ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research on recognition and detection technology of gastrointestinal precancerous lesions in endoscopic images based on deep image learning algorithm

Xueying Li[1,*]

[1] China Medical University, Shenyang, Liaoning, 210100, China

SUMMARY: *Gastrointestinal cancer, as a highly prevalent malignant disease worldwide, early and accurate identification of precancerous lesions is the key to improve the survival rate of patients. In this paper, we propose a gastrointestinal precancerous lesion recognition framework that integrates multi-task learning and improved SSD, and improves the model performance through data preprocessing optimization and algorithm innovation. The study is based on white light gastroscopy image data, and a standardized dataset is constructed after pathological biopsy confirmation. For the medical image noise problem, the denoising process of YOLOv3 detection combined with median filtering and Sym4 wavelet transform is adopted. At the algorithmic level, the improved SSD model introduces semantic segmentation branching and recursive pyramid network (RFPN), enhances the shallow semantic expression ability by fusing the Conv4_3, Conv7, and Conv8_2 layers of features, and extracts multi-scale information by combining with the modified receptive field module (RFB), which significantly improves the accuracy of the detection of small lesions. Focal Loss is adopted to alleviate the category imbalance problem, and weighted cross entropy and IoU loss are integrated to achieve multi-task co-optimization. Experiments show that the global fine-tuning strategy is significantly better than the local fine-tuning. r150×3 network, for example, the global fine-tuning has an accuracy of 98.52%, an F1 score of 94.73%, a specificity of 99.44%, and an AUC value of 0.996. Comparing with the traditional models (VGG19, ResNet50, and Inception-V3), the performance of the improved SSD in the identification of the three categories of lesions is the The best, gastric cancer recognition accuracy rate of 98.87%, false judgment rate of only 1.69%, single image recognition time is only 0.05s, efficiency than manual diagnosis to improve a hundred times. The model has the strongest ability to locate bulging lesions, with an accuracy of 90.29% when the overlap degree is ≥60%, and the localization accuracy of flat lesions is lower, which needs to be further optimized for texture feature extraction.*

KEYWORDS: *YOLOv3; endoscopic images; SSD algorithm; semantic information fusion; gastrointestinal cancer lesion recognition*

1 Introduction

Endoscopy is a more common and important examination program in clinical work, with the characteristics of safety, simplicity, reliability and effectiveness [1]. With the increasing level of endoscopic diagnosis, more and more early gastrointestinal tumors have been found, and the ensuing technical methods of endoscopic treatment have become more and more abundant and mature, and have become an important means of treating early gastrointestinal tumors [2-4].

*lxyscientificpaper@163.com
https://doi.org/10.65102/is20261081
Published: 10 June, 2026

Through endoscopic imaging can directly observe and discover the lesions of the whole digestive tract mucosa of the examinee, and it is more capable of clamping and biopsy processing of suspicious lesions for pathological examination and analysis, especially for the diagnosis and differential diagnosis of gastrointestinal tract cancers, which has an important diagnostic value [5-8].

As endoscopy generates huge images and video data, it is necessary for physicians to make accurate and error-free judgment on endoscopic images of all patients within a limited time [9, 10]. However, the manual diagnostic process is susceptible to a variety of factors, such as subjective experience, cognitive bias, time pressure, and emotional state [11, 12]. And the cumbersome image screening work will consume a lot of doctors' energy, which will easily lead to the occurrence of missed diagnosis and misdiagnosis [13]. The development of a deep learning-based artificial intelligence assisted diagnosis system relies on the powerful data processing and analysis capabilities of convolutional neural networks, and realizes the automatic identification and localization of foci in endoscopic images by constructing a deep learning model for the type of esophageal lesions, thus providing intuitive and accurate assisted diagnostic results [14-17]. It is of great significance for reducing the burden of doctors, improving diagnostic accuracy, and optimizing medical resources [18].

