scientific reports

Check for updates

OPEN

A multi-task learning model for evaluating non-tumor gastric diseases indicators in whole slide images

Mingxi Fu[1,5], Liming Liu[2,5], Fanglei Fu[1], Jingli Ouyang[1], Xueying Shi[3], Song Duan[4], Tian Guan[1], Yonghong He[1]✉, Zhiqiang Cheng[2]✉ & Lianghui Zhu[1]✉

Inflammation, acute activity, intestinal metaplasia, and atrophy are key indicators in gastric biopsy evaluations, with their grading being crucial for assessing gastric cancer progression. However, diagnostic subjectivity among pathologists and the complex interrelationships between these indicators present significant challenges. Additionally, the high resolution of whole slide images (WSIs) complicates large-scale annotation efforts. To address these issues, we propose a multi-task learning model utilizing self-supervised pre-trained weights from extensive pathological datasets. The model integrates four indicators–severity of inflammation, atrophy, acute activity, and intestinal metaplasia–training on WSIs to predict these indicators while accounting for their interrelationships. Our results show that multi-task learning outperforms single-task models, achieving higher accuracy across all indicators. This model can thus serve as an auxiliary tool for evaluating non-tumor gastric diseases and supporting diagnosis.

Gastric cancer is the fifth most common malignancy worldwide, ranking fifth in incidence and third in mortality globally [1]. According to the classical Correa cascade, gastric cancer follows a multi-stage progression: chronic gastritis → atrophic gastritis → intestinal metaplasia → dysplasia → carcinoma [2]. Helicobacter pylori infection is identified as a major driving factor [3]. Atrophic gastritis and intestinal metaplasia are considered key precancerous lesions, necessitating comprehensive pathological evaluation that includes inflammation, atrophy, acute activity, and intestinal metaplasia, which demonstrate significant clinical utility in the comprehensive assessment of non-tumor gastric diseases [4].

Pathological diagnosis remains the gold standard for gastric disease assessment[5]. Traditionally, pathologists examine tissue sections under a microscope in a time-consuming process that can take 15–30 minutes per case. Since the 1994 update of the Sydney System[6], grading of gastritis indicators has been introduced, yet a lack of quantitative assessment methods has led to low diagnostic consistency[7]. Advances in digital pathology and artificial intelligence offer solutions by enabling whole-slide image (WSI) digitization, facilitating the application of deep learning models to automate and standardize gastric disease assessment[8]. Deep learning models have demonstrated strong performance in various digital pathology tasks, including tumor detection[9], classification[10], grading[11], segmentation[12], and prognosis prediction[13]. However, WSIs' gigapixel resolution poses annotation challenges, leading researchers to adopt a weakly supervised approach known as multiple instance learning (MIL)[14–16]. MIL processes WSIs as collections of image patches (instances), allowing model training with only slide-level labels. The standard MIL framework comprises three key steps: instance feature extraction, feature aggregation, and slide-level classification[17].

Despite the effectiveness of MIL, feature extractors pre-trained on natural images lack adaptation to pathology-specific patterns, resulting in suboptimal performance[18]. Self-supervised learning (SSL) has emerged as a solution by pre-training models on large-scale unlabeled WSI datasets to improve pathology-specific feature representation[19]. Wang et al. proposed CTransPath, a hybrid CNN-Transformer model that introduced

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. [2]Department of Pathology, The Second Affiliated Hospital of Southern University of Science and Technology (Shenzhen Third People's Hospital), Shenzhen 518000, China. [3]Department of Pathology, School of Basic Medical Sciences, Peking University Third Hospital, Peking University Health Science Center, Beijing 100083, China. [4]Department of Pathology, Chongqing University Affiliated Three Gorges Hospital, Chongqing 404000, China. [5]Mingxi Fu and Liming Liu are co-first authors. ✉email: heyh@sz.tsinghua.edu.cn; chengzhiqiang2004@aliyun.com; zhulh@mails.tsinghua.edu.cn

the MoCo v3 contrastive learning framework into pathology, pre-training on 100,000 WSIs. However, its dataset covers only 20 tissue types[20]. As dataset scales expanded, the UNI model leveraged the DINO v2 algorithm with the ViT-L architecture, but it was trained on a large internal private dataset[21]. Addressing task diversity, the visual-language foundation model Conch utilized CoCa[22], pre-training on pathology images from multiple sources alongside biomedical text, incorporating image and text encoders with a multimodal fusion decoder[23]. Prov-GigaPath employed a ViT-g architecture (1.1 billion parameters) and LongNet sparse attention mechanisms, achieving whole-slide self-supervised learning on 170,000 WSIs, effectively mitigating Transformer's computational complexity growth but remaining constrained in cross-scale feature fusion across multiple magnifications[24]. Recent studies indicate a superlinear relationship between model performance and dataset size[25]. The Virchow model explored scaling across data and model dimensions, incorporating domain-informed training modifications[26]. Trained on 3.1 million WSIs, it utilized a ViT-G (1.9 billion parameters) and introduced KDE regularization and ECT data augmentation, effectively alleviating feature collapse while enhancing adaptation to different staining techniques and magnifications[27]. This model set new benchmarks in pathology AI, providing a novel paradigm for precise screening.

In the assessment of non-tumor gastric diseases, significant progress has been made in recent studies. Fang et al. employed a multiple instance learning (MIL) approach for the diagnosis and grading of intestinal metaplasia and gastric atrophy, enhancing pathologists' diagnostic accuracy[28]. Lai et al. proposed a generalized network integrating multi-scale, multi-instance, and multi-feature joint learning for the classification of intestinal metaplasia subtypes[29]. By leveraging multi-class MIL prediction and multi-feature joint learning networks, their method effectively captured inter-instance correlations and multi-scale information, thereby improving classification performance. Ma et al. developed a classification model to differentiate normal mucosa, chronic gastritis, and intestinal-type gastric carcinoma, revealing the morphological evolution of gastric mucosal lesions through feature learning[30].

Despite these advances, most deep learning techniques focus on the independent evaluation of individual indicators while often neglecting the biological correlations among multi-stage lesions as described in Correa's cascade theory. Multi-task learning (MTL) offers a new paradigm to address this challenge by leveraging shared information across multiple learning tasks to enhance their respective performance[31]. Fu et al. developed a multi-task machine learning model to predict peritoneal recurrence and disease-free survival (DFS) in gastric cancer patients, further exploring the association between these predictions and adjuvant chemotherapy[32]. Similarly, Hao et al. introduced a multi-task learning strategy that integrates abdominal CT imaging with pathological staging markers to improve gastric cancer survival prediction, demonstrating superior performance compared to conventional methods[22]. With the advancement of deep learning techniques, multi-task learning has increasingly been integrated into pathology image analysis. Multi-task deep learning frameworks typically consist of a shared feature extractor and task-specific branches to optimize the performance of multiple related tasks[33]. Hu et al. proposed a multi-task deep learning framework incorporating gastric cancer segmentation, nerve detection, and perineural invasion (PNI) decision-making to highlight diagnostically relevant PNI regions in WSIs[34]. Furthermore, Zhou et al. introduced PathM3, a multi-modal, multi-task, and multi-instance learning framework designed for WSI classification and descriptive report generation, demonstrating improvements in classification accuracy and automated pathology report generation[35].

Current studies on gastric disease assessment primarily focus on individual indicators, overlooking the interdependencies among inflammation, atrophy, acute activity, and intestinal metaplasia. Multi-task learning (MTL) provides a promising paradigm for simultaneously modeling related tasks to enhance prediction accuracy. In this study, we propose a multi-task learning framework for the assessment of non-tumor gastric diseases based on feature extraction from pathology foundation models. Our contributions can be summarized as follows:

- We construct a multi-task learning framework that jointly models the pathological correlations among the severity of inflammation, acute activity, atrophy, and intestinal metaplasia. The proposed multi-task model outperforms single-task models in predicting non-tumor gastric indicators.
- We incorporate the Virchow model, which has been pre-trained on a large-scale pathology image dataset using self-supervised learning, for feature extraction in gastric pathology images. This effectively addresses the limitations of feature extractors pre-trained on natural images, significantly enhancing the accuracy of downstream tasks.

Materials and methods

Dataset

This study is based on the Whole Slide Image (WSI) dataset from the Department of Pathology at Shenzhen Third People's Hospital, used to develop a multi-task classification system for gastritis. All images were derived from tissue samples from different patients, processed using hematoxylin and eosin (H&E) staining, and digitized at high resolution using the SQS-100 digital pathology scanning system (20× objective) developed by Shenzhen Shengqiang Technology Co., Ltd. To ensure patient privacy compliance, all cases underwent rigorous de-identification, guaranteeing the complete anonymization of personal health information. During the data preprocessing phase, the Otsu algorithm[36] based on an HSV thresholding method is employed to extract foreground tissue regions, effectively removing background noise and non-target areas. Subsequently, while preserving the original anatomical structure, tissue samples scanned at 20× magnification are segmented into non-overlapping, standardized image patches of 256×256 pixels. Data augmentation techniques are further applied to enhance dataset diversity.

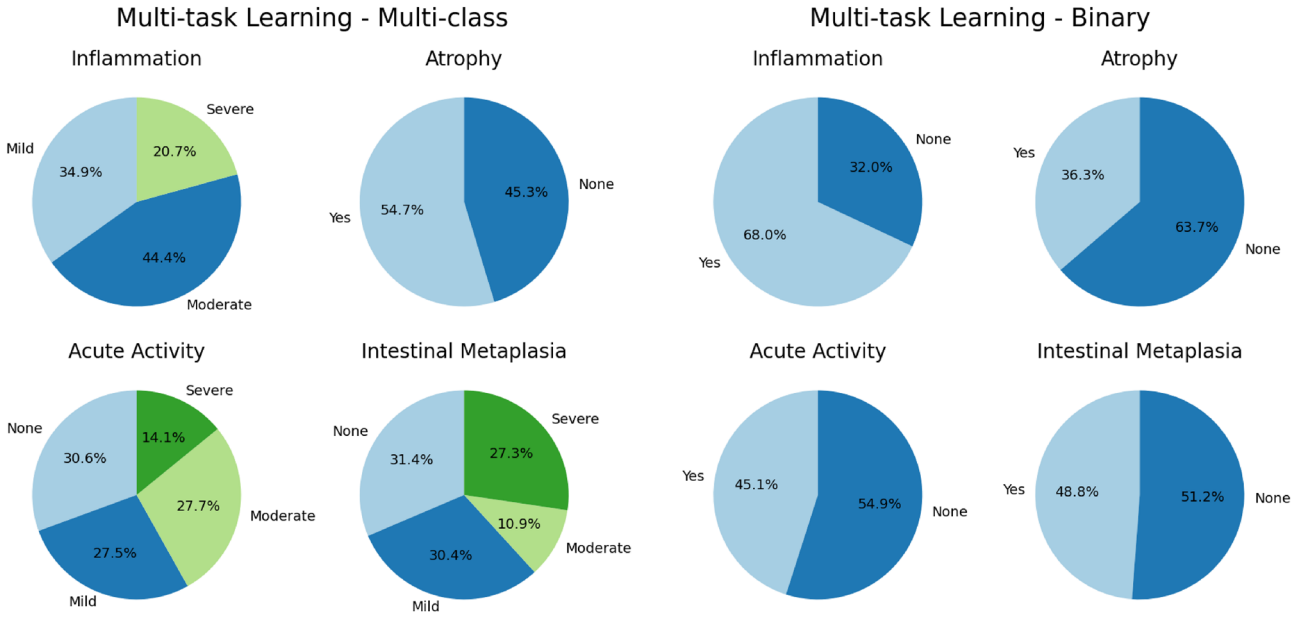


Fig. 1. Distribution of WSIs for **A** multi-class and **B** binary-class classification tasks.

