them, Chen et al. [6] developed the UNI model, which surpassed state-of-the-art (SOTA) methods in multiple single-organ and tissue subtype classifications and achieved an AUC of 0.9750 across 32 cancer types, while Wang et al. [50]'s CHIEF model demonstrated strong performance in identifying 11 cancer types, with AUCs ranging from 0.9098 to 0.9943.

Most large vision models such as UNI[6], Virchow2[56], H-optimus-0[37] employ a self-supervised learning approach called DINOv2[33] in the pretraining stage, which has been shown to yield strong, off-the-shelf representations for downstream tasks without the need for further fine-tuning with labeled data. When downstream models build upon these large vision models, substantially less data and computational resources are required[31], thereby reducing sample complexity. This characteristic closely relates to the paradigm of few-shot or low-resource learning. Few-shot learning is a specialized branch of deep learning algorithms that addresses this challenge by enabling models to learn new concepts from a few labeled examples, mimicking the human ability to generalize from limited experience[13]. Several recent large vision models [12, 6, 25] have also evaluated their performance under few-shot learning settings. In clinical practice, many histopathological tasks suffer from extremely limited annotated samples. By leveraging pre-trained models' feature representation capacity, few-shot learning can largely reduce intra-class variation, enabling models to focus on more discriminative morphological patterns[38].

1.1. Main Algorithmic Contributions

Motivated by these observations, this study leverages prior knowledge from a pathology large vision model to develop a high-performance unified screening system for both gastric and colorectal cancers, using only 76 training WSIs. It aims to reduce the clinical diagnostic workload while improving screening efficiency. The main contributions of this study are as follows:

1) Unlike existing works that focus on single-organ classification, this framework can achieve reliable performance in unified screening for both gastric and colorectal cancers even under low-resource settings.

2) This framework incorporates a gated recurrent unit (GRU) module to refine the tile-level representations extracted from pretrained pathology foundation models for enhancing the discriminative ability.

3) To address the challenge of difficult classification, this framework applies an adaptive weighted loss based on recall and the average loss over historical epochs.

2. Methods

2.1. Data Collection

In this study, a total of 149 hematoxylin and eosin (H&E)-stained slides were collected from two hospitals: slides collected from Beijing Ditan Hospital Capital Medical University (BDHCMU) were used for model development, and an independent cohort from Beijing Integrated Traditional Chinese and Western Medicine Hospital (BITCWMH) was used as an external test set to evaluate generalization across institutions. To ensure data quality, five slides with blurred regions or marker interference were excluded. The final internal cohort comprised 96 slides covering various pathological stages, including 42 gastric and 54 colorectal samples. It was randomly divided into internal training and internal testing cohorts in an approximate ratio of 8:2 prior to tile extraction. Consequently, all corresponding tiles derived from a given WSI were assigned exclusively to either the training or testing cohort. The external testing cohort consisted of 48 slides, including 15 gastric and 33 colorectal slides, also representing a range of pathological conditions. The demographic characteristics of the patients in these two cohorts are shown in Table 1. Notably, only 76 slides were used for model training in this study, which is considerably fewer than those reported in most previous studies. In the relevant literature, we investigated [48, 39, 14, 20, 32, 52, 44, 1, 51, 18, 42, 19, 43, 23, 22, 15, 2], the number of WSIs used for training in previous studies ranged from 85 to 7164. To further evaluate the generalizability of our model, an independent validation cohort comprising 100 H&E-stained slides of gastric and colorectal tissues was obtained from The Cancer Genome Atlas (TCGA) public dataset. This cohort included 50 malignant and 50 non-malignant slides. The overall cohort construction process is illustrated in Fig. 1.

Table 1. The Demographic Characteristics of the Patients

Variables		Internal Cohort	External Cohort
Sex	Female	27.40%	52.50%
	Male	72.60%	47.50%
Age	Median	56	60
	Range	30-83	31-88
	Mean $\pm$ SD	53.40 $\pm$ 11.58	60.53 $\pm$ 13.86

2.2. Data Annotation and Preprocessing

Pathologists used ASAP1.9 to annotate 144 slides at the region of interest (ROI) level and categorized them into three subclasses. Two pathologists independently labeled and diagnosed the ROIs, and a third experienced pathologist confirmed the final annotations to ensure label quality and consistency. Some examples of pathologists' labeling are shown

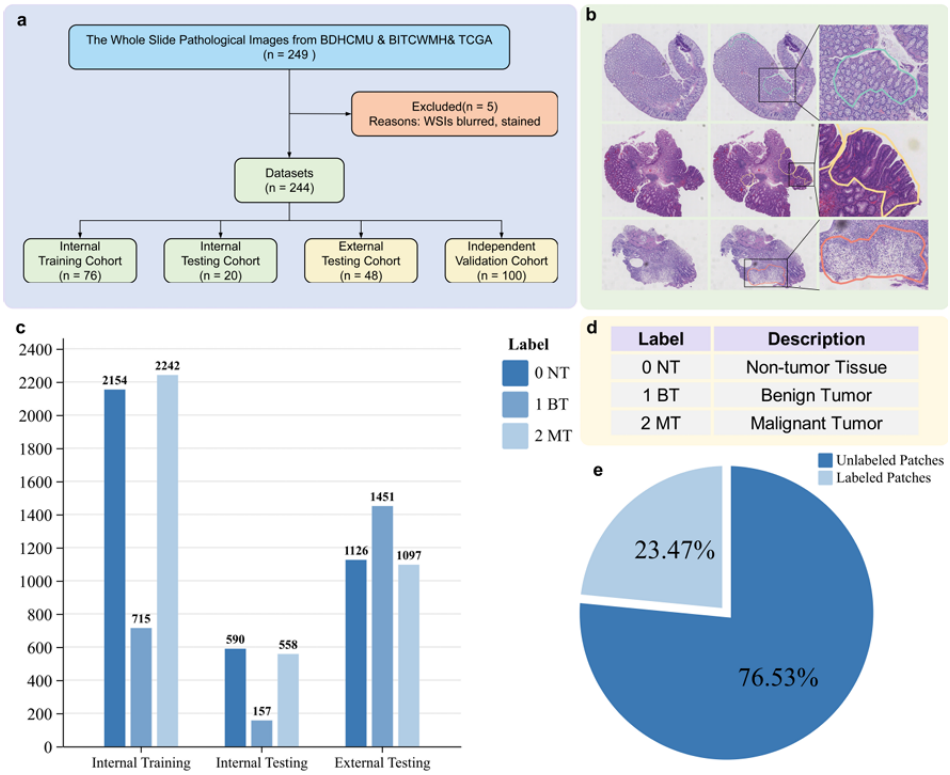

Fig. 1. Data cohort construction and its description. (a) The process of the cohort construction. (b) Some examples of pathologists' labeling. (c) The number of tiles in the internal and external cohorts. (d) Different levels of labeling and their descriptions. (e) The percentage distribution of annotated tiles.

in Fig. 1b. For the internal dataset, pathologists did not label all areas of the slide but used a selective labeling strategy. The 144 slides from BDHCMU & BITCWMH can be divided into 42,996 tiles, and the number of tiles divided after annotation is 10,090, accounting for approximately 23.47%. The proportion of annotated tiles is shown in Fig. 1e, while the total number of tiles in both the internal and external cohorts is shown in Fig. 1c. This strategy effectively controls the amount of data at the tile level while ensuring that the core information of the organization is retained, and the amount of data at the tile level is also as small as possible, thus fitting the needs of this study for limited data scenarios and providing accurate and efficient training samples for exploring the performance of the model under limited data conditions.

The different label levels and their corresponding descriptions are presented in Fig. 1d. Non-tumor tissues (Level 0, NT) include normal gastric mucosa, normal intestinal mucosa, entericized gastric mucosa, hyperplastic polyps, inflammation, fundic gland polyps, and mild atrophic glands. Benign tumors (Level 1, BT) include low-grade villous adenomas, low-grade tubular adenomas, and low-grade serrated adenomas. Malignant tumors (Level 2, MT) include high-grade tubular adenomas, signet-ring cell carcinomas, poorly differentiated adenocarcinomas, moderately differentiated adenocarcinomas, and highly differentiated adenocarcinomas. For the external dataset, pathologists assigned diagnostic labels at the WSI level to support the evaluation of the model's diagnostic performance across entire slides.