In this paper, we propose a framework for recognizing gastrointestinal precancerous lesions that integrates multi-task learning and improved SSD, the core of which includes the construction and preprocessing of datasets, the design of improved SSD detection algorithms, and multi-task learning strategies. It aims to improve the detection performance of the model through data preprocessing optimization and algorithm innovation. First, a standardized dataset is constructed by strictly screening high-quality endoscopic images, and a denoising process combining YOLOv3 target detection and multimodal filtering is proposed to ensure the reliability of the input data in response to the common noise problem in medical images. Secondly, at the level of detection algorithm, for the defects of traditional SSD model with insufficient sensitivity to small targets, semantic segmentation branch and recursive pyramid network (RFPN) are introduced, and by introducing semantic segmentation branch, the features of Conv4_3, Conv7, and Conv8_2 layers are fused, and the multiscale semantic information is extracted by using the modified receptive field module (RFB). And the expression ability of shallow features is enhanced by multi-scale semantic information fusion, which effectively improves the detection accuracy of small lesions. In addition, Focal Loss is used to replace the traditional cross-entropy loss to dynamically adjust the weights of difficult and easy samples to alleviate the category imbalance. Meanwhile, combining the weighted cross-entropy and IoU loss of segmentation task, multi-task co-optimization is realized to further strengthen the model generalization ability.

2 Gastrointestinal precancerous lesion detection method based on improved SSD with multi-task learning

2.1 Construction and denoising preprocessing of endoscopic image dataset for gastrointestinal cancer

2.1.1 Sample case selection

Gastroscopic images of patients who underwent white light gastroscopy from January 2020 to October 2024 at a hospital's gastrointestinal endoscopy center were retrospectively collected, including early gastric cancer images and non-tumor images (including gastric ulcer, chronic

gastritis, and normal images). All endoscopic images were evaluated and reviewed by 2 endoscopists, and some images were taken of the same lesion from different angles, directions, and distances. In this study, we classified early gastric cancer from both endoscopic and pathologic perspectives according to the definition of early gastric cancer staging by the Chinese Anti-Cancer Society as shown in Table 1.

Table 1: The classification of early gastric cancer

Classification	Classification	Classification description
Endoscopic classification	Type One	Protuberant early gastric cancer
	Type Two	Flat early gastric cancer, flat protruding early gastric cancer, and flat depressed early gastric cancer
	Three types	Depressed early gastric cancer
Pathological classification	Adenocarcinoma	Among them, adenocarcinoma can be further classified into tubular adenocarcinoma and mucinous adenocarcinoma
	Squamous cell carcinoma	Squamous cell carcinoma
	Medullary carcinoma	Relatively rare
	Adenosquamous carcinoma	Cancer cells contain both adenocarcinoma components and squamous cell carcinoma components

Among the inclusion criteria for the sample data were:

(1) All lesions were diagnosed by pathologic biopsy or surgical pathology, while the extent of the lesion was clearly defined.

(2) All were white light, non-magnified gastroscopic images.

Exclusion criteria were:

(1) No pathologic findings.

(2) Low-quality or low-resolution images that do not allow identification of tissue structures, such as post-biopsy hemorrhage, underinflation, extensive motion artifacts, excessively dark backgrounds, and large amounts of mucus and food residue.

(3) Non-gastroscopic images, such as those produced by devices such as magnifying endoscopes and microendoscopes. All gastroscopic images in this article were captured and recorded by endoscopes models CV-70, CV-240, CV-260, and CV-290 (Olympus Optical Co., Ltd., Tokyo, Japan) and endoscopes models EVE400, SYSTEM4400, and SP702 (Fuji).

2.1.2 Image denoising

Image denoising is the process of removing various noises contained in a digital image is called image denoising. Noise is usually introduced in images during medical imaging due to the imaging mechanism as well as the shooting angle. Noise may be generated in the process of image acquisition, storage and encryption, etc. These noises make the boundaries of the target organization in the image blurred, and some of the subtle features and structures cannot be identified by the noise interference, which greatly reduces the quality of the image, affects the medical diagnosis, and has a great impact on the model of convolutional neural network, resulting in a decline in the performance of the model. Therefore, in medical image denoising sound processing, it is very important to effectively remove the various noises contained in the image noise while retaining the structure and boundary features of the tissue intact.

The noise can be divided into different types such as pretzel noise, Gaussian noise, etc. according to the distribution characteristics. For different types of noise, different denoising processing algorithms need to be used, and common denoising methods are as follows:

Median filtering is a nonlinear filtering technique that is commonly used to remove solitary point noise and is outstanding in maintaining the edge characteristics of the image. It sets the values of all pixel points in an image to the median of all pixel values within a specified size window centered on itself.