Variable	χ^2 (df)	p-value
Inflammation vs. Atrophy	45.30 (2)	< 0.001
Inflammation vs. Acute Activity	89.14 (4)	< 0.001
Inflammation vs. Intestinal Metaplasia	67.85 (6)	< 0.001
Acute activity vs. Inflammation	32.50 (2)	< 0.001
Atrophy vs. Intestinal metaplasia	58.20 (3)	< 0.001
Acute activity vs. Intestinal metaplasia	75.60 (6)	< 0.001

Table 1. Chi-Square Test Among the Four Indicators.

The proposed multi-task classification framework covers four key non-tumor gastric diseases pathology tasks: inflammation grading, atrophy assessment, acute activity grading, and intestinal metaplasia grading. The dataset consists of 984 WSIs, distributed as follows in Fig. 1: inflammation grading (343 mild, 437 moderate, 204 severe), atrophy assessment (538 atrophic, 446 non-atrophic), acute activity grading (301 none, 271 mild, 273 moderate, 139 severe), and intestinal metaplasia grading (309 none, 299 mild, 107 moderate, 269 severe). To evaluate the clinical applicability of the model, four additional binary classification tasks are conducted: binary of inflammation, binary of atrophy, binary of acute activity, and binary of intestinal metaplasia. The dataset comprises 1094 WSIs, with positive and negative sample distributions as follows: inflammation (744 yes / 350 none), atrophy (397 yes / 697 none), acute activity (493 yes / 601 none), and intestinal metaplasia (534 yes / 560 none).

In this study, we analyze the relationships among inflammation, atrophy, acute activity, and intestinal metaplasia using the chi-square (χ^2) test. Table 1 presents the chi-square statistics (χ^2), degrees of freedom (df), and p-values for each pairwise combination of these variables. All variable combinations exhibit statistically significant differences, with p values less than 0.001, indicating strong associations among the examined indicators. These results suggest potential interactions or shared biological mechanisms underlying these non-tumor gastric pathological features.

Model

This study proposes a weakly supervised multiple instance learning (MIL) framework that leverages whole-slide image (WSI) bag-level labels (i.e., diagnostic outcomes for the entire slide) in conjunction with MIL techniques to achieve multi-task joint prediction of the severity of inflammation, atrophy, acute activity, and intestinal metaplasia in Fig. 2. This approach effectively addresses the challenges posed by the enormous size of WSIs–each of which may reach gigapixel resolution and the high cost of manual annotation, enabling the retention of global pathological information without the need for manually annotating regions of interest (ROIs).

In the MIL framework, a WSI is treated as a 'bag' comprising multiple image patches, with each patch considered an 'instance'[14]. Instead of relying on manually selected ROIs, we directly utilize WSI-level labels to

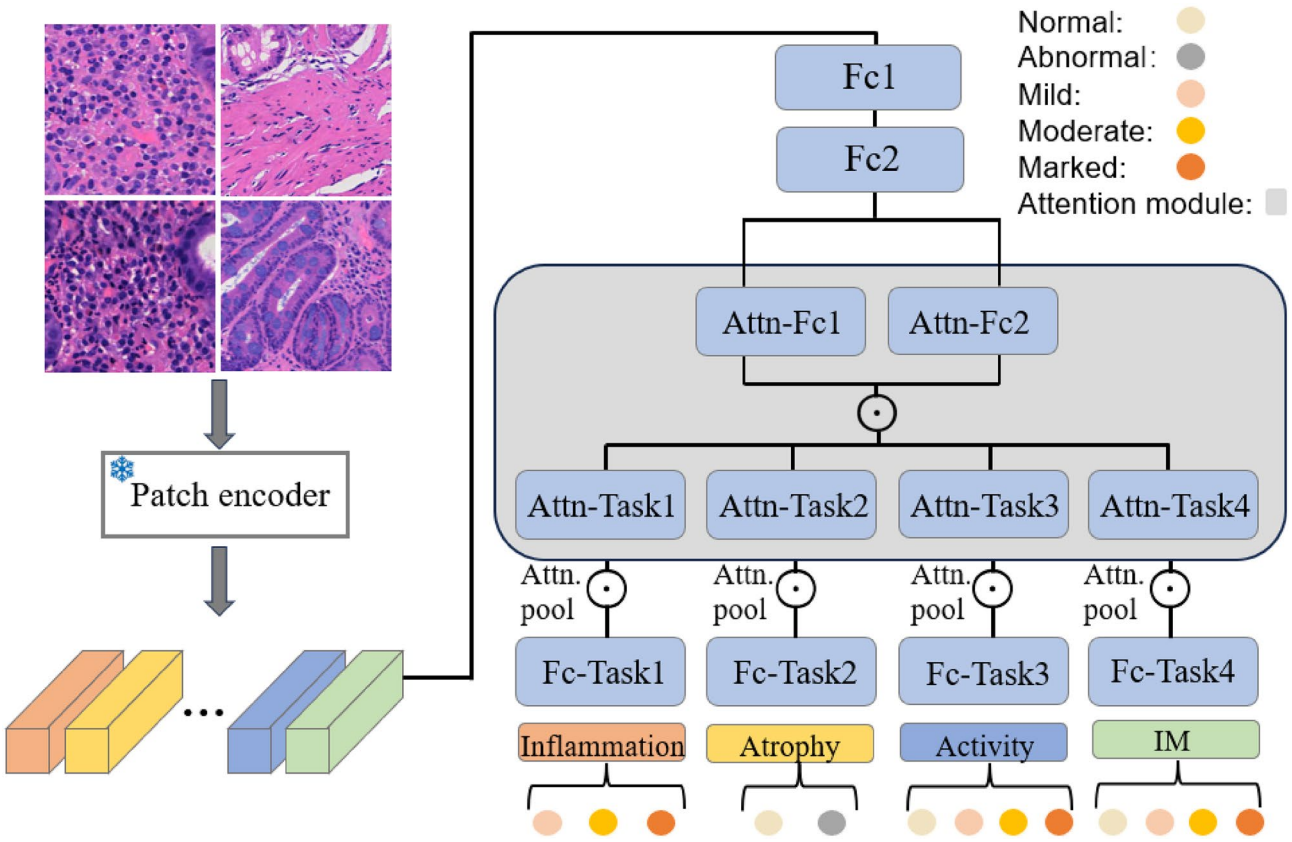


Fig. 2. Multi-task learning model framework based on the ABMIL architecture.

train a multi-task network model. This ensures that information from the entire WSI is incorporated into the learning process, thereby enhancing predictive accuracy and improving generalizability across multiple gastric pathology indicators.

Feature extraction

We employ the Virchow model, a Vision Transformer-Huge (ViT-H/14) architecture with 632 million parameters, based on DINOv2 self-supervised learning for feature extraction[27]. This model has been trained on 1.5 million H&E pathology slides using a multi-magnification strategy ($5\times/10\times/20\times/40\times$), significantly enhancing its generalization capability across diverse tissue types and staining conditions. After freezing the feature extractor, input patches are encoded into 2560-dimensional feature vectors, capturing high-level pathological semantic information. Subsequently, in the feature space, an attention-based pooling mechanism is extended across multiple tasks to aggregate information from all tissue regions within each slide. Based on these tasks, the classification layer of the network ultimately produces the final WSI-level prediction. Specifically, after feature extraction, the model incorporates two fully connected layers, *fc*1 and *fc*2, with weight and bias parameters defined as follows: $W_1 \in \mathbb{R}^{2560\times512}$, $b_1 \in \mathbb{R}^{512}$, $W_2 \in \mathbb{R}^{512\times512}$, and $b_2 \in \mathbb{R}^{512}$. Each fully connected layer is followed by an activation function, allowing the model to learn deep feature representations of pathological images. The input to this module consists of a patch-level feature embedding $z_k \in \mathbb{R}^{2560}$. The output representation h_k is computed as:

$$h_k = \mathrm{ReLU}\left(W_2\left(\mathrm{ReLU}(W_1 z_k + b_1)\right) + b_2\right). \tag{1}$$

Multi-task attention pooling

Traditional pooling methods often suffer from non-trainable limitations, but attention mechanisms assign trainable weighting coefficients to each feature vector. The attention-based pooling function is formulated as follows:

$$y_{\mathrm{attn}} = \sum_{p=1}^{P} a_p \cdot v_{pf}, \quad f \in \{1, \ldots, F\}, \tag{2}$$

The computation of a_p is defined as follows:

$$a_p = \frac{e^{w^\top \tanh(W_1 v_p^T) \odot \sigma(W_2 v_p^T)}}{\sum_{i=1}^{P} e^{w^\top \tanh(W_1 v_i^T) \odot \sigma(W_2 v_i^T)}}. \tag{3}$$

Let X and Y denote the spatial dimensions of each patch, P represents the number of patches, and F defines the feature dimension. The tuple $(v_1, v_2, \ldots, v_P)$ corresponds to the feature vectors of individual patches, where v_p represents the column vector of the pth patch, and v_{pf} denotes the fth feature value of the pth patch. The trainable vector w has a length of F, and the matrix $W_1 \in \mathbb{R}^{F \times P}$ consists of trainable parameters. To enhance non-linearity, a_p incorporates the non-linear functions $\tanh$ and σ (sigmoid).

