scientific reports

OPEN

Explainable semi-supervised model for predicting invasion depth of esophageal squamous cell carcinoma based on the IPCL and AVA patterns

Liumin Kang[1,2,4], Jinzhou Zhu[1,4], Haoxiang Ni[1], Shiqi Zhu[1], Lihe Liu[1], Jiaxi Lin[1], Yu Wang[3], Xiaohua Shi[2]✉ & Rui Li[1]✉

Evaluation of invasion depth is essential for the treatment strategy of esophageal squamous cell carcinoma (ESCC). However, the application of the Japanese Endoscopic Society classification system, based on the patterns of intravascular papillary cell layer (IPCL) and avascular area (AVA), requires a long-term training for endoscopists. We aimed to develop explainable semi-supervised models for predicting ESCC invasion depth based on the IPCL/AVA patterns. A total of 2,643 images of magnifying endoscopy with narrow-band imaging in the upstream task, self-supervised contrastive learning (*n* = 2,175), and the downstream task, fine-tuning (*n* = 468), were from Suzhou. In the fine-tuning, two approaches were adopted: the traditional blackbox or the explainable AI. Lastly, the models were evaluated in an external test dataset (Jintan, *n* = 60), in comparison with two endoscopists. The primary outcome was 3-way classification of ESCC invasion depth. The metrics included accuracy, Matthew correlation coefficient, and Cohen's kappa. Furthermore, Grad-CAM was for visualized explanation of images; local interpretation, feature importance, and partial dependence plots were conducted for classifiers; and t-SNE was for visualization of feature vectors. A Xception-backboned explainable model (accuracy 0.817) had exhibited better performance than other models and a junior endoscopist (0.733), even though it underperformed a senior (0.883) by 0.066 on accuracy. However, the endoscopists' performance was improved by AI assistance (junior 0.833 and senior 0.917). The explainable semi-supervised framework empowers AI models to achieve improved transparentness and performance, facing the opacity of traditional supervised learning and limited amounts of labelled endoscopic images.

Keywords Esophageal squamous cell carcinoma (ESCC), Intravascular papillary cell layer (IPCL), Avascular area (AVA), Self-supervised learning (SSL), Grad-CAM, Local interpretation, Feature importance, Partial dependence plots (PDP), t-SNE

Esophageal cancer, ranking as the eighth most prevalent malignancy and the sixth leading cause of cancer-related mortality, is characterized by a low five-year survival rate[1,2]. It is primarily classified into two histological subtypes: esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma, with the former constituting the majority of global cases, approximately 84% [3].

In terms of therapeutic approach, endoscopic resection is recommended for esophageal neoplasia, ranging from epithelial (EP) to minimal submucosal invasion (SM1), owing to its lower complication rate and reduced duration of hospitalization, in comparison with surgical intervention[4]. Consequently, an accurate preoperative evaluation of the depth of invasion is essential for the decision-making of the treatment strategy.

[1]Department of Gastroenterology, the First Affiliated Hospital of Soochow University, Suzhou 215006, Jiangsu, China. [2]Department of Gastroenterology, Suzhou Research Center of Medical School, Affiliated Hospital of Medical School, Suzhou Hospital, Nanjing University, Suzhou 215000, Jiangsu, China. [3]Department of Hepatobiliary Surgery, Jintan Affiliated Hospital of Jiangsu University, Changzhou 213200, Jiangsu, China. [4]Liumin Kang and Jinzhou Zhu have contributed equally to this work. ✉email: sxhsz@sina.com; lrhcsz@163.com

The Japanese Endoscopic Society (JES) recognizes magnifying endoscopy with narrow-band imaging (ME-NBI) as a highly effective technique for the preoperative assessment of invasion depth in ESCC[5,6]. In following of multiple previous classification systems, the JES classification system proposes the patterns of intravascular papillary cell layer (IPCL) and avascular area (AVA), which now has been widespread clinical adoption[6,7]. However, the accurate application of the JES classification system in routine requires a long-term training for endoscopists.

The advancements in deep learning have transformed numerous facets of clinical operations. In gastrointestinal endoscopy, artificial intelligence (AI), trained by large amount of labeled data, is progressively being incorporated into computer-aided diagnosis systems, thereby enhancing the detection and classification of lesions[8,9]. However, the procurement of such comprehensive and meticulously annotated datasets, necessitating laborious and time-intensive curation, often presents a significant impediment to the training process[10]. Self-supervised learning (SSL) represents a novel machine learning paradigm that harnesses the power of unsupervised learning to equip cutting-edge AI models with the capacity to manage the exigencies of tasks traditionally reliant on extensive datasets annotated by human experts[11,12].

As AI evolves toward higher complexity, it poses a formidable challenge for humans to grasp the logic and steps that lead an algorithm to its conclusions. The computational processes often become encapsulated within an enigmatic framework known as a "blackbox," which is inherently resistant to interpretation[13,14]. It is imperative to achieve a comprehensive grasp of AI's decision-making mechanisms through rigorous model monitoring and ensuring AI accountability. Consequently, there is a burgeoning demand for the development of explainable AI methodologies aimed at bolstering confidence in AI models. Explainable AI is designed to demystify and elucidate the operations of machine learning algorithms, deep learning structures, and neural networks. In recent years, explainable AI has emerged as a prominent area of inquiry within the domain of AI research[15,16].

In this study, we aimed to develop a series of semi-supervised models for predicting invasion depth of ESCC, based on the IPCL/AVA patterns. The models were pretrained in a large unlabeled data using self-supervised contrastive learning and fine-tuned in a small labeled data. In the fine-tuning, two approaches were adopted: the traditional blackbox or the explainable AI. Lastly, the models were evaluated in an external test dataset, in comparison with two endoscopists.

Methods

Study design

This retrospective multicenter study was conducted in two hospitals: the First Affiliated Hospital of Soochow University (Suzhou, the training dataset) and Jintan Affiliated Hospital of Jiangsu University (Jintan, the test dataset). Patients who underwent ME-NBI examinations for precancerous lesions or superficial ESCC confirmed by histology of endoscopically or surgically resected specimens between November 2016 and December 2023 were included. Each lesion offered three images in the training dataset, whereas one image in the test dataset. Non-magnified images, low-quality images, and white-light images were excluded. These images were captured using Olympus equipment (GIF-H260Z, Olympus Medical Systems, Tokyo, Japan) and saved in BMP format. Two endoscopists (L.K. and J.Z.) with more than 10 years of experience independently reviewed eligible images to ensure image quality and labeled IPCL/AVA patterns. If there was disagreement between the two endoscopists, it would be decided by X.S., with 20 years of experience. This study was approved by the ethics committee of the First Affiliated Hospital of Soochow University (Approval number 2022098)[17]. Due to the retrospective nature of the study, the need to obtain informed consent was waived by the Ethics Committee of the First Affiliated Hospital of Soochow University. All procedures involving human participants were conducted in accordance with ethical standards and the Declaration of Helsinki. Figure 1 presents the flowchart of the study. The characteristics of patients/ lesions/ images are listed in Table 1.

Based on the pathological results of invasion depth, the ME-NBI images were labeled as: (1) epithelium (EP) to lamina propria (LPM); (2) muscularis mucosa (MM) to minimal submucosal invasion (less than 200 µm; SM1); or (3) deeper submucosal invasion (200 µm or more; SM2 or deeper). The detailed information was offered in the **Supplementary Methods 1**.

Based on the patterns of IPCL/AVA, each ME-NBI image was labeled in four dimensions: (1) severe irregularity (yes or no); (2) loop-like formation (yes or no); (3) highly dilated vessels which calibers appear to be more than 3 times (yes or no); and (4) AVA (small $<$ 0.5 mm; middle 0.5–3 mm; large $\geq$ 3 mm).

The proposed semi-supervised learning framework

The proposed framework, which consists of the upstream task SSL (Fig. 2) and the downstream task fine-tuning (Fig. 3). The development of the framework was introduced in our previous study[18].