In this study, all slides were processed at a magnification of $10\times$ with a resolution of 0.8299 μm / pixel. The gastrointestinal dataset consists of ROIs annotated by pathologists. For each WSI, the annotated ROIs were independently tiled, and all resulting tiles from that WSI were used for training according to their class labels. Tiles were then fed into the model in batches with random sampling. Since the size of each WSI and the number and type of ROIs they contains vary, the number of tiles sampled from each WSI also varies, with a mean of 82 tiles per WSI (range: 5 – 1731), as shown in Table 2.

To extract the foreground from each slide and eliminate large internal cavities within the ROI, a segmentation threshold of 200 was empirically determined based on the actual characteristics of the scanned images. Subsequently, the extracted ROI is segmented into tiles of 224 $\times$ 224 pixels and saved for easy retrieval at a later stage based on the coordinate approach.

In order to achieve a relatively balanced number of tiles extracted for each type of label in the training dataset, data augmentation techniques including random rotation, horizontal flipping, and vertical flipping were applied.

2.3. Unified Screening System for Gastrointestinal Cancers

A unified screening system for gastrointestinal cancers called VGA (Virchow2-based GRU with Adaptive-weighted Loss) was developed to differentiate between non-tumor tissue, benign tumor tissue, and malignant tumor tissue. The overall workflow of the proposed screening system in this study is shown in Fig. 2.

First, since WSIs may contain up to tens of billions of pixels, feeding them directly into a neural network will be a huge challenge in terms of computational resources and storage overhead [21]. Resizing the entire image to a lower resolution would lead to the loss of cellular-level information, which will lead to a significant reduction in recognition accuracy. Therefore, it is common for researchers to adopt the strategy of dividing the WSI

into smaller tiles in order to adapt the processing capabilities of deep learning models [4, 27]. Drawing on this idea, this study analyzed WSIs at the tile level, with each tile labeled according to the pathological diagnosis of its ROI.

On the one hand, vision transformers (ViTs) have demonstrated SOTA performance in various computer vision tasks [16]. On the other hand, the large vision model prelearns collections of pathological images from different organs. Through the deep neural network structure, self-supervised learning, and other techniques, it can automatically extract meaningful features from images and construct a visual representation space with rich semantic information and strong generalization capabilities. Such pretraining enables the model to utilize existing prior knowledge when applied to downstream tasks, thereby reducing its reliance on the amount of data for the target task. Therefore, the Virchow2 large vision model [56] with ViT as the underlying architecture was selected as the feature extractor in this study, aiming at obtaining high-quality generalized image representations and providing a more discriminative and robust feature base for subsequent prediction and classification tasks. And its parameters were frozen during training.

Virchow2 is a ViT architecture with 4 registers, which consists of 32 visual transformer blocks equipped with 16 heads at the attention layer with an embedding dimension of 1280. The model was trained on a large-scale medical image dataset containing 3.1 million whole-slide histology images stained with H&E and immunohistochemistry (IHC) staining and covering tissue samples from multiple parts of the human body, including the stomach and colorectum.

After a tile-level feature is extracted by the pretrained Virchow2 encoder, a feature sequence consisting of the CLS token and the mean patch token is constructed to represent the tile. This feature sequence is then fed into an RNN for feature refinement. Specifically, the RNN model used in this study consists of two GRU [8] layers, with a hidden size of 256. The relevant hyperparameters are listed in Table 2. The GRU's update and reset gates enable adaptive interaction between global information (CLS token) and aggregated local patch tokens, producing the refined tile-level feature. The final feature is passed through a linear classifier to generate tile-level prediction. Tile-level predictions are then aggregated across all tiles to obtain ROI- and WSI-level predictions.

Tiles are extracted from the ROIs corresponding to each label category. However, due to the imbalance in the number of ROIs across different label levels, the resulting number of tiles per class is also unevenly distributed. To address this imbalance, this study introduces an adaptive weighted loss function in addition to employing data augmentation. This function is able to automatically adjust the weights of different categories according to the training state of the model, thus preventing the model from favoring easily classifiable categories and ignoring other categories. Specifically, it dynamically updates class weights based on per-class recall using an exponential moving average (EMA), ensuring that underrepresented classes receive higher emphasis. Furthermore, it combines cross-entropy and focal loss in a weighted manner, with the balance between them adjusted according to the losses from recent epochs.

Formally, let y denote the true label and p the predicted probability vector for a sample. The detailed formulation of the adaptive weighted loss is presented below:

$$L_{AW}(p, y) = w_y \times [\beta L_{CE}(p, y) + (1 - \beta) L_{\text{Focal}}(p, y)] \tag{1}$$

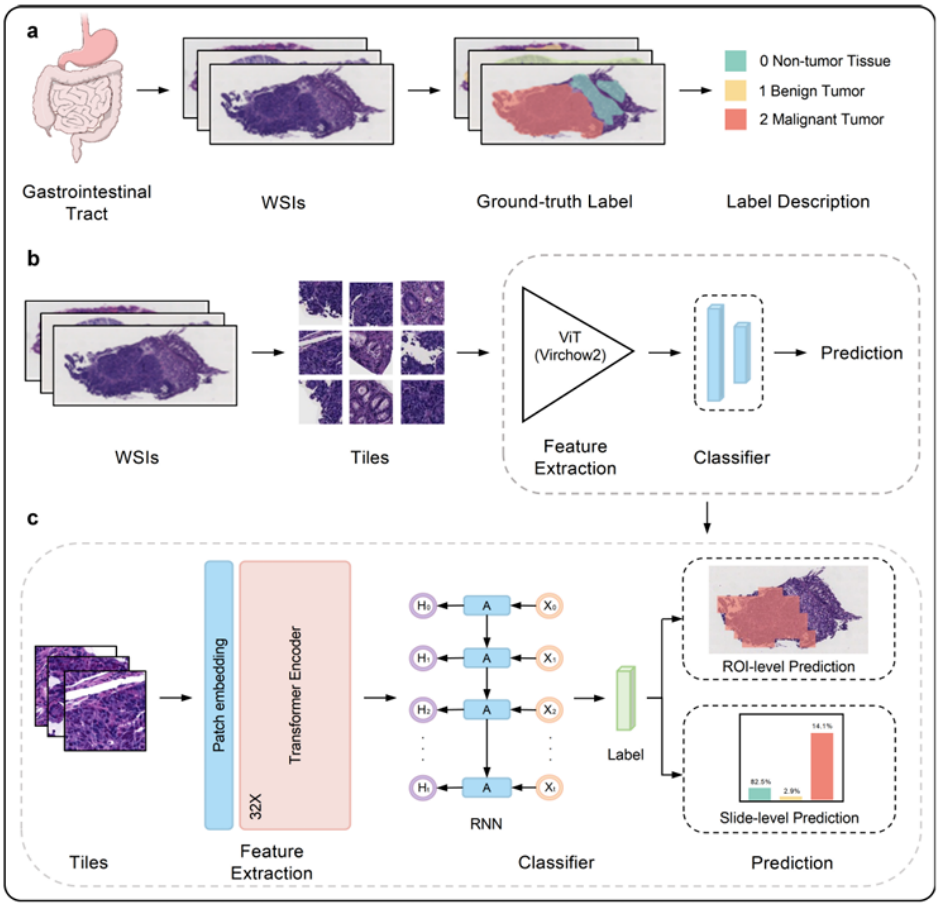

Fig. 2. The workflow of the unified gastrointestinal cancers screening system developed in this study. Subfigure (a) shows the segmentation of tissue regions on pathology slides based on expert annotations, with diagnostic labels assigned to each region, including non-tumor (blue for Level 0), benign tumor (orange for Level 1), and malignant tumor (red for Level 2). Subfigure (b) shows the extraction of tiles from the segmented tissue regions and the use of the system for training and inference. Subfigure (c) shows the training and inference process of the system. The tiles extracted from the WSI are first processed by the pre-trained Virchow2 model for feature extraction. The resulting features are then fed into a recurrent neural network (RNN)-based classifier to generate final predictions, which are subsequently compared with the pathologists' diagnoses for validation

where L_{AW} denotes the adaptive weighted loss; L_{CE} denotes the cross-entropy loss; L_{Focal} denotes the focal loss; w_y denotes the weight of the true class; β controls the relative contribution of cross-entropy loss and focal loss, dynamically adjusted based on recent epoch loss trends.