In multi-task learning, each task t contains an attention layer module. This module is composed of two fully connected layers $Attn - Fc1$ and $Attn - Fc2$, with weight parameters $W_1 \in \mathbb{R}^{384\times512}$ and $W_2 \in \mathbb{R}^{384\times512}$, which are shared across all tasks and used to perform non-linear transformations of patch features. Subsequently, each task has its own independent weight parameters $w_{a,t} \in \mathbb{R}^{512\times384}$, which are used to generate task-specific attention weights. During training, each patch is assigned an attention score $a_{k,t}$ under different tasks, indicating the information content and importance of that patch for the corresponding classification task. The specific calculation is defined as follows[37]:

$$a_{k,t} = \frac{e^{w_{a,t}^\top \tanh(W_1 v_p^T) \odot \sigma(W_2 v_p^T)}}{\sum_{i=1}^{P} e^{w_{a,t}^\top \tanh(W_1 v_i^T) \odot \sigma(W_2 v_i^T)}}. \tag{4}$$

Classification

For each task, we employ different classification layers to generate task-specific predictions:

$$p_t = \mathrm{softmax}(W_{\mathrm{cls},t} v_{p,t} + b_{\mathrm{cls},t}), \tag{5}$$

where $W_{\mathrm{cls},t}$ represents the classification layer parameters for task t, and $b_{\mathrm{cls},t}$ denotes the bias parameters for the respective task.

In our study, we define four specific classification tasks: the first task is a three-class classification for inflammation severity (mild, moderate, and severe), the second task is a binary classification for atrophy presence (atrophic and non-atrophic), the third task is a four-class classification for acute activity grading (none, mild, moderate, and severe), and the fourth task is a four-class classification for intestinal metaplasia grading (none, mild, moderate, and severe). The classification loss is computed using a cross-entropy loss function across all tasks:

$$L = -\sum_{t=1}^{4} \sum_{c=1}^{C_t} y_{ic} \log(p_{ic}), \tag{6}$$

where C_t represents the number of classes for task t, y_{ic} denotes the ground-truth label indicator, and p_{ic} is the predicted probability output by the model.

In the single-task comparison, we adopt the standard Gated Attention-Based Multiple Instance Learning (Gated ABMIL) architecture to aggregate patch-level features into WSI-level representations[14]. Specifically, a fully connected layer is first used to map the input feature dimension to a unified size of 512, followed by a hidden layer with 384 dimensions for modeling. During the input mapping process, a dropout rate of 0.25 is applied as a regularization technique. Both the input mapping and hidden layer mapping employ the *ReLU* activation function for non-linear transformation.

All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by the Ethics Committee of Shenzhen Third People's Hospital. Informed consent was obtained from all subjects and/or their legal guardians.

Experiments

The model training process employs the Adaptive Moment Estimation (Adam) optimizer, with an initial learning rate set to 0.001 and a weight decay coefficient of 1×10^{-5} to mitigate overfitting. The training is conducted for a total of 100 epochs, with an early stopping strategy dynamically optimizing model performance. Specifically, training is halted if the cross-entropy loss does not decrease for 50 consecutive epochs (patience = 50). To ensure statistical balance in data partitioning, we adopt a stratified five-fold cross-validation scheme. For external validation, the original dataset is split into a 4:1 ratio for training and testing. Within the training set, an additional 7:1 split is used for training and validation, ensuring consistent data distribution across folds. For model evaluation, we employ the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) to quantify diagnostic performance. Additionally, accuracy, sensitivity, and specificity are computed as comprehensive evaluation metrics. All statistical analyses are implemented using the scikit-learn library in Python.

Multi-class classification results

Table 2 presents the classification performance of the Virchow feature extraction model across four multi-class classification tasks–inflammation grading (GC), atrophy assessment (GA), acute activity grading (AC), and intestinal metaplasia grading (IM)–under both single-task learning (STL) and multi-task learning (MTL) frameworks. The reported metrics include accuracy, AUC, and F1-score, with a comparative analysis between

Encoder	Task	Index	Accuracy	AUC	F-score
Virchow	Single	GC	0.7246 (0.0324)	0.9027 (0.0123)	0.7220 (0.0296)
		GA	0.7846 (0.0183)	0.8578 (0.0251)	**0.7842 (0.0183)**
		AC	0.6016 (0.0607)	0.8611 (0.0349)	0.5891 (0.0729)
		IM	0.7856 (0.0679)	0.9446 (0.0397)	0.7653 (0.0663)
	Multi-task	GC	**0.7398 (0.0111)**	**0.9100 (0.0047)**	**0.7367 (0.0107)**
		GA	**0.8018 (0.0073)**	**0.8807 (0.0069)**	0.7636 (0.0369)
		AC	**0.6625 (0.0249)**	**0.8935 (0.0081)**	**0.6883 (0.0414)**
		IM	**0.8385 (0.0100)**	**0.9671 (0.0028)**	**0.8173 (0.0164)**

Table 2. Comparison of STL and MTL Performance—multi-class classification. Bold values indicate the largest metric value within the same index task.

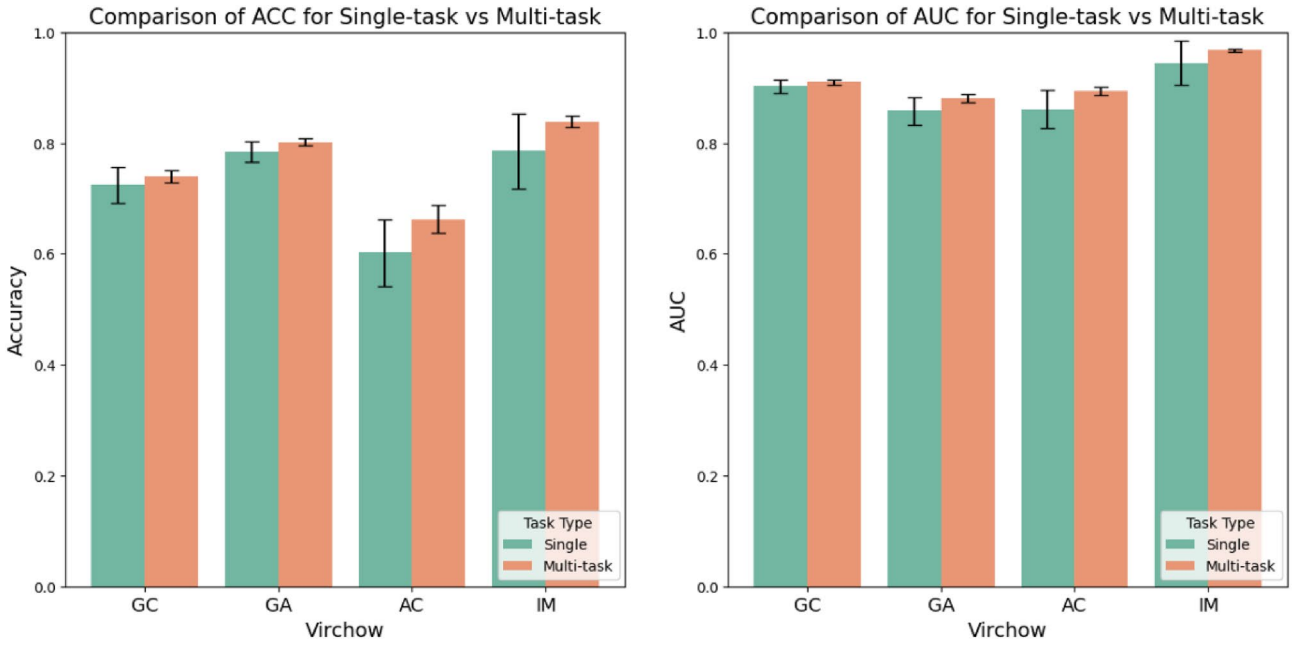


Fig. 3. Comparison of STL and MTL on ACC and AUC in the multi-class classification task.

STL and MTL results. All metrics are computed using five-fold cross-validation, and both mean and standard deviation are reported. In the single-task benchmark, performance across tasks is suboptimal (GC: 72.46%, GA: 78.46%, AC: 60.16%, IM: 78.56%). Notably, the AC task exhibits the lowest performance, likely due to class imbalance within the dataset. In contrast, multi-task learning (MTL) yields performance improvements across all tasks compared to STL (GC: +1.52%, GA: +1.72%, AC: +6.09%, IM: +5.29%). The MTL framework significantly outperforms STL in terms of accuracy, with particularly notable improvements in GA and IM tasks. Furthermore, MTL demonstrates superior performance in AUC and F1-score, suggesting that knowledge sharing across different tasks enhances the model's overall capability in Fig. 3.

To further evaluate the generalization ability of our method, we conducted external validation on two independent test cohorts: 259 cases from Peking University Third Hospital and 280 cases from Chongqing University Affiliated Three Gorges Hospital. All slides were digitized using the SQS-600P whole-slide imaging system (Shenzhen Shengqiang Technology Co., Ltd., Shenzhen, China). The detailed results on these two external centers are summarized in Tables 3 and 4. The external multi-center validation results demonstrate that the multi-task model consistently outperforms the single-task model across all tasks, providing strong evidence for the effectiveness of multi-task learning in assessing gastritis severity, atrophy, acute activity, and intestinal metaplasia.

By plotting the Receiver Operating Characteristic (ROC) curves for the four pathological assessment tasks–inflammation grading, atrophy assessment, acute activity grading, and intestinal metaplasia grading–under both single-task and multi-task learning frameworks in Fig. 4, we systematically illustrate the performance differences across classification strategies. The results demonstrate a significant advantage of multi-task learning, as the AUC values for all tasks are consistently higher under the multi-task learning framework compared to single-task learning. This improvement is particularly pronounced in the acute

Encoder	Setting	Task	Accuracy	AUC	F-score
Virchow	Single	GC	0.7250 (0.0324)	0.8770 (0.0198)	**0.7145 (0.0326)**
		GA	0.8535 (0.0479)	0.7338 (0.0557)	0.8451 (0.0290)
		AC	0.6464 (0.0233)	0.7128 (0.3988)	0.6363 (0.0289)
		IM	0.7521 (0.0224)	0.8729 (0.2207)	0.7223 (0.0892)
	Multi-task	GC	**0.7357 (0.0261)**	**0.8853 (0.0212)**	0.6825 (0.0428)
		GA	**0.8536 (0.0229)**	**0.7820 (0.0614)**	**0.8454 (0.0613)**
		AC	**0.6607 (0.0196)**	**0.8871 (0.0166)**	**0.6372 (0.0284)**
		IM	**0.7893 (0.0331)**	**0.9286 (0.0162)**	**0.7774 (0.0314)**

Table 3. Comparison of STL and MTL Performance – Chongqing University Affiliated Three Gorges Hospital. Bold values indicate the largest metric value within the same index task.