Gaussian filtering is a filtering technique that selects weights based on a Gaussian function, which can effectively deal with obeying random and normal distribution noise. In layman's terms, Gaussian filtering is a weighted average of each pixel point in an image and other pixels in its domain using weights determined by a Gaussian function.

There are mainly four types of noise in white light gastroscopy images, namely, numbers, text, intubation tube and black background, which have a great impact on the training of the model, so the image denoising process is needed. The processing flow of image denoising in this paper mainly includes two sub-processes: image cropping and image noise removal. For the noise in the image, this paper first adopts the YOLOv3 algorithm to realize the detection of the noise, and obtains the coordinate position of the noise in the bounding box of the whole image $P_i = \left(x_{ti}, y_{ti}, x_{bi}, y_{bi} \right)$, where x and y respectively denote the horizontal axis coordinates and vertical axis coordinates of the image, x_t and y_t represent the upper-left position of the noise bounding box, and x_b and y_b represent the lower-right position of the noise bounding box; and then the part of the image that contains the noise is cropped by a cropping algorithm, which adopts a cropping algorithm that retains the image region as large as possible without affecting the core region.

After cropping the image, for the noise left in the image, this paper mainly adopts the median filtering method and Sym4 wavelet transform algorithm to filter the noise in the image. First of all, the median filter is used to initially process the noise in the image, in which the filter window is 5*5; then the processed image is transformed with the Sym4 algorithm to extract the decomposed coefficients; then the decomposed coefficients are processed with the median filtering function to generate a new matrix; in accordance with the reconstruction algorithm, the image is reconstructed with the generated matrix, and then finally the wavelet thresholding is used for the reconstructed image to carry out Finally, the wavelet threshold is used to denoise the reconstructed image to generate the final denoised image.

2.2 Gastrointestinal Cancer Recognition Algorithm Based on Improved SSD

This section describes the detailed methodology of the improved SSD-based gastrointestinal cancer detection network proposed in this paper. The overall architecture of the model is shown in Fig. 1, based on the original SSD detection framework, a recursive pyramid network (RFPN) is added to facilitate the fusion of different semantic features. And a segmentation branch is proposed without changing the backbone network, which is used to do a multi-task cancer recognition system at the same time to assist small target detection. Specific details will be presented in the following sections in turn.