Each class c is assigned a weight w_c with the range of 1 to w_{max}, which is updated based on its recall using an exponential moving average:

$$w_c^{e_{\text{current}}} = \lambda w_c^{e_{\text{last}}} + (1-\lambda)\frac{1}{\mathrm{r_c} + \epsilon} \tag{2}$$

where $w_c^{e_{\text{current}}}$ denotes the w_c of the current epoch; $w_c^{e_{\text{last}}}$ denotes the w_c of the last epoch; λ denotes the EMA momentum; ϵ prevents division by zero.

The balance weight β is dynamically updated based on the ratio of the latest epoch loss to the historical average loss, which is computed over the past t epochs. If the latest epoch loss decreases by no more than τ compared to the historical average, the balance weight β is decreased by $\triangle\beta$ at each epoch, starting from 1 and with a lower bound of 0. The hyperparameters w_{max} = 2.0, λ = 0.9, t = 3, τ = 0.05, $\triangle\beta$ = 0.02 were determined based on preliminary experiments.

2.4. Training Strategy

In order to improve the robustness and generalization ability of the model training, this study uses a 3-fold cross-validation strategy to optimize the training dataset. Also, the Adam optimizer and cosine scheduler are used for learning rate and weight decay with an initial learning rate of 0.0001. During the training iterations, an adaptive weighted loss function is utilized to calculate the loss and update the model weights based on the minimum loss achieved at the end of each iteration. The training process is configured for a maximum of 200 iterations, with an early stopping mechanism implemented. If the validation loss is not further reduced after 10 consecutive training iterations, the training is terminated. These parameter settings are determined according to experimental performance and may significantly influence the final diagnostic outcomes. Therefore, selecting and tuning appropriate parameters during the experimental process is essential for achieving optimal results. All experiments were performed on an NVIDIA GeForce RTX 3090 using Python version 3.8.19 and Pytorch version 2.3.1. The model hyperparameters and settings are shown in Table 2.

2.5. Quantitative and Statistical Analysis

To evaluate the performance of multi-class classification at the tile level, this study uses class-by-class tile-level sensitivity, specificity, and precision, which are defined as follows, in conjunction with the requirements of medical research and commonly used deep learning evaluation metrics:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

Table 2. The Model Hyperparameters and Settings

Hyperparameter	Setting	Notes
Learning rate	1e-4	Initial learning rate for trainable layers
Optimizer	Adam	Weight decay = 1e-4
Batch size	32	
Maximum training epochs	200	
Early stopping patience	10	Monitored on validation loss
Number of sampled tiles per WSI	82	Average (range:5-1371)
Virchow2 encoder	Frozen	Pretrained encoder
GRU input dimension	1280	Matches feature embedding
GRU hidden dimension	256	
GRU layers	2	Number of stacked GRU layers
GRU dropout	0.3	Applied within GRU and before FC layer
GRU weight init	Xavier(input), Orthogonal(hidden)	Bias zeros, update gate bias=1
GRU output	Last hidden state	Passed to linear classifier

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

where TN, FP, TP, and FN represent true-negative, false-positive, true-positive, and false-negative situations for each category, respectively.

The evaluation at the ROI-level and WSI-level was completed by aggregating the tile-level prediction results within each ROI or WSI. During the training process, no information from the test ROI and WSI is used at all.

To evaluate the network performance at the region-of-interest level, this study draws on the idea of the paper [32] to utilize the tile-level accuracy of each ROI, with the metric defined as follows:

$$Accuracy_{ROI} = \frac{1}{|R_c|} \sum_{r \in R_c} \left(\frac{1}{|P_r|} \sum_{p \in P_r} M_p \cdot \mathbf{1} \left(\widehat{y_p} = y_p \right) \right) \tag{6}$$

where R_c is the set of samples for each region-of-interest level truth class; P_r is the set of tiles within the r_{th} ROI; $\widehat{y}_p$ is the predicted value of the p_{th} tile; y_p is the corresponding tile set label; M_p is the foreground mask, which is set to 1 when a tile accounts for more than 50% of the foreground pixels.

To evaluate the network performance at the WSI-level, this study follows clinical practice guidelines, a slide was classified as high-grade if at least one high-grade tile was identified, with the formula defined as follows:

$$Y_{slide} = \max_{p \in P_w} \mathbb{I}(\hat{y}_p = H) \tag{7}$$

where Y_{slide} is the label of the WSI; P_w is the set of tiles within the WSI; $\widehat{y}_p$ is the predicted label of the p_{th} tile; H denotes the high-grade label.

Considering the inconsistency in the number of samples in each category in the test dataset, this study used the weighted average method to find all the class-averaged assessment metrics, which were calculated using the following formula:

$$Avg = \frac{\sum_{i=1}^{C} n_i \times \text{metric}}{\sum_{i=1}^{C} n_i} \tag{8}$$

where metric is the corresponding metric value; C is the number of categories; n_i is the number of samples in the i_{th} category.

In addition, this study used AUC to compare the classification performance of different models. All metrics were calculated using the Scikit-learn package [35]. To ensure reliable performance assessment, 95% confidence intervals (CI) were calculated for the class-averaged metrics using the standard error method (mean $\pm 1.96 \times$ standard error). To evaluate whether the differences in model performance were statistically significant, the DeLong test was applied to compare AUC values between models on the internal test dataset. A two-sided p-value < 0.05 was considered statistically significant.

3. Results

3.1. Ablation Study

Table 3. Component configurations

Components	Linear Head	Mean Patch Token	GRU	Adaptive Weighted Loss
①	✓	-	-	-
②	✓	✓	-	-
③	✓	✓	-	✓
④	-	✓	✓	-
⑤	-	✓	✓	✓

To assess the contribution of each component in our model, we conducted an ablation study, and the relevant results are summarized in Table 3 and Table 4 [1]. The baseline (①) only uses the CLS token as input to the linear head for classification. Its performance provides a reasonable starting point for the ablation study. The combination of concatenating the CLS token and the mean patch token (②) shows improvements in all metrics, demonstrating that the mean patch token is useful for supplementing global information. Comparisons of combinations ③, ④, and ⑤ indicate that the use of GRU and adaptive weighted loss further enhances the model's discrimination ability. In summary, the ablation results confirm that each module has a positive contribution, and their integration is helpful for improving classification performance.

3.2. Tile-level Classification of Gastrointestinal Tissues Using VGA

In digital pathology image analysis, dividing WSI into tiles for analysis has become a widely accepted strategy in AI-assisted pathology diagnostic systems [49, 10, 5, 26]. This strategy for analysis not only significantly reduces computational complexity but also captures localized pathological features at a finer level. In addition, the tile level facilitates the generation of heat maps, accurate lesion localization, and informed region-level

[1] Note: Sens denotes sensitivity, Spec denotes specificity, Prec denotes precision, and Avg denotes average class indicator value.