Encoder	Setting	Task	Accuracy	AUC	F-score
Virchow	Single	GC	0.7377 (0.0469)	0.7322 (0.4096)	0.7277 (0.0445)
		GA	0.8996 (0.0415)	0.9218 (0.0550)	0.9012 (0.0415)
		AC	0.6720 (0.0862)	0.9092 (0.0284)	0.6750 (0.0778)
		IM	0.8765 (0.0396)	0.9737 (0.0150)	0.8675 (0.0475)
	Multi-task	GC	**0.8265 (0.0208)**	**0.9339 (0.0068)**	**0.8156 (0.0234)**
		GA	**0.8997 (0.0186)**	**0.9352 (0.0201)**	**0.8967 (0.0287)**
		AC	**0.7414 (0.0246)**	**0.9344 (0.0142)**	**0.7393 (0.0348)**
		IM	**0.8802 (0.0188)**	**0.9832 (0.0047)**	**0.8686 (0.0239)**

Table 4. Multi-task learning performance in Peking University Third Hospital. Bold values indicate the largest metric value within the same index task.

activity grading task, suggesting that multi-task learning enhances classification performance by leveraging knowledge sharing across different tasks.

Figure 5 presents the confusion matrices for four different tasks: inflammation, atrophy, acute activity and intestinal metaplasia. Each subplot displays the confusion matrix corresponding to the model's prediction performance for the respective task. From the matrices, it is evident that the values along the main diagonal (correct predictions) are significantly higher than those in other regions, indicating that the model effectively distinguishes key pathological categories across tasks. Misclassifications are predominantly concentrated between adjacent severity levels, with fewer errors occurring across distant severity levels. This pattern suggests that adjacent severity categories exhibit substantial clinical similarity, making boundary distinctions more challenging for the model.

Binary classification results

Table 5 presents the performance of the Virchow feature extraction model on four binary classification tasks under both single-task learning (STL) and multi-task learning (MTL) frameworks. The reported evaluation metrics include accuracy, AUC, and F1-score, with a comparative analysis between STL and MTL results. Similarly, it can be observed that in binary classification tasks, the multi-task learning model outperforms the single-task model across all tasks in Fig. 6. This further reinforces the notion that multi-task learning enhances the model's overall capability by leveraging knowledge sharing across different tasks.

Figure 7 compares the ROC curves for the four classification tasks under single-task and multi-task learning. Each subplot presents the ROC curve for the corresponding task and annotates the AUC values for each learning framework. However, compared to multi-class classification tasks, the AUC improvement of multi-task learning over single-task learning is less pronounced in binary classification tasks. This finding suggests that in simpler binary classification, the performance of multi-task learning is relatively limited, whereas in more complex multi-class problems, multi-task learning exhibits substantial improvements. This may be attributed to the fact that binary classification tasks often have less distinct feature boundaries, resulting in more limited benefits from shared feature learning.

Figure 8 presents the confusion matrices for four different tasks: inflammation grading, atrophy assessment, acute activity grading, and intestinal metaplasia grading. Overall, the model demonstrates high classification performance across all binary classification tasks, with particularly strong performance in inflammation and intestinal metaplasia detection. However, some misclassification remains in atrophy assessment and acute activity grading, indicating areas for potential improvement.

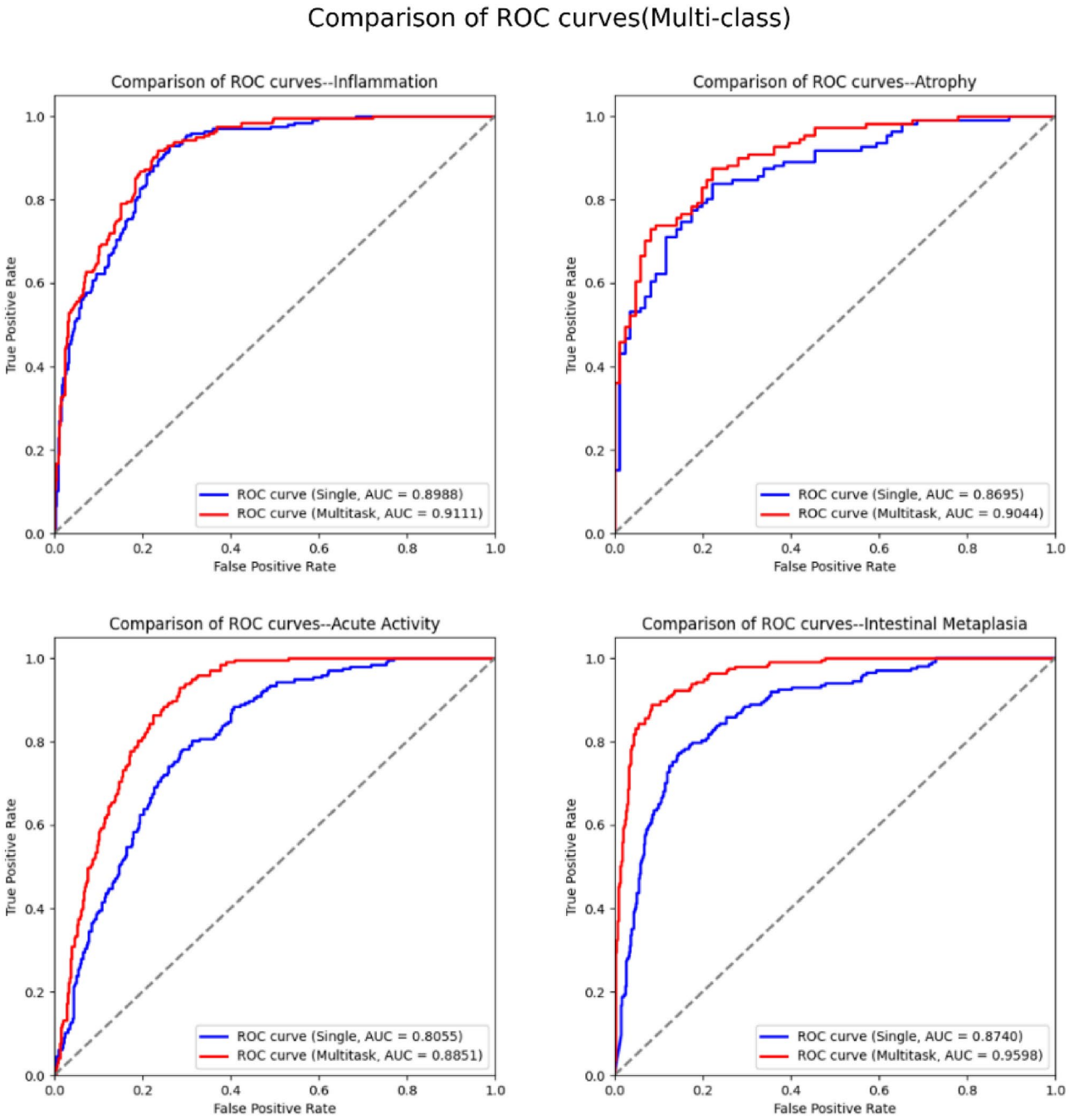


Fig. 4. Comparison of STL and MTL of ROC curves.

Ablation study

Effectiveness of different feature encoders

To ensure a fair comparison, we applied the same coordinate locations and threshold segmentation parameters for feature extraction across different feature extractors. We evaluated the impact of different feature extractors on the multi-task model predictive performance. In Table 6, the results indicate that the Virchow feature extractor outperforms other extractors in terms of overall performance across all four tasks. Similarly, from the T-SNE visualization in Fig. 9 which presents the visualization of feature embeddings after dimensionality reduction for the intestinal metaplasia task. Four colors correspond to the four severity levels of intestinal metaplasia: none, mild, moderate, and severe. As shown in the figure, the Virchow model achieves a clearer separation among these categories, whereas the other three feature extractors exhibit more overlaps and mixing across classes. This indicates that Virchow provides superior feature representation and classification performance compared to other feature encoder.

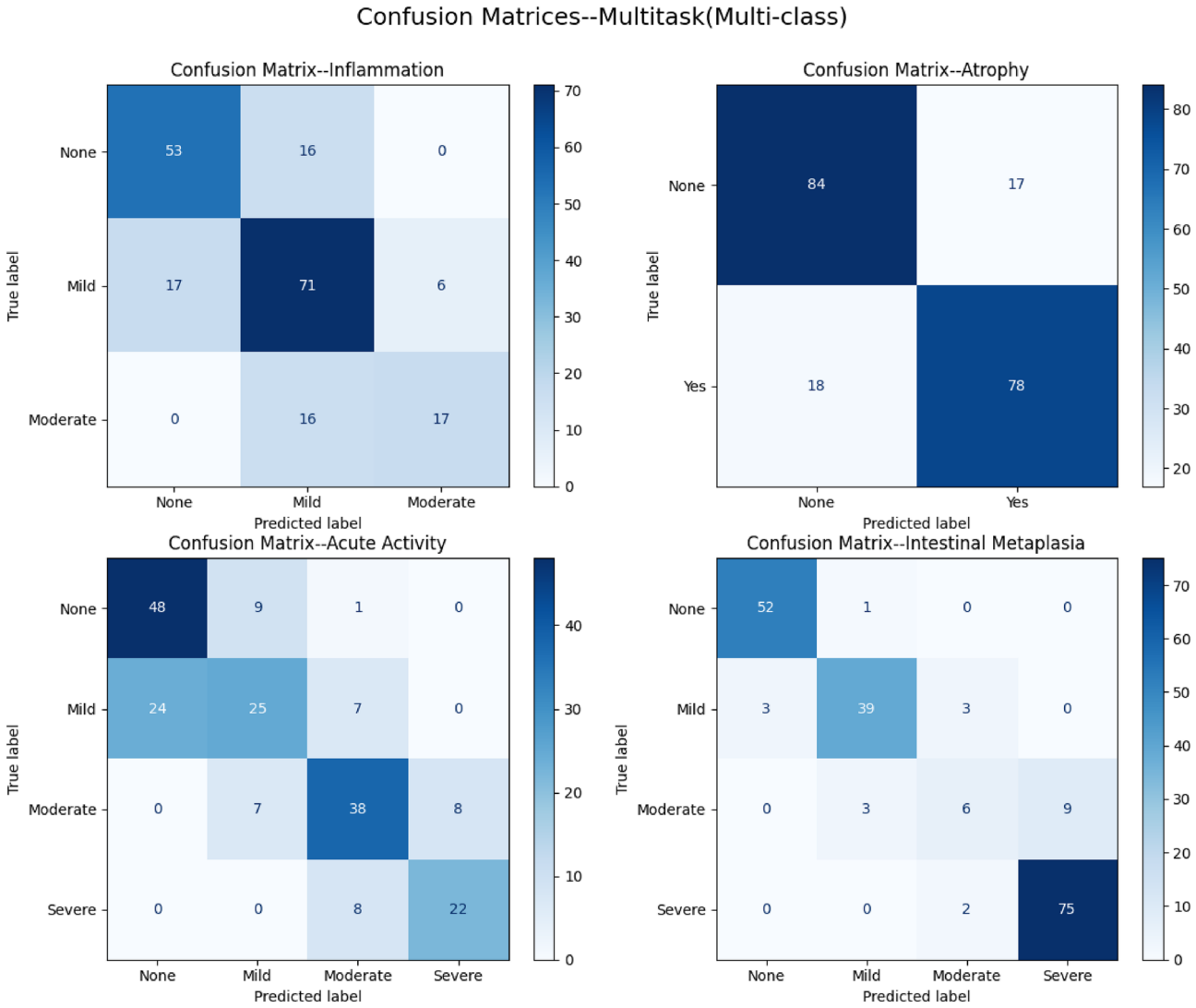


Fig. 5. Comparison of single-task and multi-task of confusion matrix.