The upstream task: self-supervised contrastive learning

In the SSL (contrastive learning), an anchor image is employed to create a positive example by applying various data augmentation strategies. In contrast, a negative example is derived from selecting another random image within the same batch. Figure 2 illustrates the SSL process, showcasing augmentation techniques such as color modification and cropping followed by resizing[18]. The process offers a simplified approach to learning visual representations contrastively, streamlining complex self-supervised learning algorithms to their core principles, eschewing the need for specialized architectural configurations or a memory bank[19]. The framework is characterized by several integral elements: (1) a data augmentation component that randomly alters an image sample to generate a related pair, as a positive match, while it generates negative matches, augmented from different images; (2) a neural network encoder that serves to extract feature vectors from the augmented data instances; (3) a concise neural network projection layer that aids in the translation of these features into a

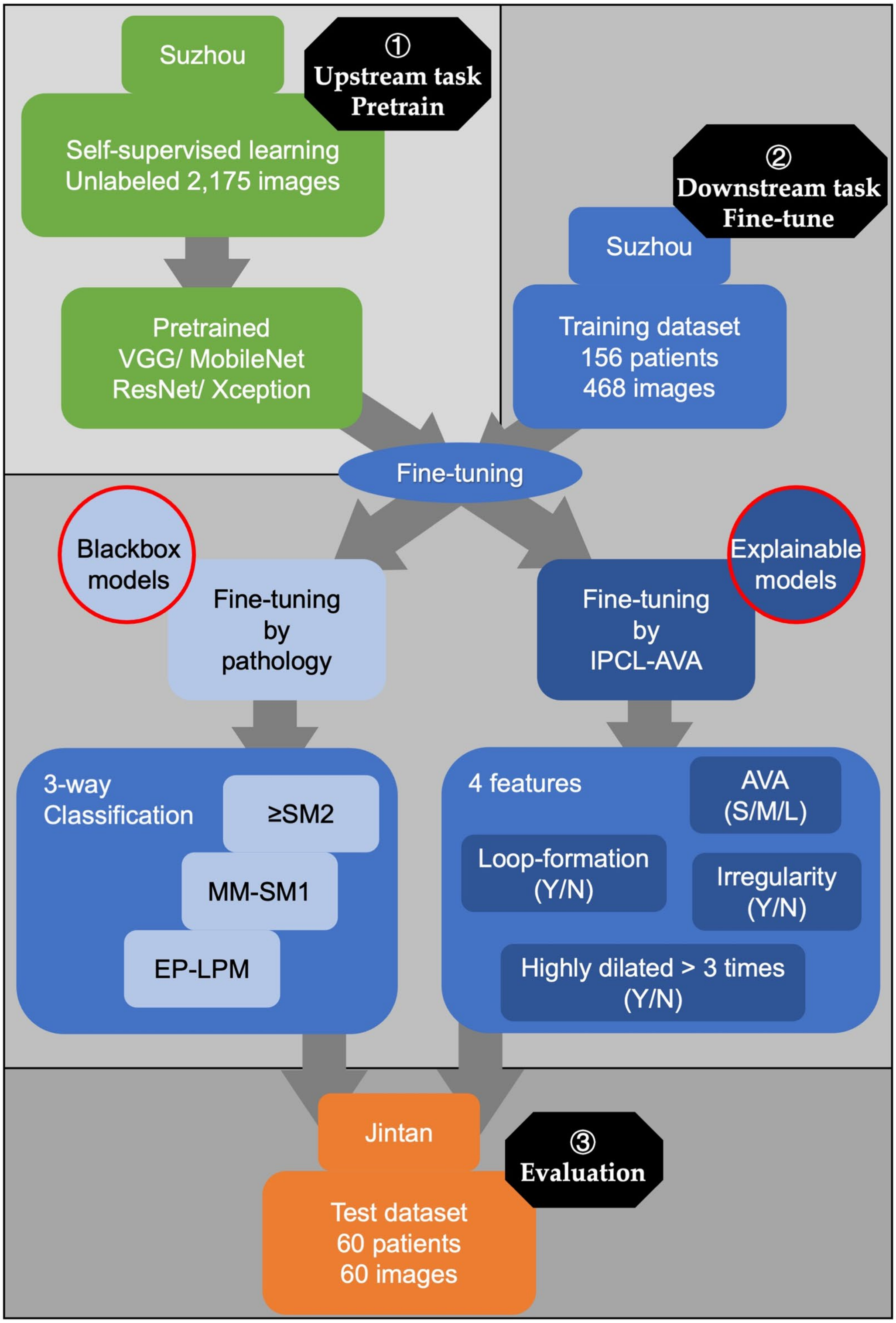

Fig. 1. The flowchart of the study. Step #1: self-supervised contrastive learning on large unlabeled images from Suzhou. Step #2: fine-tuning (two methods: blackbox or explainable) on few labeled images from Suzhou. Step #3: test on labeled images from Jintan. This retrospective multicenter study was conducted in two hospitals: the First Affiliated Hospital of Soochow University (Suzhou, the training dataset) and Jintan Affiliated Hospital of Jiangsu University (Jintan, the test dataset).

		Suzhou Training dataset ($n = 156$)	Jintan Test dataset ($n = 60$)
Age (yrs)		69 (61, 77)	67 (58, 73)
Sex (num, %)	Men	123	48
	Women	33	12
Lesion size (mm)		15 (6, 65)	18 (8, 68)
Lesions by depth (num, %)			
	EP-LPM	79	30
	MM-SM1	54	20
	≥SM2	23	10
Images by depth (num, %)			
	EP-LPM	237	30
	MM-SM1	162	20
	≥SM2	69	10

Table 1. Characteristics of patients in the study. Data of age and lesion size are presented as median (interquartile range). Each lesion offered three images in the training dataset, whereas one image in the test dataset.

space conducive to the application of the contrastive loss mechanism; and (4) a contrastive loss function that is precisely formulated to fulfill the contrastive prediction task, facilitating the refinement of the learning process.

The downstream task: Fine-tuning

As shown in Fig. 3, following the upstream task SSL, the pretrained backbone models were submitted to the downstream task fine-tuning, which were two approaches: (1) the traditional learning labeled on the pathology (blackbox) or (2) the explainable learning based on the IPCL/AVA patterns (explainable AI). The former was a 3-way supervised training based on the pathological results of the invasion depth, i.e., EP-LPM, MM-SM1, and ≥ SM2. The blackbox model outputs using a softmax classifier. The latter (explainable AI) was comprised of four feature models based on the IPCL/AVA patterns, i.e., irregularity, loop formation, dilation ≥ 3-time and size of AVA. The outputs of the four feature models were integrated by a XGBoost classifier, which was also supervised trained by pathology[18].

Evaluation

A total of 2,643 ME-NBI images in the upstream task SSL ($n = 2{,}175$) and the downstream task fine-tuning ($n = 468$) were from Suzhou. The 468 images were randomly divided into the training and the validation at a ratio of 7:3, thus 140 images were used to evaluate the models' performance during the training procedure (**Supplementary Table 2**). The images in the test were from Jintan ($n = 60$), as shown in Fig. 4. To compare with the models, images from the test data were evaluated by two independent endoscopists (junior, four years of endoscopic experience; and senior, eleven years of experience). They were blind to the collection and labelling of the images. Firstly, they labeled the test images independently, and then in the next week, they relabeled the test images in awareness of the prediction of the best AI model.

Model training

The Keras (version 3.8.0, TensorFlow version 2.8.0) was used to train the models. The training parameters are listed in **Supplementary Table 1**. Images were resized to 224 × 224 pixels and input into the framework in the form of RGB channels. The training code for SSL was inspired by that of Sayak Paul, which can be accessed at https://github.com/sayakpaul/SimCLR-in-TensorFlow-2. The training code is available at https://osf.io/t3g8n[18].

Statistical analysis and explanation

The primary outcome was 3-way classification of ESCC invasion depth. To evaluate the performance of the models and endoscopists, three metrics were calculated: accuracy, Matthew correlation coefficient (MCC), and weighted Cohen's kappa[18]. The detailed information of the metrics was offered in the **Supplementary Methods 2**. Furthermore,, Grad-CAM was conducted for visualized explanation of endoscopic images[20]; variable importance, local interpretation, and partial dependence plots (PDP) were for the XGBoost classifier[21]; and t-SNE was for the visualization of feature vectors (blackbox models vs. explainable models) in a two-dimensional space[18,22].