Table 4. Ablation study results on the internal test dataset

Components	Level	Sens	Spec	Prec	AUC
①	0 NT	0.9610	0.9919	0.9789	0.9985
	1 BT	0.8769	0.9746	0.8514	0.9892
	2 MT	0.9833	0.9610	0.9729	0.9880
	Avg(95%)	0.9604±0.0104	0.9766±0.0082	0.9610±0.0103	0.9929±0.0045
②	0 NT	0.9633	0.9954	0.9884	0.9987
	1 BT	0.8938	0.9886	0.8662	0.9902
	2 MT	0.9815	0.9864	0.9642	0.9884
	Avg(95%)	0.9627±0.0102	0.9907±0.0052	0.9634±0.0100	0.9933±0.0044
③	0 NT	0.9695	0.9961	0.9902	0.9990
	1 BT	0.9236	0.9901	0.8865	0.9956
	2 MT	0.9827	0.9897	0.9728	0.9958
	Avg(95%)	0.9696±0.0093	0.9927±0.0046	0.9703±0.0090	0.9972±0.0029
④	0 NT	0.9830	0.9952	0.9882	0.9919
	1 BT	0.9002	0.9956	0.9451	0.9780
	2 MT	0.9845	0.9872	0.9678	0.9910
	Avg(95%)	0.9808±0.0086	0.9950±0.0049	0.9743±0.0086	0.9898±0.0054
⑤	0 NT	0.9898	0.9973	0.9932	0.9997
	1 BT	0.9172	0.9968	0.9600	0.9992
	2 MT	0.9910	0.9899	0.9736	0.9962
	Avg(95%)	0.9816±0.0072	0.9941±0.0042	0.9808±0.0074	0.9981±0.0023

decision-making. It serves as a critical intermediate step between low-level feature extraction and high-level clinical inference. Therefore, this study begins with a systematic evaluation of the model's performance at the tile level to verify that it better understands the tile information.

We compared the VGA model developed in this study with three other approaches: H-optimus-0 [37], CHIEF [50], and InceptionV3 [41]. Among them, H-optimus-0 and CHIEF are two models that perform well in existing pathological large vision models [54, 6, 46, 37, 50, 56], while InceptionV3 is a backbone or baseline network that has been widely adopted in previous pathological image analysis studies [48, 10, 29, 24, 34]. All models were trained and tested using the same dataset, following identical experimental procedures to ensure a fair comparison.

The ROCs of the four methods and their AUCs on both the internal and external test datasets are shown in Fig. 3. The tile-level performance metrics of the four methods on the internal and external test datasets are summarized in Table 5 [2] and Table 6 respectively. The VGA method developed in this study achieved an average sensitivity of 0.9816 (95% [CI]: 0.9744, 0.9888) on the internal test dataset, an average specificity of 0.9941 (95% [CI]: 0.9899, 0.9983), an average precision of 0.9808 (95% [CI]: 0.9734, 0.9882) and an average AUC of 0.9981 (95% [CI]: 0.9958, 1.0000). In order to focus on the multi-category classification ability of AI systems in the clinic, we used category-averaged sensitivity as the primary evaluation metric. It can be seen that the VGA method developed in this study achieves higher performance compared to similar methods. Meanwhile, the VGA model developed in this study showed robust generalization performance during external validation, with a class-averaged sensitivity of 0.9161 (95% [CI] 0.9071,0.9251)

[2] Note: p-values are calculated using the DeLong test with VGA as the reference model.

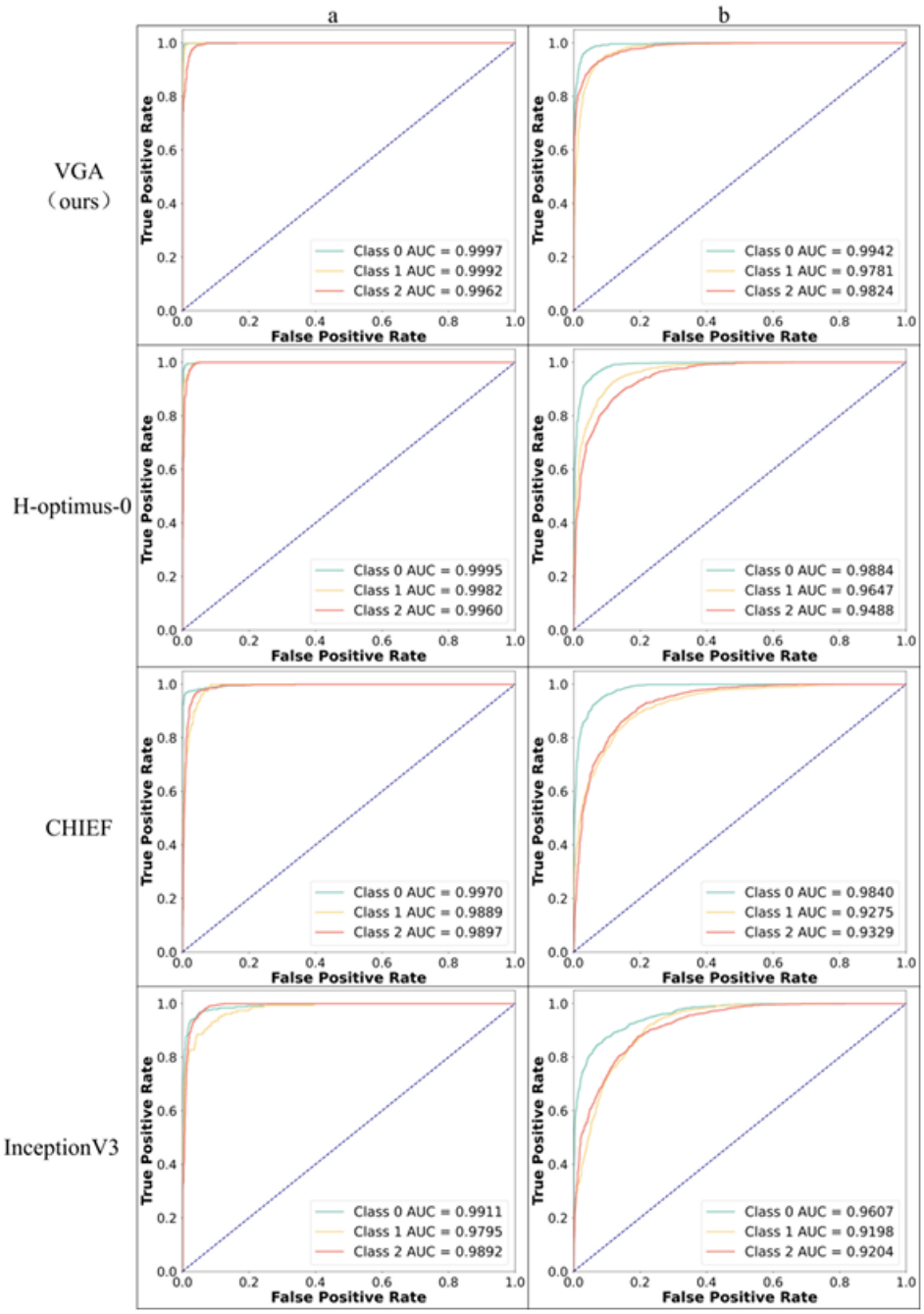

Fig. 3. Receiver operating characteristic curves (ROCs) and their AUCs for the four methods. (a) ROC and AUC of four methods on the internal test dataset. (b) ROC and AUC of four methods on the external test dataset

Table 5. Tile-level performance metrics of different methods on the internal test dataset

Models	Level	Internal Test Dataset				
		Sens	Spec	Prec	AUC	p-value
VGA (Ours)	0 NT	0.9898	0.9973	0.9932	0.9997	-
	1 BT	0.9172	0.9968	0.9600	0.9992	-
	2 MT	0.9910	0.9899	0.9736	0.9962	-
	Avg(95%)	0.9816±0.0072	0.9941±0.0042	0.9808±0.0074	0.9981±0.0023	-
H-optimus-0	0 NT	0.9847	0.9911	0.9781	0.9995	0.0073
	1 BT	0.8981	0.9958	0.9463	0.9982	0.0097
	2 MT	0.9803	0.9893	0.9716	0.9960	0.0270
	Avg(95%)	0.9724±0.0088	0.9909±0.0052	0.9715±0.0090	0.9978±0.0025	-
CHIEF	0 NT	0.9475	0.9883	0.9705	0.9970	0.0005
	1 BT	0.8662	0.9831	0.8095	0.9889	< 0.0001
	2 MT	0.9642	0.9852	0.9607	0.9897	0.0014
	Avg(95%)	0.9449±0.0123	0.9863±0.0063	0.9469±0.0118	0.9929±0.0045	-
Inception V3	0 NT	0.8695	0.9671	0.9144	0.9911	< 0.0001
	1 BT	0.7325	0.9847	0.7986	0.9795	0.0001
	2 MT	0.9409	0.9469	0.8692	0.9892	< 0.0001
	Avg(95%)	0.8835±0.0170	0.9606±0.0105	0.8811±0.0174	0.9889±0.0057	-