Encoder	Task	Index	Accuracy	AUC	F-score
Virchow	Single	GC	1 (0)	1 (0)	1 (0)
		GA	0.8592 (0.0303)	0.9348 (0.0049)	0.8601 (0.0289)
		AC	0.8273 (0.0187)	0.9347 (0.0130)	0.8266 (0.0187)
		IM	0.9671 (0.0131)	0.9875 (0.0072)	0.9671 (0.0131)
	Multi-task	GC	1 (0)	1 (0)	1 (0)
		GA	**0.8620 (0.0071)**	**0.9422 (0.0031)**	**0.8609 (0.0069)**
		AC	**0.8437 (0.0081)**	**0.9467 (0.0106)**	**0.8650 (0.0271)**
		IM	**0.9698 (0.0085)**	**0.9946 (0.0025)**	**0.9698 (0.0085)**

Table 5. Comparison of STL and MTL performance—binary classification. Bold values indicate the largest metric value within the same index task.

Effectiveness of fully connected layers

To further investigate the contribution of fully connected layers to model performance, we conducted a series of ablation experiments. The table below compares the performance of different feature extractors under two settings: (1) using only extracted features and (2) incorporating fully connected (FC) layers.

The results indicate that incorporating FC layers generally enhances the model's performance compared to the standard ABMIL model shown in Table 7. Specifically, integrating FC layers into the Virchow feature extractor significantly improves multi-task performance, yielding accuracy improvements of 0.10% in

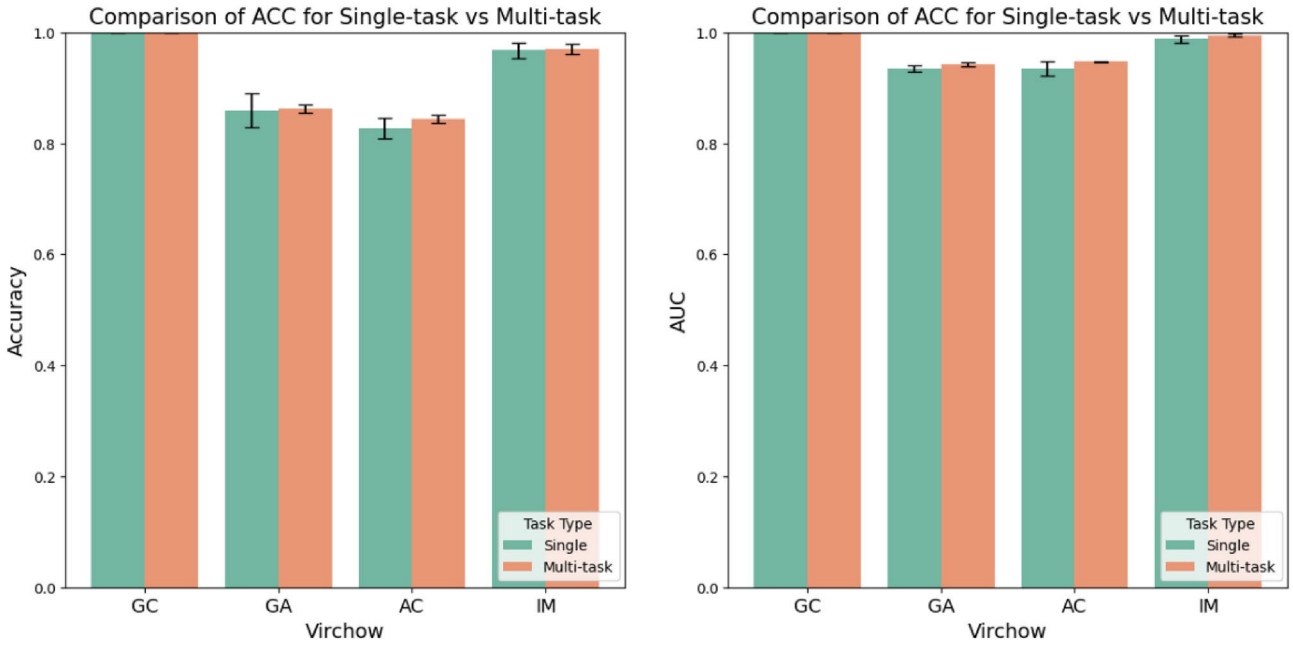


Fig. 6. Comparison of STL and MTL of ACC and AUC—binary classification.

inflammation grading, 5.18% in acute activity grading, and 4.28% in intestinal metaplasia grading. For other feature extractors, Conch shows accuracy improvements of 2.64% (inflammation), 1.23% (atrophy), 0.61% (acute activity), and 1.01% (intestinal metaplasia). Prov exhibits improvements of 0.62% (atrophy), 1.22% (acute activity), and 4.67% (intestinal metaplasia). Uni achieves gains of 1.32% (inflammation), 4.37% (acute activity), and 1.72% (intestinal metaplasia).

However, the multi-task learning model incorporating FC layers shows a decline in accuracy for the atrophy assessment task. This may be attributed to inherent differences among the four tasks. Introducing FC layers increases the number of model parameters, making training more complex. While some tasks benefit from a more intricate architecture–leading to improved accuracy–simpler tasks such as atrophy assessment (a binary classification task) may suffer from a representation bottleneck in the compressed feature space, receiving less attention during training, which ultimately leads to a decline in accuracy.

In Table 8, the results indicate that in binary classification tasks, the accuracy for inflammation classification reached 100%. The incorporation of fully connected (FC) layers led to improved accuracy in the classification of atrophy, acute activity, and intestinal metaplasia across various feature extractors. Specifically, using the Virchow feature extractor, the accuracy increased by 1.56% in atrophy classification, 1.65% in acute activity classification, and 1.91% in intestinal metaplasia classification. With the Conch feature extractor, the accuracy improved by 2.93% in atrophy classification, 2.56% in acute activity classification, and 1.64% in intestinal metaplasia classification. The Prov feature extractor yielded an increase of 3.18% in atrophy classification, 2.00% in acute activity classification, and 0.18% in intestinal metaplasia classification. Meanwhile, the Uni feature extractor showed improvements of 1.38% in atrophy classification and 1.09% in acute activity classification.

Visualization analysis

Whole-slide image (WSI)-level heatmaps can be used to verify whether the model's predictions align with the morphological features that pathologists rely on for diagnosis. The ABMIL-based multi-task learning model leverages the relationships between different tasks to identify regions of high diagnostic importance (i.e., regions with high attention scores). To visualize and interpret the relative importance of each region in the WSI, we convert the attention scores of the predicted class into percentiles and map the normalized scores to their corresponding spatial locations in the original WSI, thereby generating an attention heatmap. Additionally, by setting overlapping regions, we create fine-grained attention heatmaps for better interpretability[17].

By comparing the feature attention mechanisms of single-task learning (STL) and multi-task learning (MTL) in gastric mucosal pathology analysis, this study reveals the representational differences across various pathological features. In Fig. 10, the heatmap analysis indicates that in gastritis severity classification, single-task models primarily focus on glandular structures or edge morphology, whereas multi-task learning exhibits a significant shift in attention toward stromal regions. This observation closely aligns with the clinical practice of grading gastritis based on inflammatory cell infiltration (e.g., lymphocytes and plasma cells). Compared to single-task learning, the multi-task model demonstrates more precise and extensive attention to atrophy, covering a broader area that corresponds to key clinical markers such as glandular loss and fibrosis. In the acute activity grading task, attention is primarily directed toward

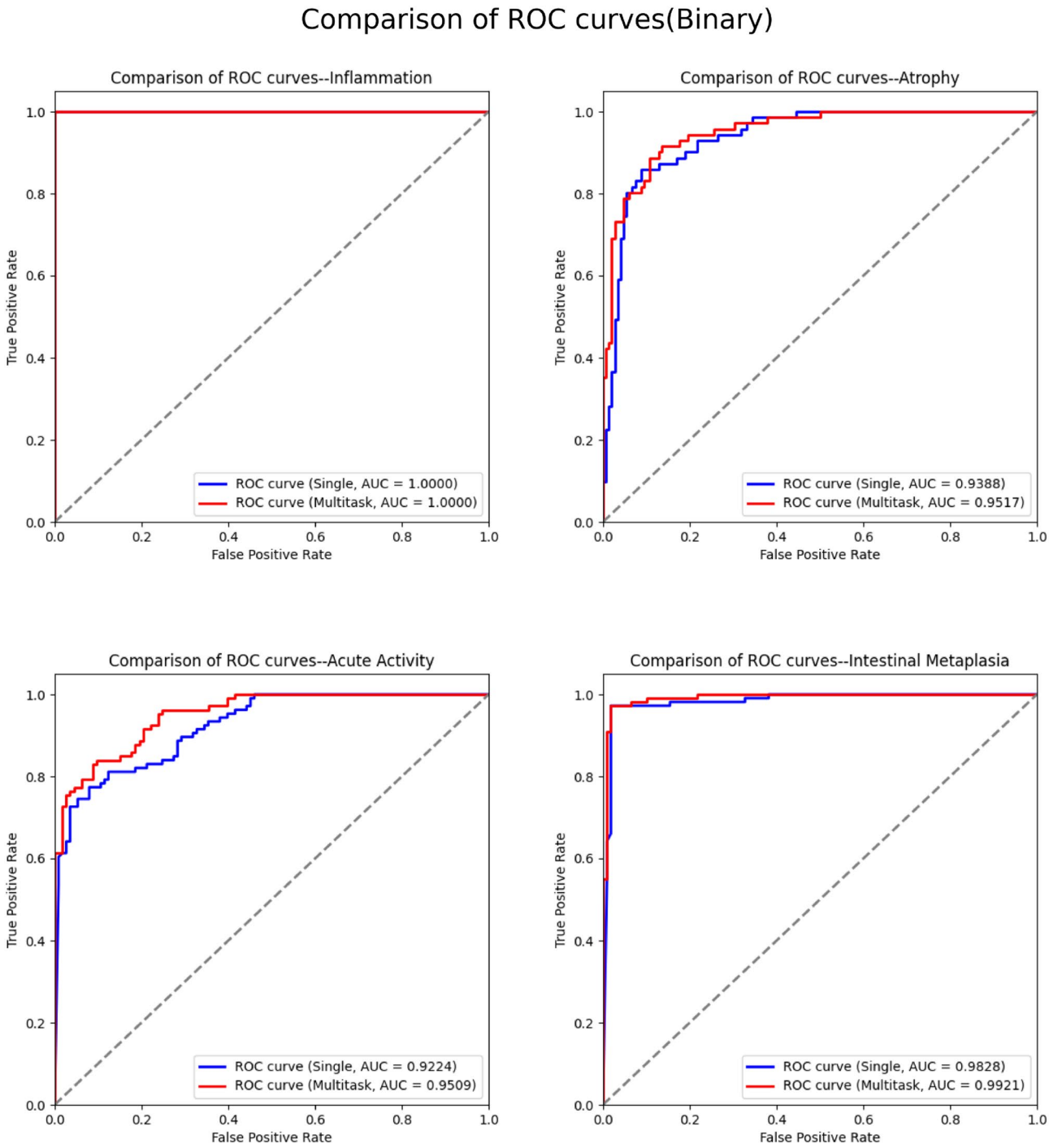


Fig. 7. Comparison of STL and MTL of ROC curves.

neutrophils. The multi-task model not only identifies a wider range of neutrophil infiltration but also achieves greater localization precision. For the analysis of intestinal metaplasia, pathologists typically focus on the presence of goblet cells within gastric glands. While the single-task model recognizes gastric glands as relevant features, it fails to accurately identify goblet cells. In contrast, the multi-task learning model significantly enhances the recognition accuracy of goblet cells, demonstrating its superior capability in capturing clinically relevant pathological features.