Results

Performance of the models in the downstream task

As shown in **Supplementary Table 2**, Xception-backboned explainable model presented the best performance, with accuracy 0.850, MCC 0.768 and weighted Cohen's kappa 0.770.

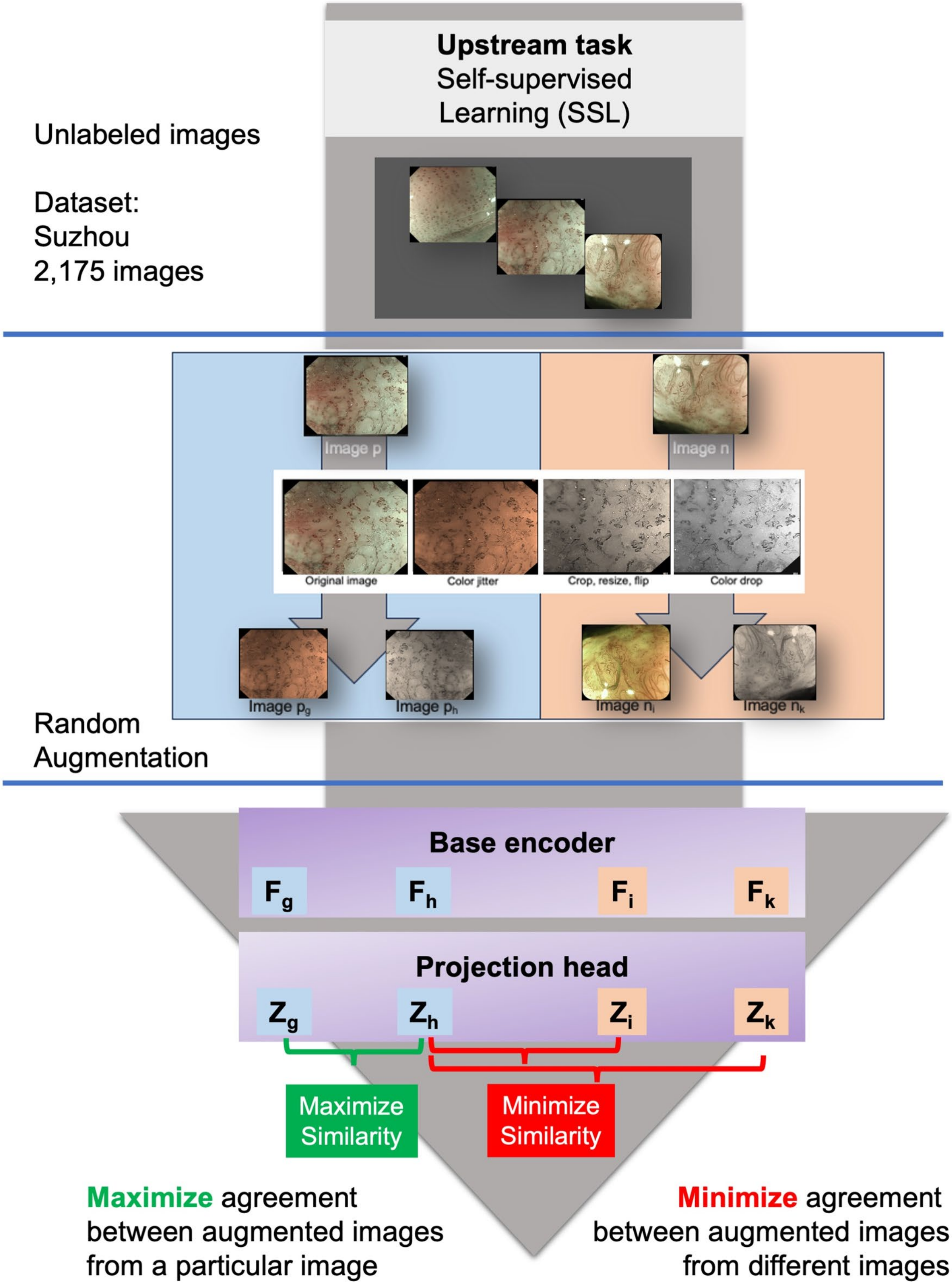

Fig. 2. The flowchart of Step #1 self-supervised contrastive learning. Self-supervised contrastive learning on large unlabeled images from Suzhou. The self-supervised contrastive learning is characterized by several integral elements: (1) a data augmentation component; (2) a neural network encoder; (3) a concise neural network projection layer; and (4) a contrastive loss function.

Performance of the models in the evaluation

The performances of the four explainable models and four blackbox models on the test set were shown in Table 2. Among the models, Xception-backboned explainable model achieved the highest accuracy (0.817), with highest MCC 0.701 and weighted Cohen's kappa 0.780. The confusion matrices are plotted in Fig. 5.

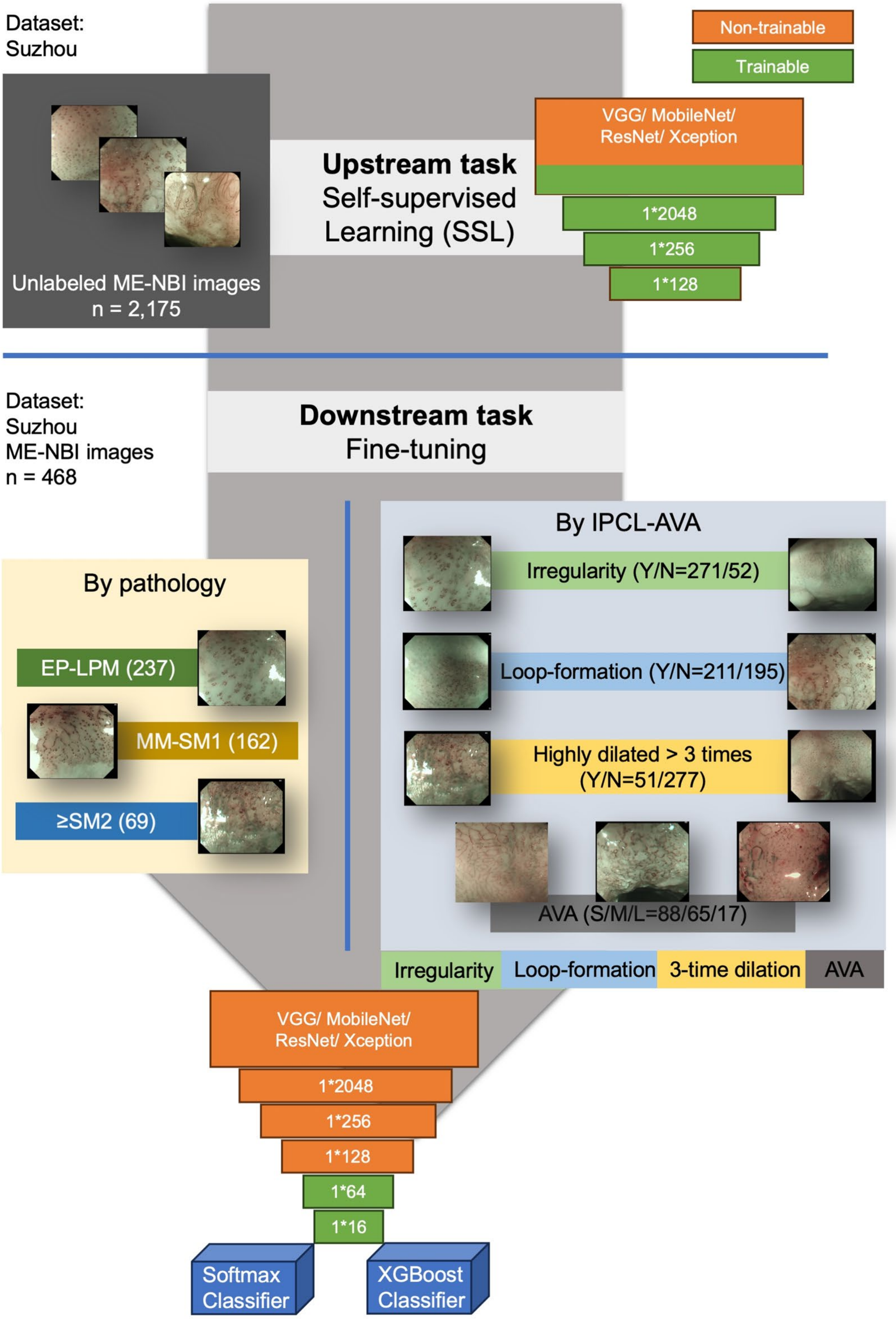

Fig. 3. The flowchart of Step #2 fine-tuning. The fine-tuning had two approaches: (1) blackbox models were trained based on traditional learning labeled on the pathology; or (2) four feature models were trained based on the IPCL/AVA patterns, and then their outputs were integrated using a XGBoost classifier in the principle of explainable AI.