Table 6. Tile-level performance metrics of different methods on the external test dataset

Models	Level	External Test Dataset			
		Sens	Spec	Prec	AUC
VGA (Ours)	0 NT	0.9218	0.9765	0.9454	0.9942
	1 BT	0.9021	0.9357	0.9015	0.9781
	2 MT	0.9289	0.9616	0.9114	0.9824
	Avg(95%)	0.9161±0.0090	0.9559±0.0066	0.9179±0.0089	0.9843±0.0040
H-optimus-0	0 NT	0.8934	0.9776	0.9464	0.9884
	1 BT	0.8980	0.9172	0.8763	0.9647
	2 MT	0.8879	0.9530	0.8895	0.9488
	Avg(95%)	0.8936±0.0100	0.9464±0.0072	0.9017±0.0096	0.9672±0.0057
CHIEF	0 NT	0.9405	0.9195	0.8403	0.9840
	1 BT	0.7733	0.9568	0.9212	0.9275
	2 MT	0.8915	0.9305	0.8453	0.9329
	Avg(95%)	0.8598±0.0110	0.9375±0.0078	0.8737±0.0107	0.9464±0.0072
Inception V3	0 NT	0.7469	0.8905	0.7509	0.9607
	1 BT	0.5782	0.9208	0.8266	0.9198
	2 MT	0.8778	0.9872	0.6454	0.9204
	Avg(95%)	0.7194±0.0140	0.9313±0.0081	0.7493±0.0138	0.9325±0.0081

and a class-averaged precision of 0.9179 (95% [CI] 0.9090, 0.9268) on the external test dataset. In contrast, the other methods exhibited a substantial decline in performance, primarily due to overfitting on the internal training dataset. The better tile-level prediction performance of the VGA model can be largely attributed to the use of the Virchow2 large vision model as a feature extractor, which was pre-trained on 3.1 million WSIs, significantly surpassing the pre-training dataset sizes of H-optimus-0 and CHIEF used in the comparison models. It is worth emphasizing that only 5,111 labeled tiles from 76 WSIs (out of a total number of 29,457 tiles) were used for training in this study. Consequently, InceptionV3 exhibited suboptimal performance in this task, which further underscores the advantages of the large vision model as a feature extractor in downstream tasks and provides strong support for the development of subsequent AI-assisted systems.

3.3. ROI-level Classification of Gastrointestinal Tissues Using VGA

Accurate classification at the ROI level is essential for identifying tumor regions and guiding precise pathological diagnoses. It facilitates localized analysis within WSIs, enabling finer-grained decision-making and improving the accuracy of clinical workflows. Table 7[3] shows the average classification accuracy of the ROI-level of the different models on the internal and external test datasets. While the H-optimus-0 model achieved the highest ROI-level class average accuracy on the internal test dataset, its performance on the external test dataset was lower than the VGA model developed in this study. Overall, the VGA model developed in this study demonstrated the stronger generalization performance among the four models.

Table 7. ROI-level performance metrics for different models on the test dataset

Model	$\text{Accuracy}_{\text{ROI}}$	
	Internal Test Set	External Test Set
VGA(Ours)	0.9613	0.9175
H-optimus-0	0.9659	0.9029
CHIEF	0.9301	0.8901
Inception V3	0.8834	0.7121

Some of the ROI-level prediction results of the four models on the test dataset are given in Fig. 4. As illustrated, the VGA model produces predictions that most closely align with the ground truth annotations, demonstrating competitive accuracy compared to the other three models. In contrast, the remaining three models show varying degrees of deviation from the true labels, with InceptionV3 exhibiting the greatest discrepancy. From the clinical point of view, the VGA model developed in this study appears more suitable for application in gastrointestinal cancers screening, where it can assist pathologists in improving diagnostic efficiency and accuracy.

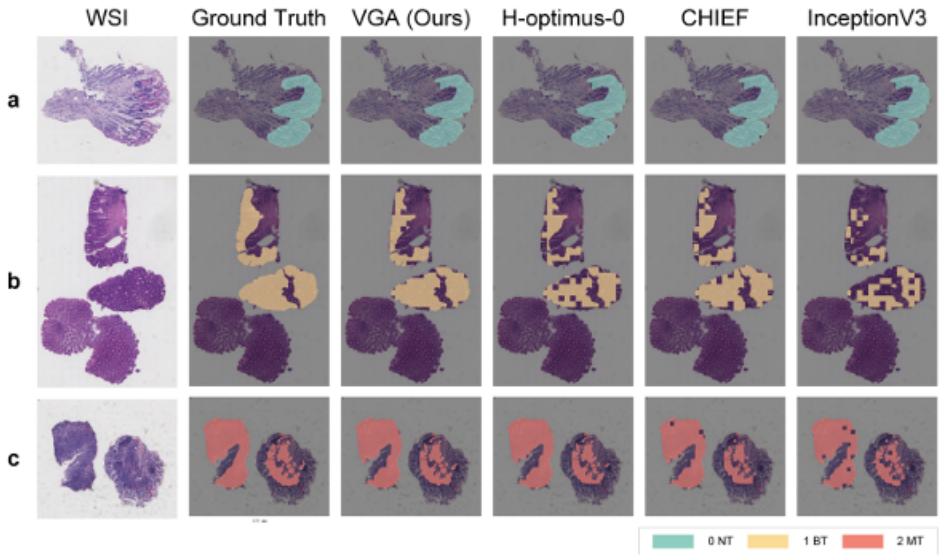

Fig. 4. ROI-level prediction results for different models on the test dataset

3.4. VGA's WSI-level Cancer Screening Capabilities

To evaluate the generalizability of the proposed algorithm across multi-center datasets, this study performed WSI-level testing on an external test dataset and an independent validation dataset selected from TCGA. Following clinical practice guidelines, a slide was classified as high-grade if at least one high-grade tile was identified. Among the 48 WSIs in the external test dataset, 28 were classified as non-malignant and 20 as malignant. At the same time, we invited a pathologist to review these 48 WSIs, confirming that among the malignant predictions, 5 were non-malignant, while the remaining were correctly classified. In the WSI-level test, the sensitivity of the VGA model for malignant tumor slide detection was 100%, and the specificity was 84.85%, indicating its ability to accurately exclude non-malignant tumor slides. In the independent validation dataset (100 slides), the VGA model achieved 98.00% sensitivity and 82.00% specificity for malignant tumor slide detection, further demonstrating its robust generalization performance. Fig. 5a and Fig. 5b illustrate the confusion matrix plots of the VGA model on the external test dataset and the independent validation dataset. Meanwhile, the average time for the VGA model to analyze a WSI is 25 seconds. This means that in combination with the initial screening results from the VGA model, pathologists may focus on high-risk areas in WSIs in a short period of time, thus improving the efficiency and accuracy of gastrointestinal cancers screening.

Fig. 6 shows a heatmap generated from a subset of slides during WSI-level testing of the VGA model on the external dataset. More heatmaps are presented in Fig.1 and Fig.2 of the Supplementary Materials. The diagnostic predictions of the VGA model are based on the classification probability outputs of all tiles within each slide, which can be used

[3] Note: $\text{Accuracy}_{\text{ROI}}$ indicates the average accuracy of the ROI class.