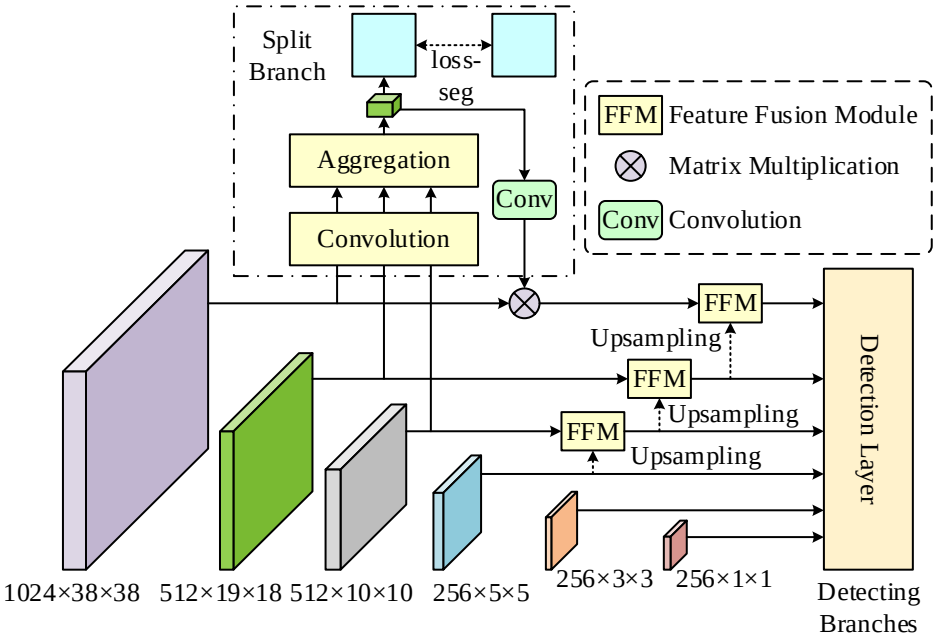

Figure 1: Improved SSD model

2.2.1 Semantic information fusion to assist small target detection

The original SSD framework uses six different scales of features to detect targets of different sizes. Although this method is much better than using only one size feature to do the detection uniformly, it often misses or misdetects the detection of small target objects. Therefore, in this paper, a semantic segmentation branch is added to the detection framework, in addition to using the mask information obtained from the segmentation branch to assist in comparing the detection results, we also propose a novel fusion semantic information detection method, which will utilize the strong semantic information from the semantic segmentation branch fused to the shallow features (Conv4_3) to better and more accurately detect the small target objects.

In order to reduce the number of parameters to achieve real-time stomach cancer recognition, a partial decoder is used to fuse only the features of Conv4_3, Conv7, and Conv8_2 as a segmentation branch, and the segmentation branch network structure is shown in Fig. 2. The feature maps after Conv8_2 are not utilized, mainly because the size of the feature maps after Conv8_2 becomes very small, which does not contribute much to get a good segmentation result. The semantic segmentation branch will eventually get the result of stomach cancer segmentation and will calculate the loss with real labels as part of the loss function of the whole network. In order to further utilize the semantic information brought by the semantic segmentation branch, the semantic features of the segmentation branch are used to incorporate them into the first detection layer, which is Conv4_3.

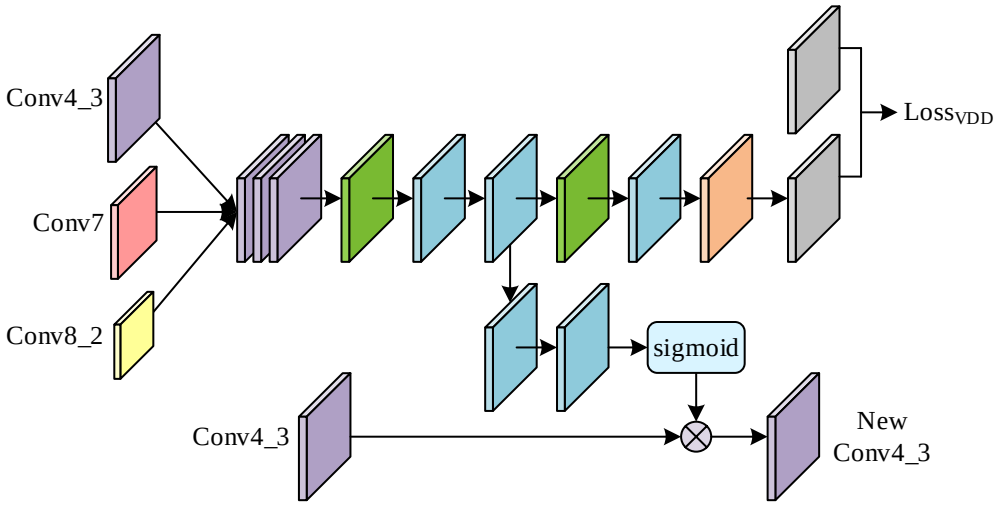

Figure 2: Divide the branch network structure

In the semantic segmentation branch, the original feature maps of Conv4_3, Conv_7, and Conv8_2 are first input into the modified Receptive Field Block (RFB), which simulates the human visual perception system and improves the robustness and discriminative properties of the features. The modified RFB module is shown in Fig. 3, which has four branches in the upper part of the module, utilizing different convolution kernels and null rate sizes to obtain multi-scale information, and finally stitching these four $32\times H\times W$ feature maps together to obtain a 128×H×W feature map. This feature is then directly summed with the initial feature obtained by a 1×1 convolution to obtain a feature map with the same size and the number of channels all changed to 32. Then the new feature map generated by inputting the Conv4_3, Conv_7, and Conv8_2 layers of features to the modified RFB module is stitched together, and the final gastric cancer segmentation mask is obtained after a number of convolution and pooling layers.