Computational cost

Although the Virchow model has 632M parameters, our benchmarking results indicate that its practical computational requirements are relatively modest. As shown in Table 9, Virchow required 30.54 seconds and 2765.63 MB of peak memory to process a whole-slide image during feature extraction. These values

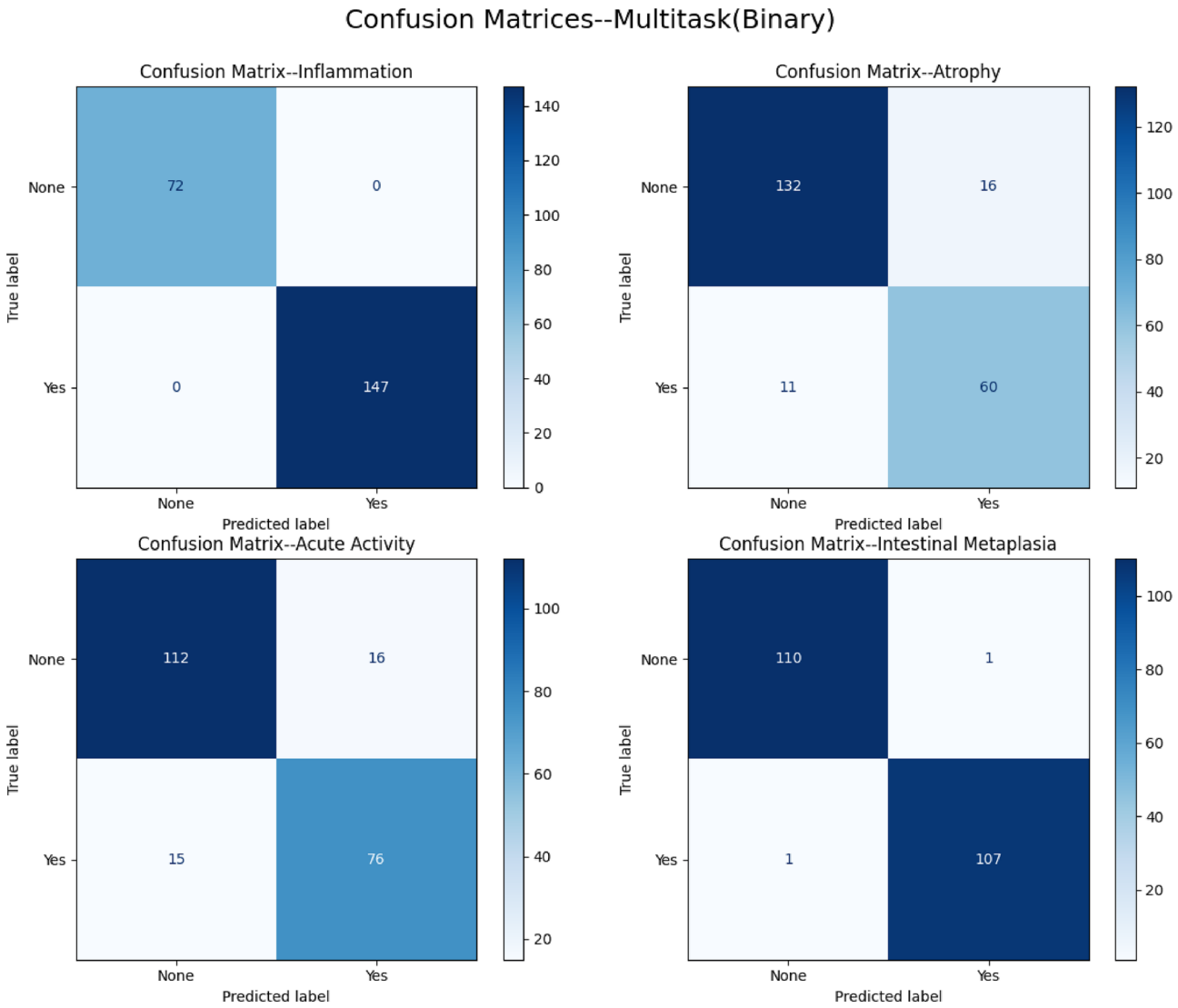


Fig. 8. Comparison of single-task and multi-task of confusion matrix.

are comparable to other encoders (e.g., Uni and Conch) and remain within a reasonable range for clinical deployment. Moreover, Prov exhibited the highest memory consumption (4677.77 MB), whereas Virchow's requirements were substantially lower. These results suggest that, despite its larger parameter count, Virchow can be deployed in practice without prohibitive computational overhead when Virchow has higher accuracy.

Discussion

Gastric biopsy is a core diagnostic tool for gastric cancer screening, and histological evaluation is crucial for diagnosing non-tumor conditions such as inflammation, atrophy, acute activity, and intestinal metaplasia. However, the lack of standardized grading criteria among pathologists worldwide, coupled with the subjective nature and poor reproducibility of assessments, affects the stability and efficiency of clinical diagnosis. This study addresses these challenges by developing a multi-task deep learning model for the assessment of non-tumor gastric diseases, focusing on four key pathological indicators: inflammation grading, atrophy assessment, acute activity grading, and intestinal metaplasia grading. By integrating artificial intelligence (AI) into the diagnostic workflow, the model enables objective quantification of pathological features while leveraging the interrelations among different indicators. Experimental results demonstrate that modeling based on pathological associations significantly enhances generalization and validates the superiority of the multi-task learning (MTL) framework in both binary and multi-class classification tasks. Compared to single-task models, the MTL framework achieves overall performance improvements, particularly in complex tasks, suggesting that cross-task knowledge transfer mechanisms help capture latent feature correlations, thereby improving classification accuracy. Additionally, the MTL model learns deep representations more effectively, leading to better performance in complex prediction tasks.

Encoder	Task	Accuracy	AUC	F-score
Virchow	GC	0.7398 (0.0111)	0.9100 (0.0047)	0.7367 (0.0107)
	GA	**0.8018 (0.0073)**	**0.8807 (0.0069)**	**0.7636 (0.0369)**
	AC	**0.6625 (0.0249)**	0.8935 (0.0081)	**0.6883 (0.0414)**
	IM	**0.8385 (0.0100)**	**0.9671 (0.0028)**	**0.8173 (0.0164)**
Conch	GC	**0.7581 (0.0174)**	**0.9174 (0.0062)**	**0.7545 (0.0185)**
	GA	0.7917 (0.0059)	0.8770 (0.0058)	0.7521 (0.0399)
	AC	0.6544 (0.0236)	**0.8956 (0.0076)**	0.6725 (0.0394)
	IM	0.8272 (0.0130)	0.9653 (0.0039)	0.8120 (0.0086)
Prov	GC	0.7530 (0.0035)	0.9057 (0.0035)	0.7382 (0.0086)
	GA	0.7754 (0.0120)	0.8607 (0.0152)	0.7304 (0.0385)
	AC	0.6321 (0.0094)	0.8869 (0.0042)	0.6672 (0.0405)
	IM	0.8228 (0.0110)	0.9640 (0.0035)	0.8172 (0.0068)
Uni	GC	0.7368 (0.0120)	0.9034 (0.0042)	0.7323 (0.0093)
	GA	0.7815 (0.0060)	0.8605 (0.0123)	0.7395 (0.0354)
	AC	0.6280 (0.0255)	0.8725 (0.0106)	0.6567 (0.0421)
	IM	0.8191 (0.0064)	0.9604 (0.0024)	0.8096 (0.0082)

Table 6. Performance comparison across different feature encoders. Bold values indicate the largest metric value within the same task of different feature encoders.