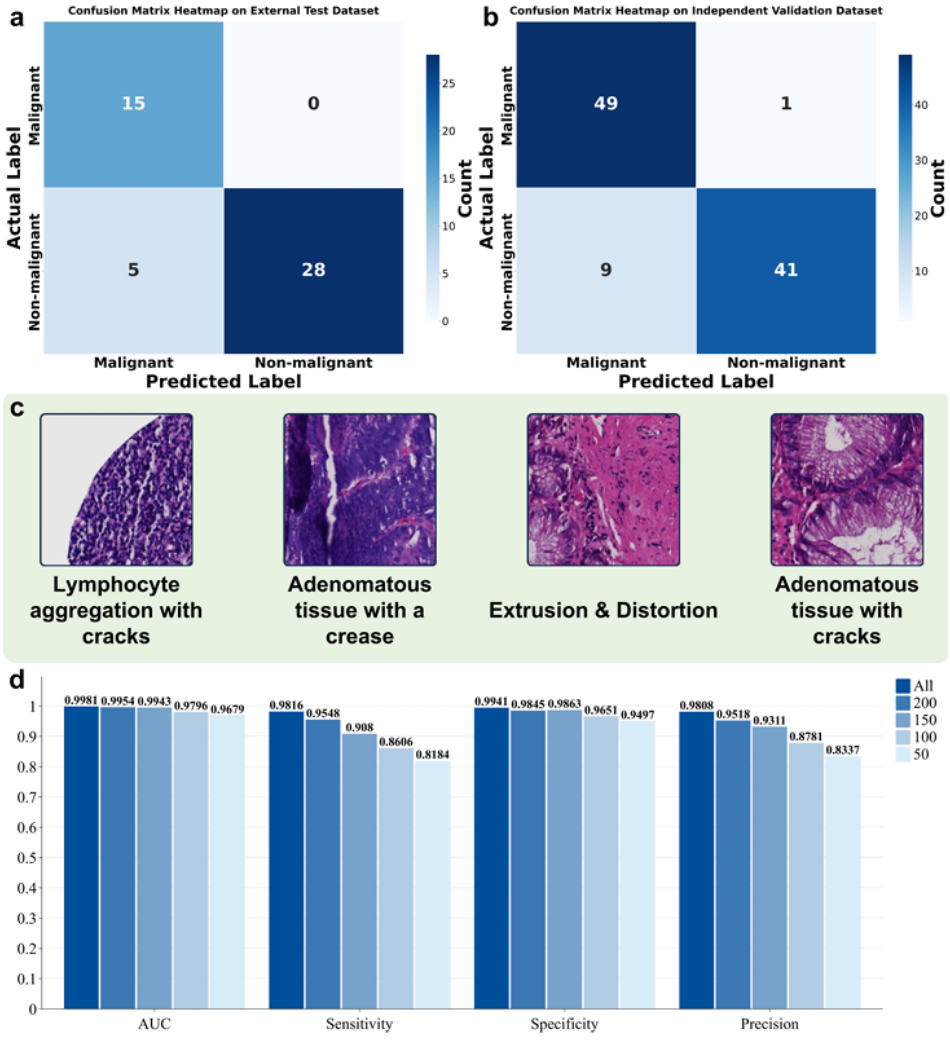

Fig. 5. Results visualization. (a) Confusion matrix for the external test dataset. (b) Confusion matrix for the independent validation dataset. (c) Representative error tiles for the VGA model when tested at the WSI level. (d) Performance of the VGA model on the internal test dataset under different sample-scarcity conditions (All indicates that the VGA model is trained using the full training dataset, 200 indicates that 200 samples of each class in the training dataset are taken in equal quantities to form a new training dataset to train the VGA model, and so on; all the metrics in the graphs are class averages)

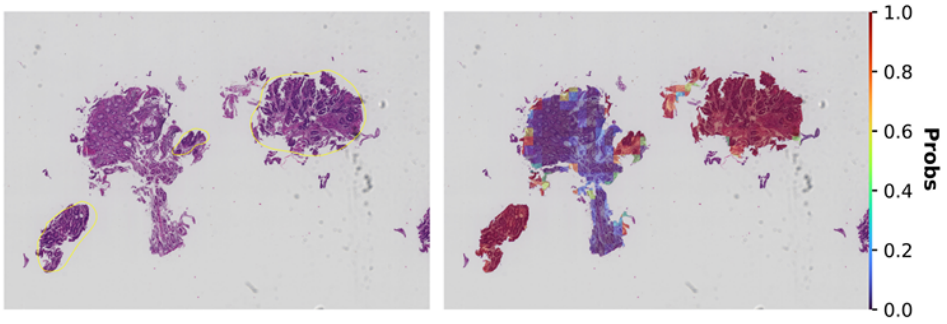

Fig. 6. The heatmap of a slide during WSI-level testing of the VGA model.

to visualize the localization of highly suspicious foci on malignant tumor slides. In the heatmap, warmer colors indicate areas where themodel assigns a higher probability of malignant tumor presence.

In this study, we recorded the error cases during WSI-level testing and further analyzed the frequently occurring misclassifications. In the five slides that were misdiagnosed as malignant, extrusion, distortion, creasing, or cracking of the tissue had a significant impact on the diagnostic results. As shown in Fig. 5c, an area of lymphocyte aggregation accompanied by cracks was misdiagnosed as malignant. Similarly, the adenomatous tissue was also misdiagnosed as malignant due to the presence of an obvious crease. In another case, the morphological deformation of the fibrous mesenchyme after extrusion led to the misdiagnosis. Additionally, the adenomatous tissue was again misdiagnosed as malignant due to the accompanying cracks. The results indicate that the VGA model developed in this study is susceptible to errors when dealing with poor-quality slides or cases involving partially benign tumors and requires adjunctive examination by a pathologist for accurate determination.

3.5. Sample Less Scenario Testing

To further validate the applicability and robustness of the proposed model under low-resource conditions, we conducted sample adaptation evaluation experiments. Specifically, we randomly selected 50, 100, 150, and 200 tiles per category from the internal training dataset to simulate clinical scenarios with varying levels of data availability. The number of WSIs used has also gradually increased from 9 to 20 accordingly. In each setting, we evaluated the classification performance and confusion matrix performance of the model in a three-level organizational activity classification task to analyze the trend of the impact of changes in data volume on model performance. Fig. 5d shows the performance metrics of the VGA algorithm for the internal test dataset for the three organizational activity levels under different sampling conditions. As the number of samples per class increased from 50 to 200, the classification performance of the model also improved accordingly. When using 200 samples per class, the model achieved an average AUC of 0.9954 (95% [CI] 0.9917,0.9991), an average sensitivity of 0.9548 (95% [CI] 0.9436,0.9660), an average specificity of 0.9845 (95% [CI] 0.9778,0.9912), and a mean precision of 0.9518 (95% [CI] 0.9402,0.9634). Combined with Table 5, it can be seen

that the performance of the model on the internal test dataset at this point is close to the results at full training, with small differences observed in the evaluation metrics. Increasing the number of tiles further yields marginal perfomance gains. Therefore, we limit the few-shot evaluation to up to 200 tiles per class.

Beyond the empirical results, it is also important to understand why the model remains effective when the number of training tiles is limited. This can be explained in part by the strong representation reuse enabled by large-scale pretraining. Through the pretrained feature extractor, the encoder retains general visual representations learned from extensive pretraining, which can be effectively reused for downstream classification tasks[36]. Furthermore, freezing the encoder while training only a lightweight classification head reduces the number of trainable parameters, leading to faster training and lower computational cost compared to fine-tuning the entire model. The reduced number of trainable parameters also helps mitigate the risk of overfitting on smaller target datasets[17]. Consequently, when the model is trained using only a small number of image tiles for each category, it can maintain stable performance rather than experiencing a sharp decline.

4. Discussion

Timely diagnosis and screening of gastrointestinal cancers are important to improve patient outcomes and survival. Currently, histopathologic analysis remains the gold standard for the diagnosis of gastrointestinal cancers. However, on the one hand, there is a shortage of pathologists and a long training period for pathologists all over the world, including the United States and low- and middle-income countries [3]. On the other hand, insufficient diagnostic experience can lead to missed diagnoses and misdiagnoses, significantly impacting subsequent treatment [55]. Therefore, these challenges highlight the urgent need for reliable tools to assist in pathological image analysis and gastrointestinal cancers screening, with the goal of enhancing diagnostic efficiency. At the same time, advances in digital pathology have created a practical foundation for the integration and deployment of AI models in clinical diagnostic workflows [28].