Comparison with endoscopists

The performances of the junior and senior endoscopists were listed in Table 2. The senior endoscopist showed higher accuracy, MCC and weighted Cohen's kappa coefficient (0.883, 0.809 and 0.890) than Xception-backboned explainable model. The junior endoscopist had accuracy 0.733, MCC 0.571 and weighted Cohen's kappa 0.750.

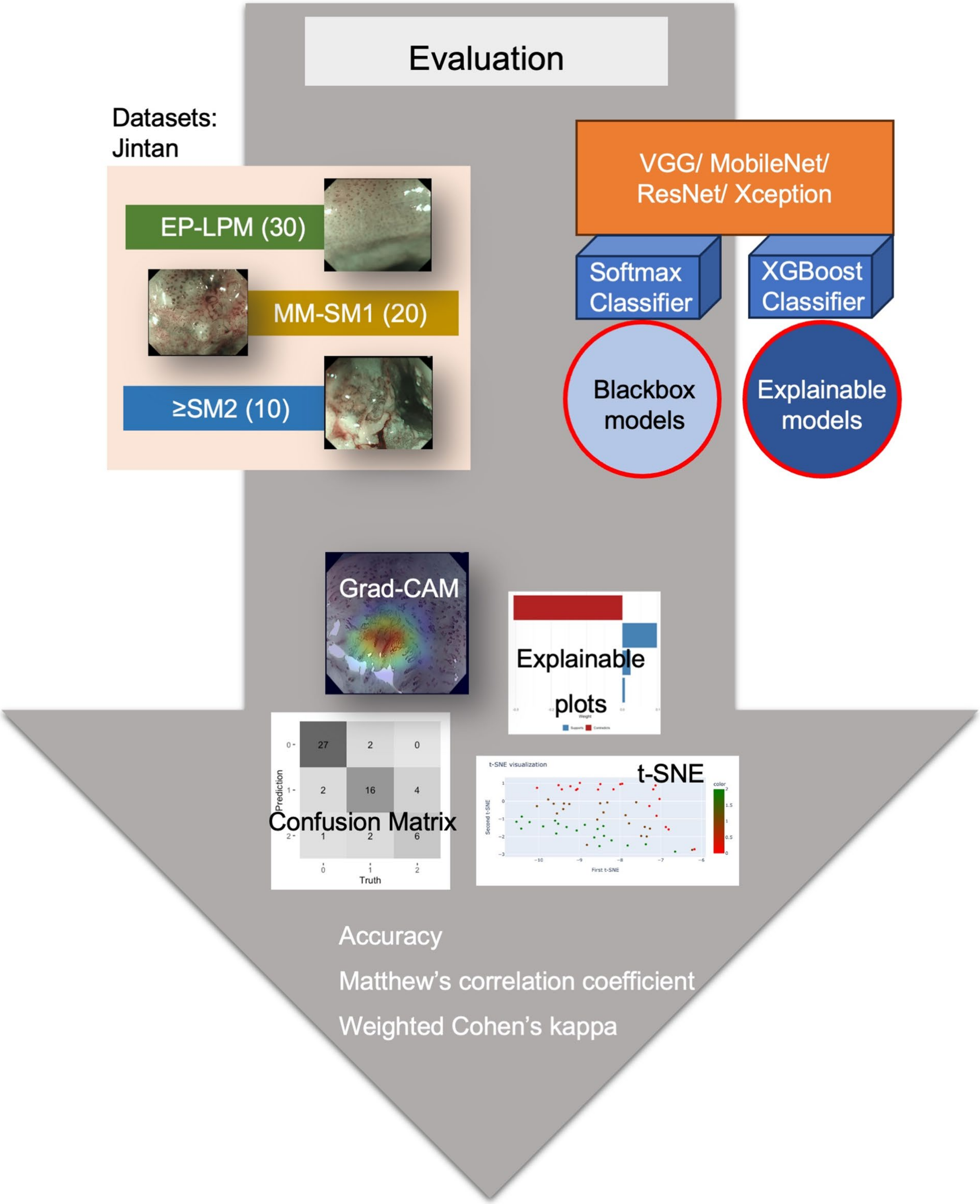

Fig. 4. The flowchart of Step #3 test. Models were evaluated on an external test dataset and compared with endoscopists. The metrics included accuracy, Matthew correlation coefficient (MCC), and weighted Cohen's kappa. Furthermore, for visualized explanation, Grad-CAM was conducted for computer vision models on endoscopic images; local interpretable model-agnostic explanation (LIME), SHapley Additive exPlanations (SHAP), partial dependence plots (PDP) were conducted for the XGBoost classifier; and t-SNE was conducted for visualize feature vectors (blackbox models vs. explainable models) in a two-dimensional space.

Models / Endoscopists		Accuracy	MCC	Weighted Cohen's kappa
Blackbox models	VGG16	0.617	0.405	0.480 [0.270–0.690]
	MobileNet	0.650	0.454	0.470 [0.250–0.690]
	ResNet50	0.733	0.573	0.600 [0.400–0.790]
	Xception	0.767	0.626	0.710 [0.570–0.860]
Explainable models	VGG16	0.700	0.531	0.660 [0.510–0.800]
	MobileNet	0.750	0.599	0.680 [0.530–0.830]
	ResNet50	0.783	0.651	0.720 [0.570–0.870]
	Xception	**0.817**	**0.701**	**0.780 [0.680–0.880]**
Junior endoscopist		0.733	0.571	0.750 [0.680–0.820]
Senior endoscopist		**0.883**	**0.809**	**0.890 [0.880–0.900]**
Junior endoscopist + AI		0.833	0.728	0.850 [0.820–0.870]
Senior endoscopist + AI		**0.917**	**0.865**	**0.920 [0.910–0.930]**

Table 2. The performance of models and endoscopists in the evaluation. MCC, Matthew's correlation coefficient.