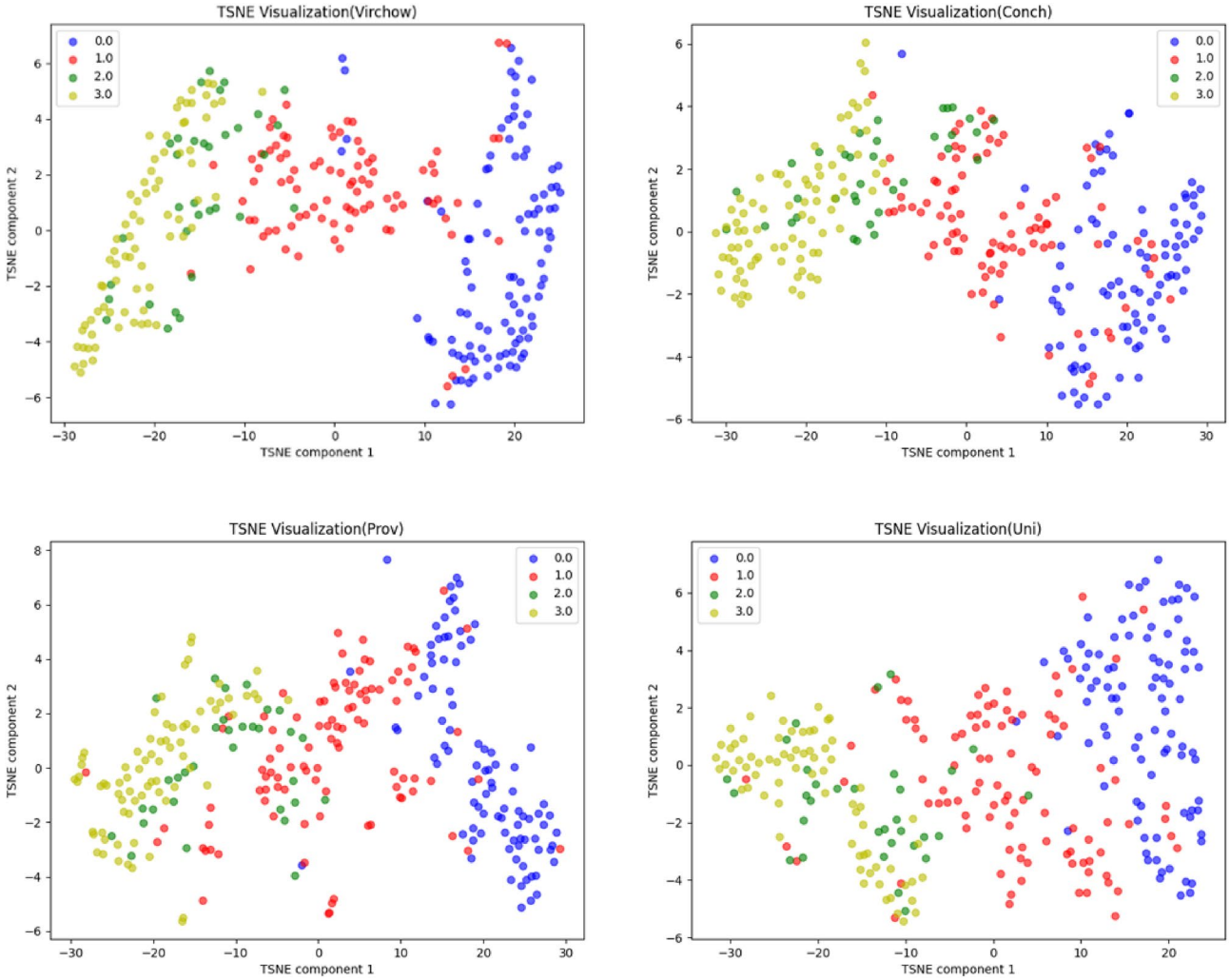


Fig. 9. T-SNE visualization of different feature encoders.

Encoder	Task	Accuracy	AUC	F-score
Virchow	GC	0.7388 (0.0048)	0.8931 (0.0054)	0.7343 (0.0059)
	GA	**0.8059 (0.0059)**	0.8667 (0.0024)	0.7584 (0.0437)
	AC	0.6107 (0.0141)	0.8645 (0.0078)	0.6286 (0.0340)
	IM	0.7957 (0.0147)	0.9437 (0.0054)	0.7842 (0.0065)
	GC-fc	**0.7398 (0.0111)**	**0.9100 (0.0047)**	**0.7367 (0.0107)**
	GA-fc	0.8018 (0.0073)	**0.8807 (0.0069)**	**0.7636 (0.0369)**
	AC-fc	**0.6625 (0.0249)**	**0.8935 (0.0081)**	**0.6883 (0.0414)**
	IM-fc	**0.8385 (0.0100)**	**0.9671 (0.0028)**	**0.8173 (0.0164)**
Conch	GC	0.7317 (0.0175)	0.9002 (0.0100)	0.7229 (0.0190)
	GA	0.7794 (0.0028)	0.8592 (0.0040)	0.7447 (0.0440)
	AC	0.6483 (0.0192)	0.8882 (0.0065)	0.6658 (0.0331)
	IM	0.8171 (0.0085)	0.9629 (0.0038)	0.7882 (0.0149)
	GC-fc	**0.7581 (0.0174)**	**0.9174 (0.0062)**	**0.7545 (0.0185)**
	GA-fc	**0.7917 (0.0059)**	**0.8770 (0.0058)**	**0.7521 (0.0399)**
	AC-fc	**0.6544 (0.0236)**	**0.8956 (0.0076)**	**0.6725 (0.0394)**
	IM-fc	**0.8272 (0.0103)**	**0.9635 (0.0033)**	**0.8139 (0.0097)**
Prov	GC	**0.7591 (0.0203)**	**0.9057 (0.0080)**	**0.7587 (0.0196)**
	GA	0.7692 (0.0153)	0.8455 (0.0096)	**0.7327 (0.0268)**
	AC	0.6199 (0.0161)	0.8740 (0.0045)	0.6431 (0.0284)
	IM	0.7815 (0.0106)	0.9479 (0.0021)	0.7646 (0.0107)
	GC-fc	0.7530 (0.0035)	0.9025 (0.0055)	0.7508 (0.0046)
	GA-fc	**0.7754 (0.0120)**	**0.8607 (0.0152)**	0.7304 (0.0358)
	AC-fc	**0.6321 (0.0094)**	**0.8869 (0.0042)**	**0.6672 (0.0405)**
	IM-fc	**0.8228 (0.0110)**	**0.9640 (0.0035)**	**0.8172 (0.0068)**
Uni	GC	0.7236 (0.0135)	0.8892 (0.0093)	0.7169 (0.0142)
	GA	**0.7917 (0.0118)**	0.8598 (0.0156)	**0.7416 (0.0470)**
	AC	0.5843 (0.0116)	0.8491 (0.0080)	0.6086 (0.0366)
	IM	0.8019 (0.0171)	0.9470 (0.0061)	0.7757 (0.0157)
	GC-fc	**0.7368 (0.0120)**	**0.9034 (0.0042)**	**0.7323 (0.0093)**
	GA-fc	0.7815 (0.0060)	**0.8605 (0.0123)**	0.7395 (0.0354)
	AC-fc	**0.6280 (0.0255)**	**0.8702 (0.0112)**	**0.6567 (0.0421)**
	IM-fc	**0.8191 (0.0064)**	**0.9604 (0.0024)**	**0.8096 (0.0082)**

Table 7. Performance comparison before and after incorporating fully connected layers. Bold values indicate the larger metric value when comparing results with and without the fc layer for the same task under the same encoder.

On the technical front, this study employs the Virchow model, the DINOv2 self-supervised pre-trained Vision Transformer (ViT-H/14), as the feature extractor. The multi-magnification training strategy ($5\times/10\times/20\times/40\times$) enhances the model's adaptability to multi-scale pathological features. With 632 million parameters, the Virchow model provides sufficient representational capacity for learning complex pathological patterns. Ablation studies further validate the model's architecture. Since atrophy is typically observed at lower magnifications, whereas acute activity requires high magnification for detailed examination, the Virchow feature extractor outperforms other models due to its hybrid magnification training strategy. Although adding fully connected layers improves feature representation, it introduces competition for attention resources in simpler binary classification tasks such as atrophy classification, leading to a slight drop in accuracy. This finding highlights the need for dynamic alignment between model complexity and task difficulty, offering insights for future lightweight optimization.

From a clinical perspective, the proposed model standardizes output and reduces inter-pathologist subjectivity in histological assessment. Testing in primary healthcare settings indicates that pathologists can focus only on cases flagged as abnormal by the model, thus reducing diagnostic time and improving efficiency. Additionally, heatmap visualizations provide insight into pathologically relevant regions, offering an interpretable basis for clinical decision-making.

Despite its promising results, this study has several limitations. The current dataset size ($n = 984$) does not encompass all subtypes, limiting the model's ability to infer causal relationships. Furthermore, the "black-box" nature of deep learning models complicates the interpretation of cross-task feature interactions. Although gated attention mechanisms partially reveal associations, the direct interaction pathways between pathological indicators remain unclear. Future research should integrate multi-center prospective cohort studies ($n \geq 5000$) and multimodal data to further investigate the molecular mechanisms underlying these indicators.

Encoder	Task	Accuracy	AUC	F-score
Virchow	GC	1 (0)	1 (0)	1 (0)
	GA	0.8464 (0.0075)	0.9310 (0.0071)	0.8360 (0.0184)
	AC	0.8272 (0.0121)	0.9373 (0.0040)	0.8533 (0.0270)
	IM	0.9507 (0.0036)	0.9918 (0.0012)	0.9507 (0.0037)
	GC-fc	1 (0)	1 (0)	1 (0)
	GA-fc	**0.8620 (0.0071)**	**0.9422 (0.0031)**	**0.8609 (0.0069)**
	AC-fc	**0.8437 (0.0081)**	**0.9467 (0.0106)**	**0.8650 (0.0271)**
	IM-fc	**0.9698 (0.0085)**	**0.9946 (0.0025)**	**0.9698 (0.0085)**
Conch	GC	1 (0)	1 (0)	1 (0)
	GA	0.8464 (0.0075)	0.9310 (0.0071)	0.8360 (0.0184)
	AC	0.8272 (0.0121)	0.9373 (0.0040)	0.8533 (0.0270)
	IM	0.9507 (0.0036)	0.9918 (0.0012)	0.9507 (0.0037)
	GC-fc	1 (0)	1 (0)	1 (0)
	GA-fc	**0.8757 (0.0059)**	**0.9419 (0.0050)**	**0.8638 (0.0075)**
	AC-fc	**0.8528 (0.0033)**	**0.9461 (0.0007)**	**0.8778 (0.0251)**
	IM-fc	**0.9671 (0.0060)**	**0.9943 (0.0019)**	**0.9671 (0.0060)**
Prov	GC	1 (0)	1 (0)	1 (0)
	GA	0.8402 (0.0035)	0.9212 (0.0067)	0.8269 (0.0123)
	AC	0.8081 (0.0125)	0.9291 (0.0045)	0.8123 (0.0191)
	IM	0.9525 (0.0070)	**0.9876 (0.0033)**	0.9250 (0.0072)
	GC-fc	0.9991 (0.0009)	1 (0)	1 (0)
	GA-fc	**0.8720 (0.0076)**	**0.9408 (0.0045)**	**0.8689 (0.0097)**
	AC-fc	**0.8281 (0.0125)**	**0.9270 (0.0092)**	**0.8542 (0.0310)**
	IM-fc	**0.9543 (0.0079)**	0.9765 (0.0084)	**0.9543 (0.0079)**
Uni	GC	1 (0)	1 (0)	1 (0)
	GA	0.8601 (0.0035)	0.9330 (0.0034)	0.8445 (0.0172)
	AC	0.8346 (0.0125)	0.9279 (0.0092)	0.8370 (0.0428)
	IM	0.9644 (0.0081)	0.9876 (0.0039)	0.9643 (0.0082)
	GC-fc	1 (0)	1 (0)	1 (0)
	GA-fc	**0.8739 (0.0098)**	**0.9440 (0.0032)**	**0.8740 (0.0122)**
	AC-fc	**0.8455 (0.0044)**	**0.9436 (0.0107)**	**0.8700 (0.0428)**
	IM-fc	**0.9662 (0.0070)**	**0.9892 (0.0032)**	**0.9662 (0.0070)**

Table 8. Binary classification performance comparison before and after incorporating fully connected (FC) layers. Bold values indicate the larger metric value when comparing results with and without the fc layer for the same task under the same encoder.