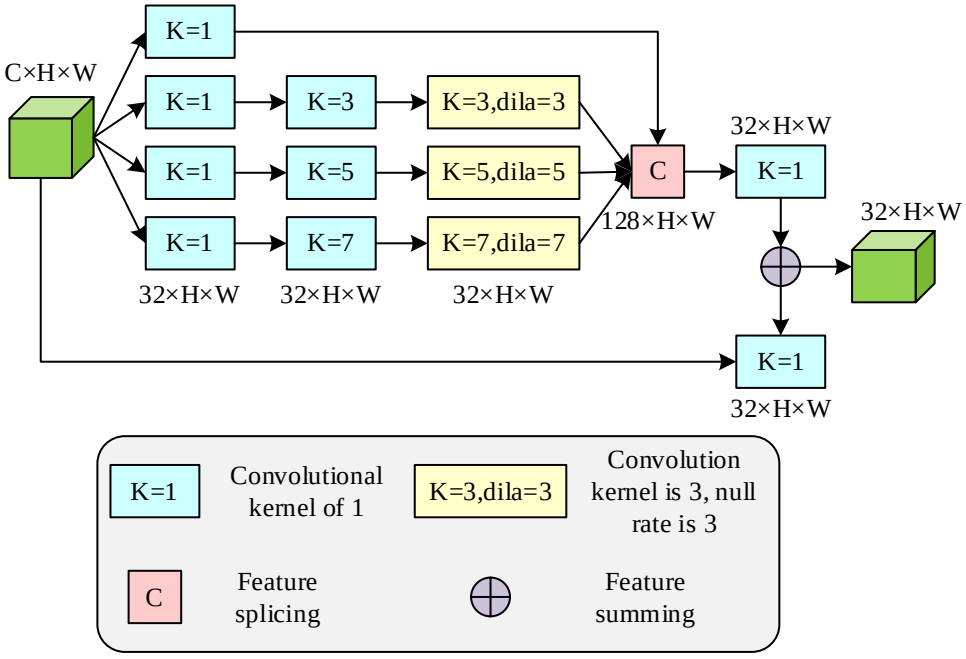

Figure 3: Modified RFB module

The feature maps obtained from the split branch acquire high-level semantic information because the split branch mask supervision allows each pixel point to be assigned to a category, in this paper's task cancer or normal binary classification. In feature maps deeper than Conv4_3, sufficient semantic information has already been learned from the previous feature maps, so there is no need to include the semantic information obtained from the segmentation branch. In addition, the size of the higher-level feature graphs is getting smaller and smaller, and it becomes more and more difficult to add a segmentation branch to the higher-level features. Therefore, only the segmentation branch is added to Conv4_3 and the semantic information is added to be fused into it to assist the target detection at this layer, i.e., the semantic feature maps are used to activate the original low-level detection feature maps, i.e., a pixel-by-pixel multiplication is done, as shown in Eq:

$$feature_{map_{Conv4_3}} = feature_{map_{Conv4_3}} \odot attention_{seg} \tag{1}$$

where $\odot$ is a pixel-by-pixel dot product operation, and $attention_{seg}$ is the attention weight obtained from the features in front of the last convolutional layer before the semantic segmentation branch to get the mask, which is passed through a convolutional layer and a Sigmoid function. By dot product operation, a new feature map $feature\_map_{Conv4\_3}$ is obtained that incorporates semantic information and has low-level features such as texture, shape, etc., and then the original feature map is replaced with this semantically-integrated feature map.

By adding a semantic segmentation branch, this network is turned into a multi-task learning problem. The information from the semantic segmentation branch can be added to the shallow features for better detection of small targets, and it can also be used as auxiliary information for cancer recognition to compare the detection results as a reference. In addition, after adding the semantic segmentation branch, the number of parameters does not increase significantly because the same backbone network is used to extract features, but better detection results can be obtained.