WSIs from the stomach and colorectum were collected from different institutions for training, testing, and external validation of a universal screening model for gastrointestinal cancers, referred to as VGA, developed in this study. Unlike previous studies that either trained separate models for stomach and colorectum data using the same network architecture or used a model trained on one organ to validate its generalization on the other, this study simultaneously incorporated data from both organs for model training and performed classification to achieve unified screening.

The model was trained using a simplified three-class system (NT, BT, MT) to facilitate rapid screening. However, this simplification overlooks important clinical distinctions. For example, the NT class includes both normal mucosa and inflammatory conditions, which have different clinical implications, while the MT class groups high-grade adenomas with adenocarcinomas, which would normally require more granular subtyping. This simplified classification is suitable for initial screening, but future work will focus on extending the model or integrating it into workflows that require finer diagnostic resolution to better reflect clinical complexity.

The results showed that the AI screening model developed in this study achieved a class-averaged sensitivity of over 0.9161 and a class-averaged precision of over 0.9179

on both the internal and external test datasets, demonstrating improvements over the other three baseline networks. Meanwhile, in the WSI-level test on the external dataset, the sensitivity of malignant tumor slides reached 100%, which demonstrates the model's potential to assist pathologists in confirming diagnoses or prompting further investigation when discrepancies arise in the initial assessment. The AI screening model developed in this study could potentially be integrated into the digital pathology workflow by automatically predicting the diagnosis and highlighting the suspected malignant areas after slide scanning. This enables pathologists to prioritize suspicious cases, thereby speeding up the diagnosis. Additionally, we observed that occasional misdiagnoses tended to occur when it processed slides of poor quality, such as those with tissue extrusion, distortion, creases, or cracks. Future work could consider marking such slides to prompt pathologists to re-evaluation, rescanning or recutting when necessary. However, it cannot be denied that the performance of the current model relies on high-quality input.

Despite the promising performance of the proposed AI screening model, several limitations and deployment assumptions should be acknowledged. First, the current framework operates at the tile level and is not yet capable of directly interpreting individual ROIs or entire WSIs. Future work will focus on developing effective tile-level feature aggregation methods to enable WSI-level prediction and interpretation. Second, this study was limited to gastrointestinal organs including the stomach and colorectum only, and data from more organs will be included in the future to construct a generalized screening model for gastrointestinal cancers. Third, the dataset used was derived from only three medical centers, and further validation on more diverse datasets is required to assess robustness and reduce potential dataset bias.

From a deployment perspective, several practical considerations also exist. The model assumes access to high-quality WSIs; variations in staining protocols, scanning resolution, or slide preparation may affect prediction accuracy. Furthermore, real-time integration into digital pathology workflows may face computational constraints, such as inference speed and hardware requirements, which could affect scalability in high-throughput settings. Finally, the current framework relies on frozen encoders with pretrained representations, which may limit adaptability to rare or out-of-distribution cases. Addressing these limitations through broader multi-center datasets, improved WSI-level modeling, prospective clinical validation and optimization for real-time clinical deployment will be important directions for future work.

5. Conclusion

In summary, previous AI-assisted systems have tended to be organ-specific tasks or for pan-cancer detection. We have developed a unified screening model for gastrointestinal cancers named VGA, which is capable of reliably classifying WSIs from the stomach and colorectum in multiple categories. With the help of the developed AI system, pathologists can perform more rapid screening of gastrointestinal cancers and lay the foundation for subsequent treatment.

Code Availability. The code are available from the corresponding author upon reasonable request.

References

1. Barmpoutis, P., Waddingham, W., Yuan, J., Ross, C., Kayhanian, H., Stathaki, T., Alexander, D.C., Jansen, M.: A digital pathology workflow for the segmentation and classification of gastric glands: Study of gastric atrophy and intestinal metaplasia cases. Plos one 17(12), e0275232 (2022)
2. Bilal, M., Tsang, Y.W., Ali, M., Graham, S., Hero, E., Wahab, N., Dodd, K., Sahota, H., Wu, S., Lu, W., et al.: Development and validation of artificial intelligence-based prescreening of large-bowel biopsies taken in the uk and portugal: a retrospective cohort study. The Lancet Digital Health 5(11), e786–e797 (2023)
3. Black-Schaffer, W.S., Morrow, J.S., Prystowsky, M.B., Steinberg, J.J.: Training pathology residents to practice 21st century medicine: a proposal. Academic pathology 3, 2374289516665393 (2016)
4. Cai, C., Shi, Q., Li, J., Jiao, Y., Xu, A., Zhou, Y., Wang, X., Peng, C., Zhang, X., Cui, X., et al.: Pathologist-level diagnosis of ulcerative colitis inflammatory activity level using an automated histological grading method. International Journal of Medical Informatics 192, 105648 (2024)
5. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine 25(8), 1301–1309 (2019)
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine 30(3), 850–862 (2024)
7. Cheung, J., Savine, S., Nguyen, C., Lu, L., Yasin, A.S.: Transfer learning from one cancer to another via deep learning domain adaptation (2026)
8. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://aclanthology.org/D14-1179/
9. Choi, S., Kim, S.: Artificial intelligence in the pathology of gastric cancer. Journal of Gastric Cancer 23(3), 410 (2023)
10. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine 24(10), 1559–1567 (2018)
11. Da, Q., Wang, S., Wang, W., Yang, C., Wang, B., Ruan, M., Fu, Z., Xu, Y., Zhou, Y., Wang, C., et al.: Progress and challenges of pathological artificial intelligence in the era of large models. Zhonghua bing li xue za zhi= Chinese journal of pathology 54(3), 305–309 (2025)
12. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., Williamson, D.F.K., Robertson, H., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Chen, C.S., Komura, D., Kawabe, A., Ochi, M., Sato, S., Yokose, T., Miyagi, Y., Ishikawa, S., Gerber, G., Peng, T., Le, L.P., Mahmood, F.: A multimodal whole-slide foundation model for pathology. Nature Medicine 31(11), 3749–3761 (Nov 2025)
13. Du, Y., Liu, X., Yue, L., Feng, L., Tao, P., Jing, Q.: Minidigpath: A new standard for pathology images few-shot learning classification. In: 2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE). pp. 136–140 (2023)
14. Fu, B., Zhang, M., He, J., Cao, Y., Guo, Y., Wang, R.: Stohisnet: A hybrid multi-classification model with cnn and transformer for gastric pathology images. Computer Methods and Programs in Biomedicine 221, 106924 (2022)

15. Griem, J., Eich, M.L., Schallenberg, S., Pryalukhin, A., Bychkov, A., Fukuoka, J., Zayats, V., Hulla, W., Munkhdelger, J., Seper, A., et al.: Artificial intelligence–based tool for tumor detection and quantitative tissue analysis in colorectal specimens. Modern Pathology 36(12), 100327 (2023)
16. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence 45(1), 87–110 (2022)
17. Hasan, K.R., Kim, S., Cho, J., Han, H.S.: Prototypical few-shot learning for histopathology classification: Leveraging foundation models with adapter architectures. IEEE ACCESS 13, 86356–86379 (2025)
18. Hinata, M., Ushiku, T.: Detecting immunotherapy-sensitive subtype in gastric cancer using histologic image-based deep learning. Scientific reports 11(1), 22636 (2021)
19. Huang, B., Tian, S., Zhan, N., Ma, J., Huang, Z., Zhang, C., Zhang, H., Ming, F., Liao, F., Ji, M., et al.: Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: A retrospective multicentre study. EBioMedicine 73 (2021)
20. Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., Tsuneki, M.: Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Scientific reports 10(1), 1504 (2020)
21. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal 16, 34–42 (2018)
22. Korbar, B., Olofson, A.M., Miraflor, A.P., Nicka, C.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A., Hassanpour, S.: Deep learning for classification of colorectal polyps on whole-slide images. Journal of pathology informatics 8, 30 (2017)
23. Lan, J., Chen, M., Wang, J., Du, M., Wu, Z., Zhang, H., Xue, Y., Wang, T., Chen, L., Xu, C., et al.: Using less annotation workload to establish a pathological auxiliary diagnosis system for gastric cancer. Cell Reports Medicine 4(4) (2023)
24. Le Page, A.L., Ballot, E., Truntzer, C., Derangère, V., Ilie, A., Rageot, D., Bibeau, F., Ghiringhelli, F.: Using a convolutional neural network for classification of squamous and non-squamous non-small cell lung cancer based on diagnostic histopathology hes images. Scientific Reports 11(1), 23912 (2021)
25. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A visual-language foundation model for computational pathology. Nature Medicine 30(3), 863–874 (Mar 2024)
26. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering 5(6), 555–570 (2021)
27. Ma, M., Zeng, X., Qu, L., Sheng, X., Ren, H., Chen, W., Li, B., You, Q., Xiao, L., Wang, Y., et al.: Advancing automatic gastritis diagnosis: An interpretable multilabel deep learning framework for the simultaneous assessment of multiple indicators. The American Journal of Pathology 194(8), 1538–1549 (2024)
28. Moxley-Wyles, B., Colling, R.: Artificial intelligence and digital pathology: where are we now and what are the implementation barriers? Diagnostic Histopathology (2024)
29. Mudeng, V., Farid, M.N., Ayana, G., Choe, S.w.: Domain and histopathology adaptations–based classification for malignancy grading system. The American Journal of Pathology 193(12), 2080–2098 (2023)
30. Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., Cree, I.A., et al.: The 2019 who classification of tumours of the digestive system. Histopathology 76(2), 182 (2019)
31. Neidlinger, P., El Nahhas, O.S.M., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., Rocken, C., Foersch, S., Truhn, D., Marra, A., Saldanha, O.L., Kather, J.N.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature Biomedical Engineering (Oct 2025)