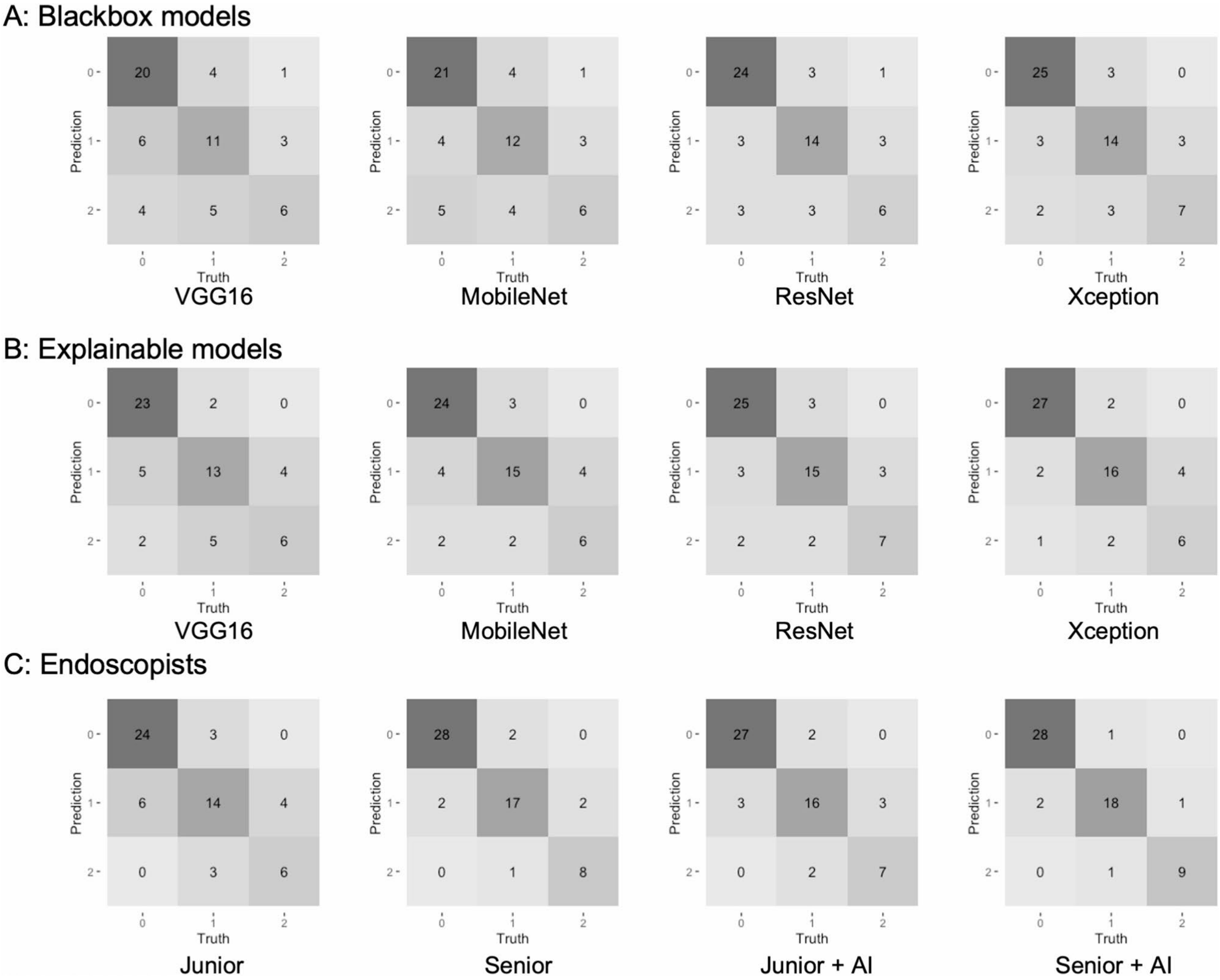

Fig. 5. The confusion matrices of the models and endoscopists in the test dataset. (**A**) blackbox models. (**B**) explainable models. (**C**) endoscopists and AI-assisted endoscopists.

Performance of AI-assisted endoscopists

In awareness of Xception-backboned explainable model's prediction, the performance of endoscopists were improved. The performance (accuracy) of the senior arrived at 0.917, which improved 3.85%. In the meantime, the junior's accuracy arrived at 0.833, which improved 13.64%.

Visualized interpretation of the models

As shown in Fig. 6, t-SNE visualization of feature vectors revealed distinct clustering patterns between models. For the blackbox model, classes exhibited partial overlap, particularly between EP-LPM and MM-SM1, suggesting ambiguity in distinguishing early invasive depths. In contrast, the explainable model produced

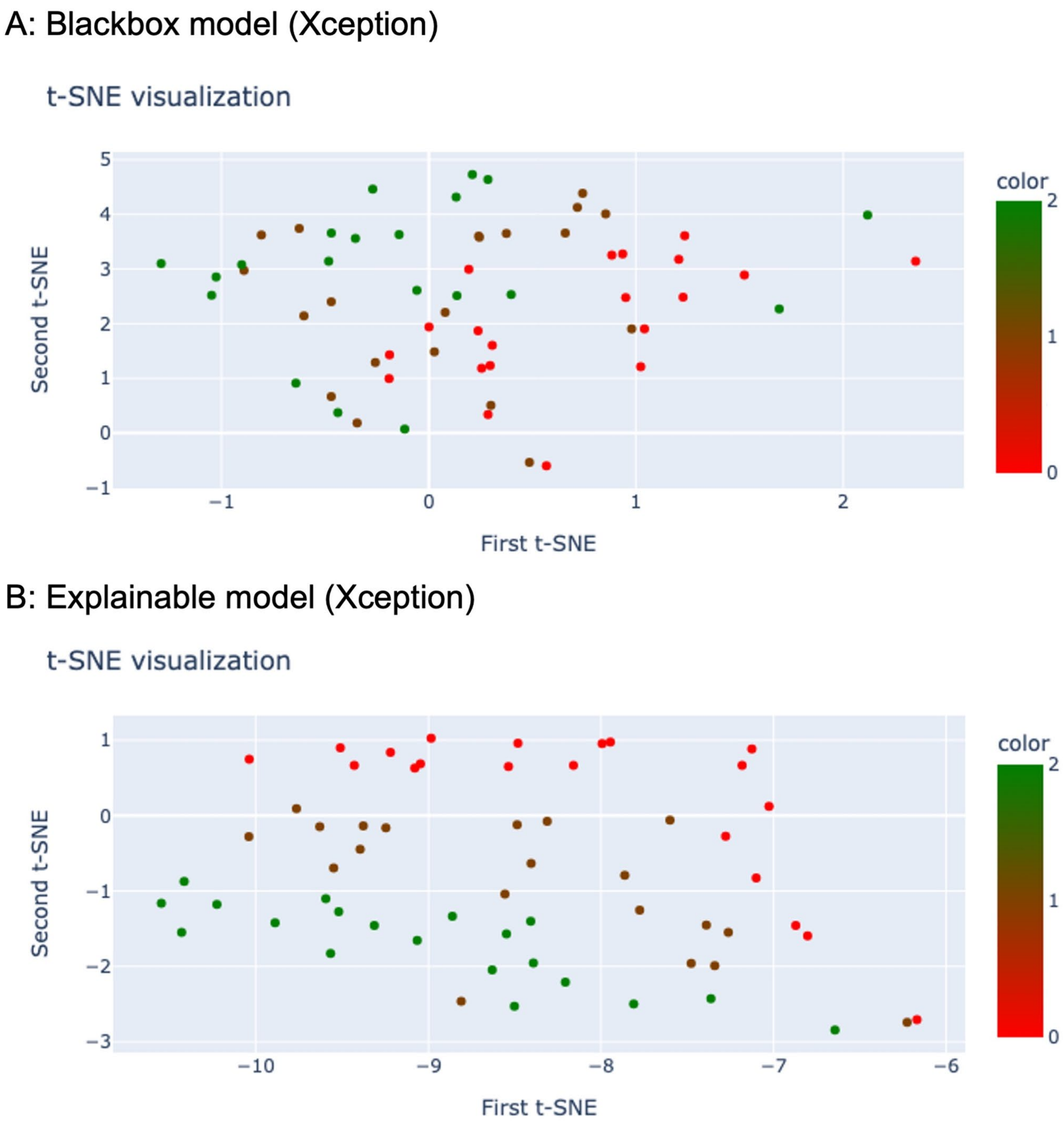

Fig. 6. t-SNE visualization of feature vectors in models in the test. t-SNE is an unsupervised machine learning algorithm for dimensionality reduction. In the study, it was used to map high-dimensional data to a two-dimensional space. Each point represents an image, and the distance between points reflects the similarity between images in the reduced space. (**A**) Xception-backboned blackbox model. (**B**) Xception-backboned explainable model.

markedly separable clusters. The findings confirm that integrating IPCL/AVA patterns during fine-tuning enforces pathologically relevant feature learning, reducing diagnostic uncertainty in borderline lesions.

Figure 7 was to visualize the association between the IPCL/AVA patterns and the prediction of the XGBoost classifier within the explainable AI.