2.2.2 Loss function

There are two losses in the original SSD, one is the localization loss, which uses the L1 loss function, and the other is the classification loss, which uses the cross-entropy loss. However, a picture with thousands of prior frames, only a few of them are positive example samples. Thus, it brings the problem of too many negative examples, resulting in an imbalance of positive and negative samples. So the original SSD needs to use the method of difficult case mining to control the number of negative samples in the ratio of 1:3. Such methods can control the positive and negative ratio, but the mined samples still ignore the negative samples that are easy to categorize, which is not the best solution. Therefore, this paper adopts a new loss function, Focal Loss, which is an improvement on the original cross-entropy loss function.

$$L_{focal} = \begin{cases} -\alpha(1-\hat{y})^{\gamma}\log(\hat{y}), & y=1 \\ -(1-\alpha)\hat{y}^{\gamma}\log(1-\hat{y}), & y=0 \end{cases} \tag{2}$$

where $\hat{y}$ is the output of the model, γ is used to regulate the rate of weight reduction of simple samples, α is used to balance the weight of positive and negative samples, generally set α to 0.25. It can be seen that the Focal Loss will not completely ignore simple negative samples, but to give a weight to each sample, simple samples low weight, complex and difficult

to distinguish samples high weight, so that the model can be more focused on difficult to classify the model performance. The weight of simple samples is low, and the weight of complex and difficult-to-distinguish samples is high, so that the model can focus more on difficult-to-categorize samples and improve the performance of the model. In addition, the segmentation branch in this paper adopts the weighted cross-entropy loss and IoU loss, so the total loss value is equal to the sum of the segmentation loss and the detection loss value. A hyperparameter α is set in the segmentation loss part to control the weight of segmentation loss. As shown in Eq:

$$L_{seg} = L_{IoU}^{w} + L_{BCE}^{w} \tag{3}$$

$$L_{all} = \alpha L_{seg} + \left(L_{cls} + L_{reg} \right) \tag{4}$$

3 Performance analysis and clinical application of improved SSD model in gastrointestinal cancer lesion detection

Based on the gastrointestinal precancerous lesion detection framework fusing multi-task learning and improved SSD proposed in Chapter 2, this chapter comprehensively analyzes the model's detection capability in complex clinical scenarios through systematic experimental validation and multi-dimensional performance evaluation. By constructing a standardized dataset, designing a multi-scale semantic information fusion strategy, and optimizing the loss function, a methodological foundation is laid for the experimental validation in this chapter.

3.1 Experimental preparation

3.1.1 Data acquisition and evaluation indicators

The data sample, as described in Section 2.1 Case selection, contained white light gastroscopy images from a hospital's Gastrointestinal Endoscopy Center from January 2020 to October 2024, covering 354 cases of early gastric cancer (GC), 437 cases of gastric ulcer (GU), and 488 cases of chronic gastritis (CG), as well as 683 normal samples, for a total of 1,962 samples. All images were confirmed by pathological biopsy or surgical pathology to ensure that the extent of the lesion was clear. Inclusion criteria included white light non-magnified gastroscopy images, excluding images that were low-resolution, blurred, or contained interfering factors (e.g., bleeding, mucus residue).

Five main evaluation metrics were used: accuracy, F1 score, sensitivity and specificity, and AUC.

AUC is the medical term for “area under the curve” and is often used to assess the discriminatory ability of a diagnostic test or predictive model, especially to measure the accuracy of a model by the area under the ROC curve. The closer the AUC value is to 1, the better the model's ability to distinguish between “disease” and “non-disease”.

3.1.2 Experimental setup

In this paper, we use seven backbone network structures based on improved SSDs: r50×1, r100×1, r50×2, r100×2, r50×3, r100×3, and r150×3. All of these network structures utilize the RFPN pyramid network architecture, except that the network's depth (number of layers) and width (number of nodes/units in each layer) are different. The letter r in the network structure represents the RFPN architecture, the number after the letter r represents the depth of the network, and the number after the multiplication sign represents the width of the network. A

width of 1 represents the original width of the RFPN architecture, while a width of 2 means that the number of nodes/units per layer in the original RFPN architecture is increased to two times.

3.2 Global versus local fine-tuning

Training experiments for global and local fine-tuning were conducted for models based on each of the seven backbone networks. Table 2 demonstrates the results of the comparison of the four evaluation metrics of the local and global fine-tuning models on the test set for the seven network structures, and the global fine-tuning model with the same network structure has better evaluation metrics than the local fine-tuning model.