32. Oh, Y., Bae, G.E., Kim, K.H., Yeo, M.K., Ye, J.C.: Multi-scale hybrid vision transformer for learning gastric histology: Ai-based decision support system for gastric cancer treatment. IEEE journal of biomedical and health informatics 27(8), 4143–4153 (2023)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024)
34. Park, S.y., Ayana, G., Wako, B.D., Jeong, K.C., Yoon, S.D., Choe, S.w.: Vision transformers for low-quality histopathological images: A case study on squamous cell carcinoma margin classification. Diagnostics 15(3), 260 (2025)
35. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830 (2011)
36. Rahman, T., Baras, A.S., Chellappa, R.: Evaluation of a task-specific self-supervised learning framework in digital pathology relative to transfer learning approaches and existing foundation models. Modern Pathology 38(1), 100636 (2025)
37. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0
38. Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Comput. Surv. 55(13s) (Jul 2023)
39. Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., et al.: Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nature communications 11(1), 4294 (2020)
40. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis 67, 101813 (2021)
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
42. Tsuneki, M., Kanavati, F.: Weakly supervised learning for poorly differentiated adenocarcinoma classification in gastricendoscopic submucosal dissection whole slide images. Technology in Cancer Research & Treatment 21, 15330338221142674 (2022)
43. Tung, C.L., Chang, H.C., Yang, B.Z., Hou, K.J., Tsai, H.H., Tsai, C.Y., Yu, P.T.: Identifying pathological slices of gastric cancer via deep learning. Journal of the Formosan Medical Association 121(12), 2457–2464 (2022)
44. Veldhuizen, G.P., Röcken, C., Behrens, H.M., Cifci, D., Muti, H.S., Yoshikawa, T., Arai, T., Oshima, T., Tan, P., Ebert, M.P., et al.: Deep learning-based subtyping of gastric cancer histology predicts clinical outcome: a multi-institutional retrospective study. Gastric Cancer 26(5), 708–720 (2023)
45. Vinay Kumar, Abul K˙ Abbas, D.J., Aster, J.C.: Robbins & Cotran Pathologic Basis of Disease. Elsevier, Illinois, USA (2020)
46. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine 30(10), 2924–2935 (2024)
47. Wang, J., Liu, X.: Medical image recognition and segmentation of pathological slices of gastric cancer based on deeplab v3+ neural network. Computer methods and programs in biomedicine 207, 106210 (2021)
48. Wang, K.S., Yu, G., Xu, C., Meng, X.H., Zhou, J., Zheng, C., Deng, Z., Shang, L., Liu, R., Su, S., et al.: Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. BMC medicine 19, 1–12 (2021)

49. Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.A.: Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification. Medical image analysis 58, 101549 (2019)
50. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634(8035), 970–978 (2024)
51. Wang, Z., Peng, H., Wan, J., Song, A.: Identification of histopathological classification and establishment of prognostic indicators of gastric adenocarcinoma based on deep learning algorithm. Medical molecular morphology pp. 1–13 (2024)
52. Wei, J.W., Suriawinata, A.A., Vaickus, L.J., Ren, B., Liu, X., Lisovsky, M., Tomita, N., Abdollahi, B., Kim, A.S., Snover, D.C., et al.: Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA network open 3(4), e203398–e203398 (2020)
53. Xie, Y., Shi, L., He, X., Luo, Y.: Gastrointestinal cancers in china, the usa, and europe. Gastroenterology report 9(2), 91–104 (2021)
54. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature 630(8015), 181–188 (2024)
55. Yang, Z., Wei, T., Liang, Y., Yuan, X., Gao, R., Xia, Y., Zhou, J., Zhang, Y., Yu, Z.: A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. Nature Communications 16(1), 2366 (2025)
56. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)

Lijue Liu is a associate professor at the School of Automation, Central South University, with research interests in image processing, data mining, and intelligent information processing.

Fangjie Yin is a master's candidate at the School of Automation, Central South University, with research interests in image processing.

Genjian Yang is an Engineer at China Electronic Product Reliability and Environmental Testing Research Institute. His research interests include multimodal algorithms and game learning.

Qi Li is an attending physician in the Department of Pathology, Beijing Integrated Traditional Chinese and Western Medicine Hospital. Her main research focuses on the pathological diagnosis of digestive tract, thyroid, gynecological and other diseases.

Siya Li is a researcher at CAS Blue Bay Cloud Technology (Guangdong) Co., Ltd focusing on artificial intelligence and medical research and development, with interests in AI-driven medical image analysis and AI applications in precision medicine.

Teng Pan has dedicated her master's, doctoral, and postdoctoral research to investigating the tumor microenvironment and angiogenesis in breast cancer, with extensive research experience in breast cancer biology. She obtained her PhD in Oncology from Tianjin

Medical University. Her research work has been published in several journals, including Cancer Letters, British Journal of Cancer, and Theranostics.

Ting Liu is employed at Beijing Ditan Hospital, with research interests including data mining and hepatocellular carcinoma-related studies.

Jin Tang is a professor at the School of Automation, Central South University, with research interests in computer vision, industrial intelligence, and application of large language models.

Ruijie Ming is a physician in the Department of Oncology at Chongqing University Three Gorges Hospital, with research interests in tumor therapy resistance, the tumor microenvironment, and tumor multi-omics analysis.

Yu Song is an associate chief physician and deputy director in the Department of Otolaryngology Head and Neck Surgery, Peking University First Hospital, with research interests in surgical treatment of allergic rhinitis and diagnosis and treatment of rhinobasal diseases.

Xue Feng is an attending physician in the Department of Respiratory and Critical Care Medicine, Tianjin Chest Hospital, with research interests in respiratory critical care medicine and respiratory central regulation.

Dan Wang is a postdoc at Richard Dimbleby Laboratory of Cancer Research, Randall Division and Division of Cancer and Pharmaceutical Sciences, King's College London, with interests in the application of AI in medicine, tumor microenvironment.

Xingang Zhou is a chief physician in the Department of Pathology, Beijing Ditan Hospital, Capital Medical University, specializing in digital and artificial intelligence pathology as well as liver pathology.

Wenbai Chen is a professor at the Beijing Information Technology Science and University, with research interests in include multimodal algorithms and pattern recognition.

Jinhai Deng specialized in the tumor microenvironment throughout his master's and doctoral studies, and possesses extensive research expertise in tumor biology and tumor immunology. He obtained his PhD from King's College London, UK. His research findings have been published in journals including EMBO Molecular Medicine, British Journal of Cancer, Theranostics and so on.

Received: November 30, 2025; Accepted: March 30, 2026.