This study presents the first multi-task learning framework for histological assessment of non-neoplastic gastric diseases, departing from conventional single-task modeling paradigms by leveraging the interrelationships among pathological indicators to inform model design. The proposed framework not only offers a reproducible and generalizable AI-assisted diagnostic tool for gastric pathology but also introduces a novel approach for the early detection of precancerous gastric lesions, highlighting both its scientific significance and translational potential. Importantly, the model's quantitative evaluation of key histological features–namely inflammation, acute activity, atrophy, and intestinal metaplasia–serves as a valuable auxiliary method for monitoring pharmacological efficacy and assessing treatment response. It is particularly applicable to clinical scenarios involving Helicobacter pylori eradication therapy, acid-suppressive medications, and mucosal protective agents. The model's standardized scoring mechanism and interpretable heatmap outputs may help establish a more objective and consistent decision-making system, demonstrating broad applicability in future clinical trials and real-world pharmacological research.

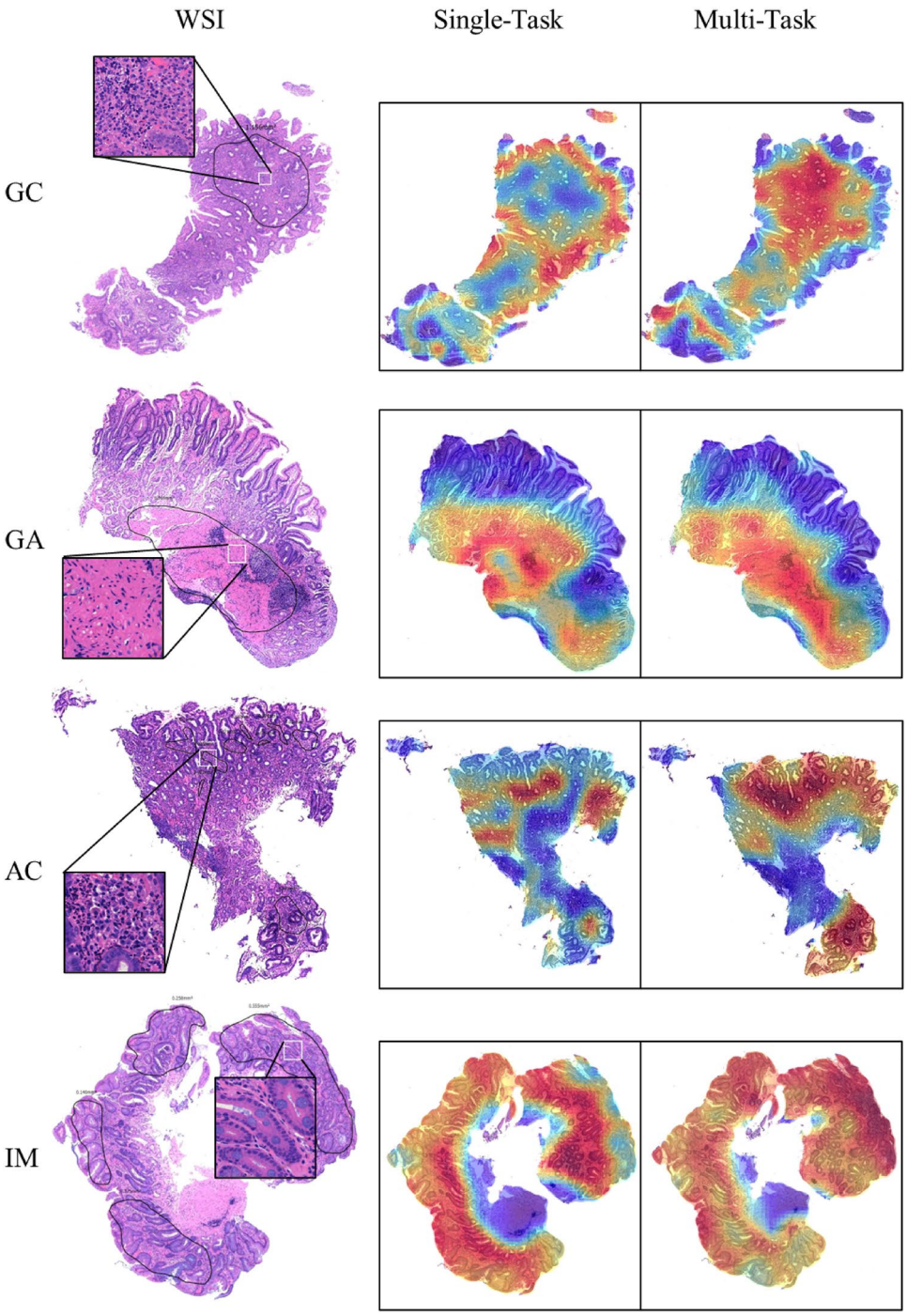


Fig. 10. Visualization analysis of different indicators in WSIs.

Encoder	Inference time (s)	Peak memory (MB)
Virchow	30.54	2765.63
Uni	16.77	1337.08
Prov	38.52	4677.77
Conch	16.79	1335.64

Table 9. Inference efficiency comparison of different feature extractors.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 19 May 2025; Accepted: 5 September 2025
Published online: 18 December 2025

References

1. Bray, F. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer J. Clin.* **74**, 229–263 (2024).
2. Correa, P. & Piazuelo, M. B. The gastric precancerous cascade. *J. Dig. Dis.* **13**, 2–9 (2012).
3. He, J. et al. Helicobacter pylori infection induces stem cell-like properties in correa cascade of gastric cancer. *Cancer Lett.* **542**, 215764 (2022).
4. Kim, N. & Park, Y. H. Atrophic gastritis and intestinal metaplasia. In *Helicobacter pylori* 229–251 (Springer, 2024).
5. Berlth, F., Bollschweiler, E., Drebber, U., Hoelscher, A. H. & Moenig, S. Pathohistological classification systems in gastric cancer: diagnostic relevance and prognostic value. *World J. Gastroenterol.: WJG* **20**, 5679 (2014).
6. Dixon, M. F., Genta, R. M., Yardley, J. H., Correa, P., the Participants in the International Workshop on the Histopathology of Gastritis, H. Classification and grading of gastritis: The updated Sydney system. *Am. J. Surg. Pathol.* **20**, 1161 (1996).
7. Nam, S. et al. Introduction to digital pathology and computer-aided pathology. *J. Pathol. Transl. Med.* **54**, 125–134 (2020).
8. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology-new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
9. Jahan, I. et al. Deep learning and vision transformers-based framework for breast cancer and subtype identification. *Neural Comput. Appl.* **37**(16), 9311–30 (2025).
10. Gao, C. et al. Transformer based multiple instance learning for wsi breast cancer classification. *Biomed. Signal Process. Control* **89**, 105755 (2024).
11. Bhattacharyya, R., Roy, P., Banerji, S. & Mitra, S. Efficient grading of prostate cancer WSI with deep learning. In *Medical Imaging 2024: Digital and Computational Pathology*, vol. 12933, 20–26 (SPIE, 2024).
12. Rashmi, R., Sudhamsh, G. V. S. & Girisha, S. A semi-supervised learning approach for tissue semantic segmentation in whole slide images. *IEEE Access: Pract. Innov. Open Solut.* **12**, 120482–120497 (2024).
13. Nahhas, O. S. et al. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. *Nat. Protoc.* **20**, 293–316 (2025).
14. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136 (PMLR, 2018).
15. Shao, Z. et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Process. Syst.* **34**, 2136–2147 (2021).
16. Li, H. *et al.* Long-mil: scaling long contextual multiple instance learning for histopathology whole slide image analysis. *arXiv preprint* arXiv:2311.12885 (2023).
17. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
18. Tang, W. *et al.* Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11343–11352 (2024).
19. Tang, K. et al. Self-supervised representation distribution learning for reliable data augmentation in histopathology WSI classification. *IEEE Trans. Med. Imaging* **44**, 462–474 (2025).
20. Wang, X. *et al.* TransPath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, vol. 12908, 186–195 (Springer International Publishing, Cham, 2021).
21. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
22. Hao, D., Li, Q., Zhang, Y. & Wu, S. Medical knowledge-enabled multi-task learning for gastric cancer survival prediction. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5 (IEEE, 2024).
23. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
24. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
25. Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12104–12113 (2022).
26. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
27. Zimmermann, E. *et al.* Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint* arXiv:2408.00738 (2024).
28. Fang, S. et al. Diagnosing and grading gastric atrophy and intestinal metaplasia using semi-supervised deep learning on pathological images: development and validation study. *Gastric Cancer* **27**, 343–354 (2024).
29. Lai, Q. et al. Multi-scale multi-instance multi-feature joint learning broad network (m3jlbn) for gastric intestinal metaplasia subtype classification. *Knowl.-Based Syst.* **249**, 108960 (2022).
30. Ma, B. et al. Artificial intelligence-based multiclass classification of benign or malignant mucosal lesions of the stomach. *Front. Pharmacol.* **11**, 572372 (2020).
31. Zhang, Y. & Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **5**, 30–43 (2018).
32. Fu, M. et al. Multitask machine learning-based tumor-associated collagen signatures predict peritoneal recurrence and disease-free survival in gastric cancer. *Gastric Cancer* **27**, 1242–1257 (2024).
33. Graham, S. et al. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Med. Image Anal.* **83**, 102685 (2023).
34. Hu, Z. et al. A multi-task deep learning framework for perineural invasion recognition in gastric cancer whole slide images. *Biomed. Signal Process. Control* **79**, 104261 (2023).
35. Zhou, Q. et al. Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 373–383 (Springer, 2024).
36. Otsu, N. et al. A threshold selection method from gray-level histograms. *Automatica* **11**, 23–27 (1975).
37. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).

Acknowledgements

Thank you for all authors and grants from National Natural Science Foundation of China (NSFC)(82430062), the Shenzhen Engineering Research Center (XMHT20230115004), and the Jilin FuyuanGuan Food Group Co.,

Ltd..

Author contributions

M.F. and L.L. conceived the study, developed the model, and wrote the main manuscript text. F.F. conducted the experiments. J.O. prepared the figures. X.S. and S.D. helped collect the pathological images and complete data collation. T.G. provided technical support and resources. Y.H., Z.C., and L.Z. supervised the project and revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This work was supported by National Natural Science Foundation of China (NSFC)(82430062), the Shenzhen Engineering Research Center (XMHT20230115004), and the Jilin FuyuanGuan Food Group Co., Ltd.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-19195-0.

Correspondence and requests for materials should be addressed to Y.H., Z.C. or L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025