Table 2: Comparison of evaluation indicators of various models in different networks

Network	Local fine-tuning				Global fine-tuning			
structure	Accuracy/%	F1/%	Sensitivity/%	Specificity/%	Accuracy/%	F1/%	Sensitivity/%	Specificity/%
r50×1	92.79	86.19	84.24	95.04	94.88	88.85	88.09	96.58
r100×1	93.15	87.49	85.24	95.28	94.79	90.09	88.56	97.14
r50×2	93.73	87.14	87.43	95.76	95.53	91.84	89.63	97.42
r100×2	94.57	88.22	84.04	96.71	96.80	92.98	90.36	97.38
r50×3	95.22	89.04	83.46	95.84	96.92	92.94	90.72	97.03
r100×3	95.25	89.33	85.35	96.99	97.20	93.22	91.97	98.47
r150×3	96.87	90.78	83.07	97.69	98.52	94.73	92.12	99.44

The data show that the global fine-tuning model significantly outperforms the local fine-tuning model in the vast majority of cases. Taking the network structure r50×1 as an example, the global fine-tuning improved the accuracy from 92.79% to 94.88%, the F1 score from 86.19% to 88.85%, and the sensitivity and specificity from 84.24% and 95.04% to 88.09% and 96.58%, respectively. Similar trends prevail in all network structures, e.g., the global fine-tuning accuracy of r100×3 (97.20%) improves by nearly 2 percentage points from 95.25% for local fine-tuning, the F1 score improves from 89.33% to 93.22%, and the sensitivity and specificity improve from 85.35% and 96.99% to 91.97% and 98.47%, respectively. Particularly noteworthy is that with the increase of network depth and width (e.g., r150×3), the advantage of the global fine-tuning model further expands, and its accuracy, F1 score, and specificity reach 98.52%, 94.73%, and 99.44%, respectively, all of which are the highest values among all networks. In addition, the global fine-tuning strategy excelled in sensitivity, a key medical metric, e.g., the sensitivity of r50×3 and r100×3 improved to 90.72% and 91.97%, respectively, which were significantly better than the 83.46% and 85.35% of local fine-tuning. The only exception is that the 83.07% of the local fine-tuning sensitivity of r150×3 is slightly lower than that of other networks, but it still improves to 92.12% after global fine-tuning, indicating that global parameter optimization can effectively alleviate the feature degradation problem of deep networks. In summary, global fine-tuning significantly improves the model's ability to detect gastrointestinal precancerous lesions by fully adjusting the model parameters, and the comprehensive performance is more robust especially in complex scenarios (e.g., small target detection).

Figure 4 shows the subject operating characteristic (ROC) curves and their area under the curve (AUC) of the local fine-tuning models of the seven network structures on the test set, and Figure 5 shows the ROC curves and their AUCs of the global fine-tuning models of the seven network structures on the test set.The AUCs of the seven networks under the local fine-tuning and the global fine-tuning models are calculated, and the data are pooled into the data, and the AUCs of the seven networks under the local fine-tuning and the global AUCs under the fine-tuning models are shown in Table 3.