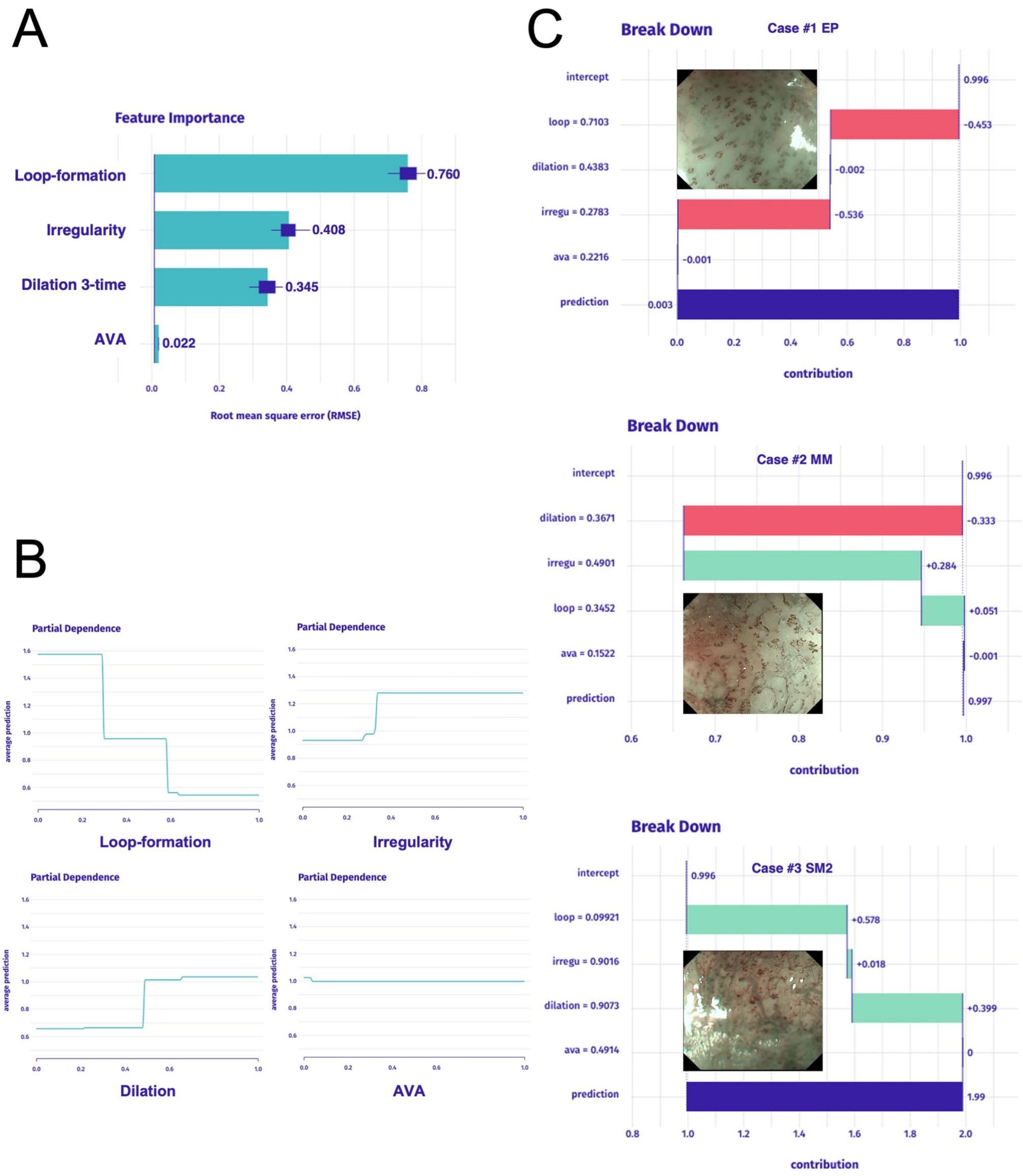

Fig. 7. Visualized explanation for XGBoost classifier in the explainable model. (**A**) feature importance plots; (**B**) partial dependence plots (PDP); (**C**) local Interpretation plots. The feature importance plotting indicated the general association between the IPCL/AVA features and the prediction, as well as the PDPs. The local Interpretation plots reflected the local association between the IPCL/AVA features and the prediction within individual cases.

In Fig. 7A, the feature importance indicated the general association between the IPCL/AVA patterns and the prediction, as well as the PDPs in Fig. 7B. They showed that loop-formation significantly contributed to invasion depth, as well as irregularity and dilation. However, the contribution of AVA was non-significant. The break down plotting (local interpretation, Fig. 7C) reflected the contribution of features to the prediction, within individual cases. In the case #1 with invasion depth EP, XGBoost XAI model's prediction was 0.003 (category 0 = EP-LPM). The most important variable was irregularity (prediction value = 0.278), which decreased the general prediction of XGBoost by 0.536. The second and third most important variables were loop-formation (0.710) and dilation (0.438), which decreased the prediction by 0.453 and 0.002. In the case #2 (invasion depth MM), XGBoost prediction was 0.997 (category 1 = MM-SM1). Dilation (0.367), irregularity (0.490) and loop-formation (0.345) were the key features for the prediction. Their contribution to the prediction were − 0.333, + 0.284 and + 0.051. Moreover, in the case #3 (invasion depth SM2), XGBoost prediction was 1.990 (category 2 = ≥ SM2). Loop-formation (0.099), Dilation (0.907) and irregularity (0.902) increased the prediction by 0.578, 0.399 and 0.018.

Lastly, based on the outputs of the four feature models within Xception-backboned explainable AI, Grad-CAM was conducted for inferential explanation as shown in Fig. 8. The highlighted areas of the four feature models were their inferential evidence.

Discussion

In this study, we presented explainable semi-supervised models developed for predicting invasion depth of ESCC based on the IPCL/AVA patterns. The novel framework empowers AI models to achieve improved transparentness and performance, facing the opacity of traditional supervised learning and limited amounts of labelled endoscopic images.

Deep supervised learning algorithms are often contingent upon a substantial corpus of labeled data to attain optimal performance levels[23]. Nonetheless, the assembly and annotation of such datasets can entail considerable financial and temporal expenditures. SSL emerges as a niche within the unsupervised learning spectrum, dedicated to the extraction of informative features from data that lacks human-provided labels[24]. As a prevalent approach within SSL, contrastive learning facilitates the training of encoders on expansive datasets that are devoid of labels. It operates by enhancing the congruence between varied, augmented perspectives of identical data instances, within the latent space, through the optimization of a contrastive loss function[19].

The imperative for explainable AI is driven by the inherent complexity and obscurity of traditional AI models, which frequently operate as impenetrable black boxes[25]. They generate predictions grounded in input data yet fail to elucidate the rationale underpinning these forecasts. The elucidation algorithm serves as the pivotal element within explainable AI, tasked with furnishing clarity and revealing the salient and impactful factors that inform the model's predictive outcomes. This mechanism can draw upon diverse methodologies within explainable AI, encompassing techniques such as feature significance, contribution assessment, and data visualization, thereby imparting profound insights into the inner workings of machine learning models[26].

The subtle endoscopic manifestations of early ESCC lesions are frequently overlooked, with literature indicating a considerable miss rate for upper gastrointestinal tract cancers during endoscopic examinations. The precise identification of ESCC lesions is essential for the prediction histology and invasion depth and consequently guides therapeutic interventions[6]. Mucosal lesions exhibit a low propensity for local lymph node metastasis (less than 2%) when compared to those that have invaded the submucosa (ranging from 8 to 45.9%), making them suitable candidates for endoscopic treatment[2,27]. In accordance with guidelines from Japan and Europe, lesions confined to the EP-LPM are clear indications for ER, while those invading MM-SM1 are considered relative indications. For lesions with ≥ SM2, esophagectomy or chemoradiotherapy is the recommended course of treatment[5].

A variety of endoscopic classification criteria have been proposed for the diagnosis of ESCC, including mucosal surface characteristics, the JES's classification based on IPCL and AVA[5,6]. Within this classification system, Type A vessels are indicative of normal mucosa or low-grade intraepithelial neoplasia, while Type B1, B2, and B3 vessels suggest progression to high-grade intraepithelial neoplasia or invasion into EP-LPM, MM-SM1 and ≥ SM2, respectively. However, the patterns of IPCL and AVA are highly dependent on the expertise of endoscopists and is subject to interobserver variability. Thus, there is a need for computer-aided diagnostic (CADx) approaches that can reduce the complexity and variability inherent in IPCL/AVA classification.

The past five years witnessed a series of AI studies concerned deep learning in endoscopic diagnosis and detection of ESCC. In 2019, Everson et al.[28] collected a dataset comprising 7046 ME-NBI images from 17 subjects, including 10 with ESCC and 7 controls, to train a convolutional neural networks (CNN) model to classify 2-way IPCL patterns. It achieved a high level of accuracy, correctly distinguishing abnormal from normal IPCL patterns in 93.7% of cases. In 2020, Fukuda et al.[29] developed a CADx system for differentiating cancerous from non-cancerous SCC on NBI/BLI images and reported an accuracy rate of 88%. Similarly, a CADx system by Guo et al.[30] reported remarkable sensitivity (98.04%) and specificity (95.03%) in endoscopic NBI images. Tokai et al. developed an AI-diagnostic system to determine the invasion depth of ESCC. This system analyzed 279 white-light images, accurately estimating the invasion depth of ESCC with a sensitivity of 84.1% and an overall accuracy of 80.9% within 6 s. Uema et al.[31] constructed a ResNeXt-101 backboned model to classify microvessels in ESCC. With a dataset of 2,524 ME-NBI images encompassing 393 lesions, the system achieved a microvessel classification accuracy of 84.2%, surpassing the average accuracy of eight endoscopists (77.8%). Wang and colleagues[32] proposed an AI-assisted endoscopic diagnostic approach for the detection and localization of IPCLs in early-stage ESCC using ME-BLI and ME-NBI images. They employed an enhanced Faster region-based CNN with a polarized self-attention-HRNetV2p backbone for automatic IPCL detection. The methods showed promising results with a recall of 79.25%, precision of 75.54%, F1-score of 0.764, and a mean average precision of 74.95%. In the meantime, Zhang et al.[33] collected 5,119 ME-NBI images from 581 ESCC patients and developed

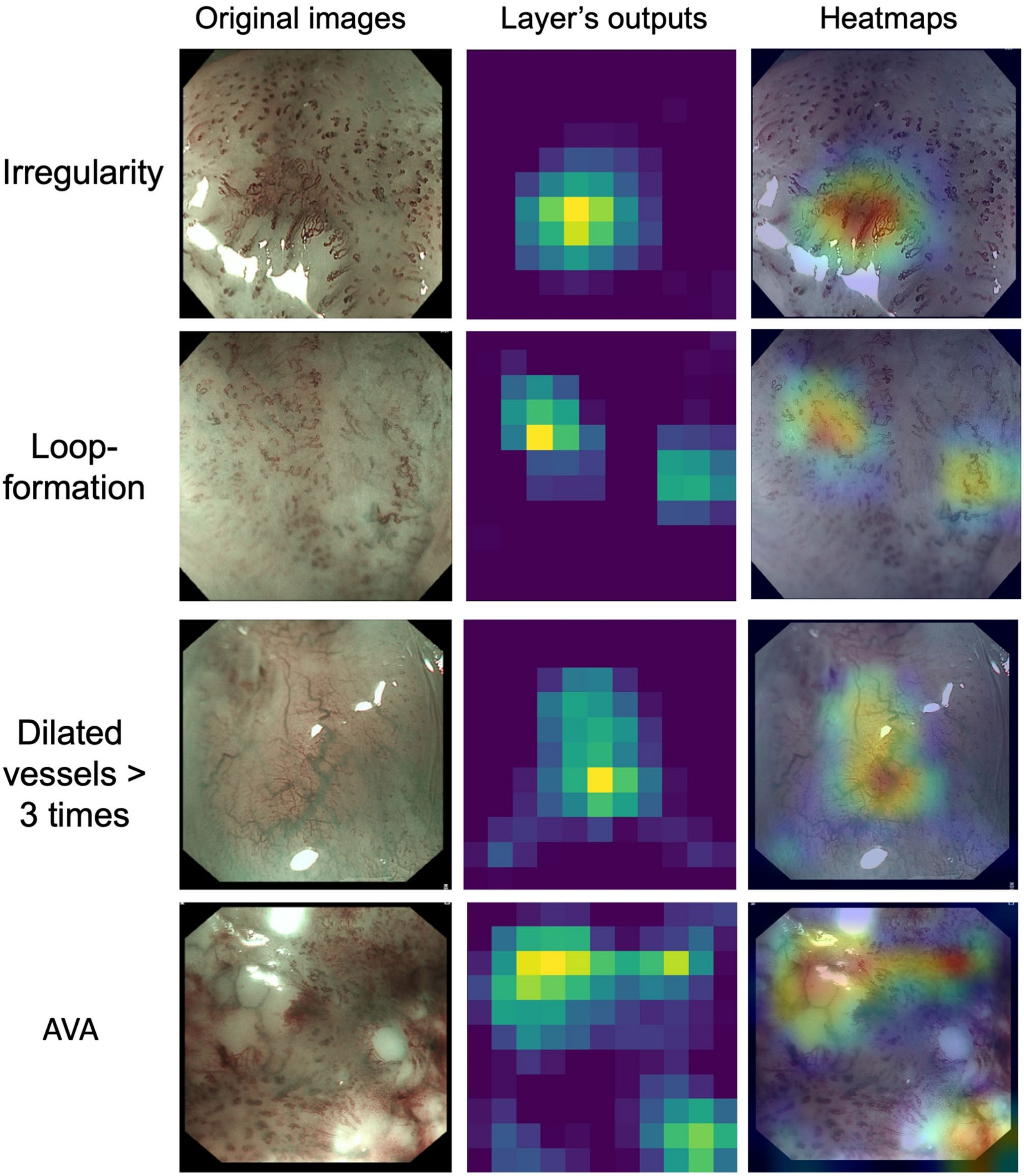

Fig. 8. Visualized inference of four feature models within the explainable fine-tuning via Grad-CAM. Left column: the original endoscopic images; Middle column: heatmaps based on the outputs of the feature models' last layer; Right column: the Grad-CAM heatmap covering the original images, highlighting inferential evidence of the models.

a multi-model diagnostic system for feature extraction and integration. This diagnostic system, grounded in a variety of endoscopic diagnostic methods, outperformed traditional DL models and endoscopists, achieving sensitivity, specificity, and accuracy rates of 85.7%, 86.3%, and 86.2% in image validation, and 87.5%, 84%, and 84.9% in consecutive video analysis, respectively, for distinguishing SM2-3 lesions.

The study has some limitation. To begin with, the dataset employed for training and testing was of insufficient size, which may undermine the robustness and generalizability of the findings. Second, methodological diversity and dataset heterogeneity hinder comparative analysis with the previous reports. Future efforts should prioritize standardized evaluation frameworks and multicenter collaborative datasets to enable robust benchmarking and clinical translation.

The study introduces a novel framework that combines semi-supervised learning with explainable AI to address the challenges of data scarcity and model interpretability in endoscopic assessment of ESCC. By leveraging semi-supervised learning, the model reduces its reliance on large labeled datasets, effectively utilizing abundant unlabeled ME-NBI images to enhance performance. This approach maintains competitive accuracy while providing interpretable predictions, addressing the traditional “blackbox” critique of deep learning models. Clinically, the model demonstrates real-world utility by significantly enhancing endoscopists’ diagnostic accuracy, aligning with treatment guidelines for ESCC stratification. Technically, this study pioneers the integration of self-supervised contrastive learning with multi-feature explainable AI for ESCC invasion prediction, introducing innovative visualization methods such as t-SNE for feature clustering and Grad-CAM for region-of-interest localization, tailored specifically to endoscopic IPCL/AVA patterns. These advancements collectively position the model as a powerful tool for improving diagnostic accuracy and trust in AI-driven endoscopic practices.

Data availability

The code used to train self-supervised models can be found in an open-accessed website (https://osf.io/t3g8n).

Received: 9 November 2024; Accepted: 6 June 2025
Published online: 02 July 2025

References

1. Morgan, E. et al. The global landscape of esophageal squamous cell carcinoma and esophageal adenocarcinoma incidence and mortality in 2020 and projections to 2040: new estimates from GLOBOCAN 2020. *Gastroenterology* **163**(e642), 649–658. https://doi.org/10.1053/j.gastro.2022.05.054 (2022).
2. Deboever, N., Jones, C. M., Yamashita, K., Ajani, J. A. & Hofstetter, W. L. Advances in diagnosis and management of cancer of the esophagus. *BMJ* **385**, e074962. https://doi.org/10.1136/bmj-2023-074962 (2024).
3. Zhang, M., Gong, L., Chen, Y., Ding, R. & Yang, Z. Disease burden of esophageal cancer attributable to low fruit intake in China and globally from 1990 to 2019. *Am. J. Transl Res.* **16**, 3182–3190. https://doi.org/10.62347/FEFU5237 (2024).
4. Wong, M. C. S. et al. Performance of screening tests for esophageal squamous cell carcinoma: a systematic review and meta-analysis. *Gastrointest Endosc* 96, 197–207 e134, (2022). https://doi.org/10.1016/j.gie.2022.04.005
5. Santi, E. G. et al. Microvascular caliber changes in intramucosal and submucosally invasive esophageal cancer. *Endoscopy* **45**, 585–588. https://doi.org/10.1055/s-0033-1344228 (2013).
6. Oyama, T. et al. Prediction of the invasion depth of superficial squamous cell carcinoma based on microvessel morphology: magnifying endoscopic classification of the Japan esophageal society. *Esophagus* **14**, 105–112. https://doi.org/10.1007/s10388-016-0527-7 (2017).
7. Fan, M. et al. Exploration of an effective training system for diagnosis of superficial esophageal squamous cell carcinoma with magnifying narrow-band imaging: prospective research. *Dig. Endosc.* **33**, 770–779. https://doi.org/10.1111/den.13865 (2021).
8. Sanchez-Peralta, L. F., Bote-Curiel, L., Picon, A., Sanchez-Margallo, F. M. & Pagador, J. B. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artif. Intell. Med.* **108**, 101923. https://doi.org/10.1016/j.artmed.2020.101923 (2020).
9. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60. https://doi.org/10.1038/s41586-023-06221-2 (2023).
10. Chen, Y., Mancini, M., Zhu, X. & Akata, Z. Semi-Supervised and unsupervised deep visual learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell PP*. https://doi.org/10.1109/TPAMI.2022.3201576 (2022).
11. Huang, S. C. et al. Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *NPJ Digit. Med.* **6**, 74. https://doi.org/10.1038/s41746-023-00811-0 (2023).
12. Park, S., Lee, E. S., Shin, K. S., Lee, J. E. & Ye, J. C. Self-supervised multi-modal training from uncurated images and reports enables monitoring AI in radiology. *Med. Image Anal.* **91**, 103021. https://doi.org/10.1016/j.media.2023.103021 (2024).
13. Neri, E. et al. Explainable AI in radiology: a white paper of the Italian society of medical and interventional radiology. *Radiol. Med.* **128**, 755–764. https://doi.org/10.1007/s11547-023-01634-5 (2023).
14. Yang, G., Ye, Q. & Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion.* **77**, 29–52. https://doi.org/10.1016/j.inffus.2021.07.016 (2022).
15. Kalyakulina, A., Yusipov, I., Moskalev, A., Franceschi, C. & Ivanchenko, M. eXplainable artificial intelligence (XAI) in aging clock models. *Ageing Res. Rev.* **93**, 102144. https://doi.org/10.1016/j.arr.2023.102144 (2024).
16. de Pahud, A. et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *NPJ Digit. Med.* **7**, 195. https://doi.org/10.1038/s41746-024-01190-w (2024).
17. Yin, M. et al. Identification of gastric signet ring cell carcinoma based on endoscopic images using few-shot learning. *Dig. Liver Dis.* https://doi.org/10.1016/j.dld.2023.07.005 (2023).
18. Wang, Y. et al. A Semi-Supervised learning framework for classifying colorectal neoplasia based on the NICE classification. *J. Imaging Inf. Med.* https://doi.org/10.1007/s10278-024-01123-9 (2024).
19. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. E. A simple framework for contrastive learning of visual representations. *ArXiv* (2020). abs/2002.05709.
20. Wang, Y. et al. Automated multimodal machine learning for esophageal variceal bleeding prediction based on endoscopy and structured data. *J. Digit. Imaging.* **36**, 326–338. https://doi.org/10.1007/s10278-022-00724-6 (2023).
21. Yu, C. et al. Automated machine learning in predicting 30-Day mortality in patients with Non-Cholestatic cirrhosis. *J. Pers. Med.* **12** https://doi.org/10.3390/jpm12111930 (2022).
22. Liu, J. & Vinck, M. Improved visualization of high-dimensional data using the distance-of-distance transformation. *PLoS Comput. Biol.* **18**, e1010764. https://doi.org/10.1371/journal.pcbi.1010764 (2022).
23. Yin, M. et al. Development and validation of a multimodal model in predicting severe acute pancreatitis based on radiomics and deep learning. *Int. J. Med. Inf.* **184**, 105341. https://doi.org/10.1016/j.ijmedinf.2024.105341 (2024).
24. Azizi, S. et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat. Biomed. Eng.* **7**, 756–779. https://doi.org/10.1038/s41551-023-01049-7 (2023).
25. Klauschen, F. et al. Toward explainable artificial intelligence for precision pathology. *Annu. Rev. Pathol.* **19**, 541–570. https://doi.org/10.1146/annurev-pathmechdis-051222-113147 (2024).
26. Joyce, D. W., Kormilitzin, A., Smith, K. A. & Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digit. Med.* **6**, 6. https://doi.org/10.1038/s41746-023-00751-9 (2023).

27. Tanaka, I. et al. The sub-classification of type B2 vessels according to the magnifying endoscopic classification of the Japan esophageal society. *Dig. Endosc.* **32**, 49–55. https://doi.org/10.1111/den.13459 (2020).
28. Everson, M. et al. Artificial intelligence for the real-time classification of intrapapillary capillary loop patterns in the endoscopic diagnosis of early oesophageal squamous cell carcinoma: A proof-of-concept study. *United Eur. Gastroenterol. J.* **7**, 297–306. https://doi.org/10.1177/2050640618821800 (2019).
29. Fukuda, H. et al. Comparison of performances of artificial intelligence versus expert endoscopists for real-time assisted diagnosis of esophageal squamous cell carcinoma (with video). *Gastrointest. Endosc.* **92**, 848–855. https://doi.org/10.1016/j.gie.2020.05.043 (2020).
30. Guo, L. et al. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). *Gastrointest. Endosc.* **91**, 41–51. https://doi.org/10.1016/j.gie.2019.08.018 (2020).
31. Uema, R. et al. Use of a convolutional neural network for classifying microvessels of superficial esophageal squamous cell carcinomas. *J. Gastroenterol. Hepatol.* **36**, 2239–2246. https://doi.org/10.1111/jgh.15479 (2021).
32. Wang, J. et al. AI-assisted identification of intrapapillary capillary loops in magnification endoscopy for diagnosing early-stage esophageal squamous cell carcinoma: a preliminary study. *Med. Biol. Eng. Comput.* **61**, 1631–1648. https://doi.org/10.1007/s11517-023-02777-3 (2023).
33. Zhang, L. et al. Human-Like artificial intelligent system for predicting invasion depth of esophageal squamous cell carcinoma using magnifying Narrow-Band imaging endoscopy: A retrospective multicenter study. *Clin. Transl Gastroenterol.* **14**, e00606. https://doi.org/10.14309/ctg.0000000000000606 (2023).

Acknowledgements

The authors thank all the participating patients and investigators.

Author contributions

Kang L., Ni H., Lin J. and Zhu S. contributed to manuscript drafting and data analysis. Kang L., Ni H. and Wang Y. contributed to data acquisition. Li R., Liu L. and Shi X. contributed to manuscript revision. Kang L., Li R. and Zhu K. contributed to endoscopic images reading. Li R., Zhu J. and Shi X. contributed to the design of the work. All authors contributed to final approval of the completed version.

Funding

This study was supported by Frontier Technologies of Science and Technology Projects of Changzhou Municipal Health Commission (QY202309), Medical Education Collaborative Innovation Fund of Jiangsu University (No. JDY2022018), Suzhou Health Committee (DZXYJ202301), Joint construction by four parties (ML12202323, ML12203223), Clinical research program of the first affiliated hospital of Soochow University (BXLC003), and Science and Technology Plan of Jiangsu (BE2023710).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This retrospective study was approved by the ethics committee of the First Affiliated Hospital of Soochow University (Approval number 2022098).

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-06172-w.

Correspondence and requests for materials should be addressed to X.S. or R.L.

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.