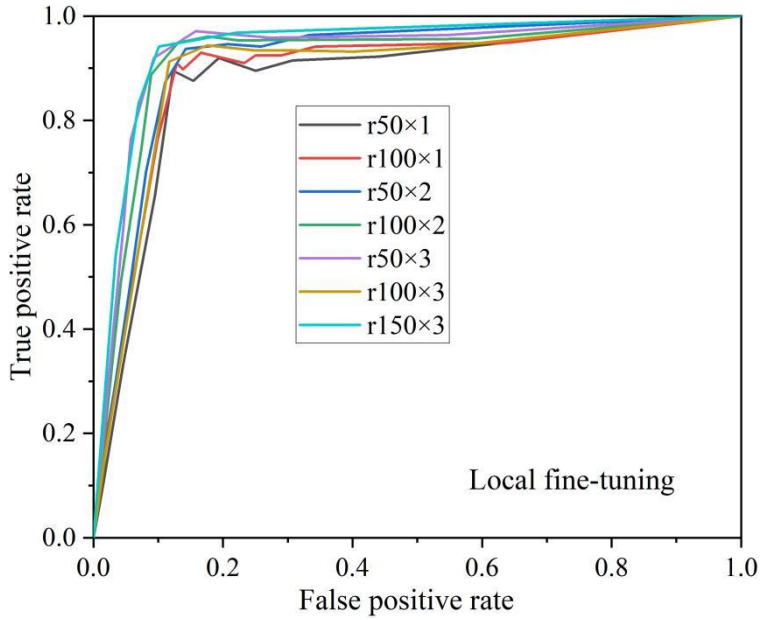

Figure 4: The ROC curves of the local fine-tuning models of 7 network structures

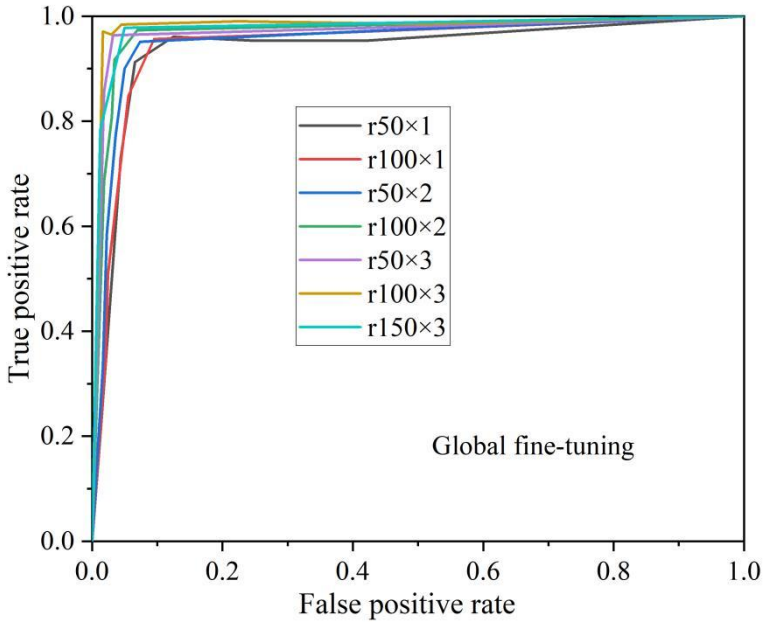

Figure 5: The ROC curves of the global fine-tuning models of 7 network structures

Table 3: The AUC of 7 networks under local fine-tuning and global fine-tuning models

Network structure	Local fine-tuning	Global fine-tuning
r50×1	0.930	0.988
r100×1	0.943	0.983
r50×2	0.958	0.992
r100×2	0.956	0.994
r50×3	0.963	0.996
r100×3	0.979	0.998
r150×3	0.986	0.996

Comparison of Figures 4 and 5 reveals that the global fine-tuning model is superior to the local fine-tuning model. Table 3 further verifies the advantage of global fine-tuning, and the global fine-tuning AUCs are all close to or over 0.99, with the r50×3 and r100×3 networks having AUCs of 0.996 and 0.998 under global fine-tuning, respectively, indicating that the model's ability to discriminate between lesions is close to perfect. The local fine-tuning AUC fluctuates more, for example, the local fine-tuning AUC of the r150×3 network is 0.986, but the global fine-tuning improves it to 0.996, indicating that the global parameter optimization can release the model potential more fully.

3.3 Comparative results of model confusion experiments

Based on the significant advantages of the global fine-tuning strategy, this section introduces a side-by-side comparison between the classical model and the improved SSD model, and analyzes the misclassification patterns of the models in the classification of gastric cancer, gastric ulcer, and chronic gastritis through the confusion matrix.

The models VGG19, ResNet50 and Inception-V3, which were the winners in the ILSVRC tournament in the past few years, are chosen as the comparative training models to be compared with the improved SSD-based gastrointestinal cancer recognition algorithm in this paper. The experiment is based on 1279 lesion samples to recognize 3 types of diseases, gastric cancer (GC), gastric ulcer (GU) and chronic gastritis (CG).The confusion matrices of the 4 models are shown in Table 4.

Table 4: The confusion matrix results of four models for three types of diseases

Model	Predicate/Actually	GC	GU	CG
VGG19	GC	317	14	28
	GU	14	418	16
	CG	17	43	402
ResNet50	GC	322	19	16
	GU	9	428	21
	CG	14	18	432
Inception-V3	GC	338	10	15
	GU	4	423	8
	CG	10	17	454
Improve SSD	GC	350	5	6
	GU	2	430	5
	CG	10	7	464

The confusion matrix results for each model are shown in Figures 6-Figures 9, respectively, after converting the confusion matrix results in Table 4 into percentages.

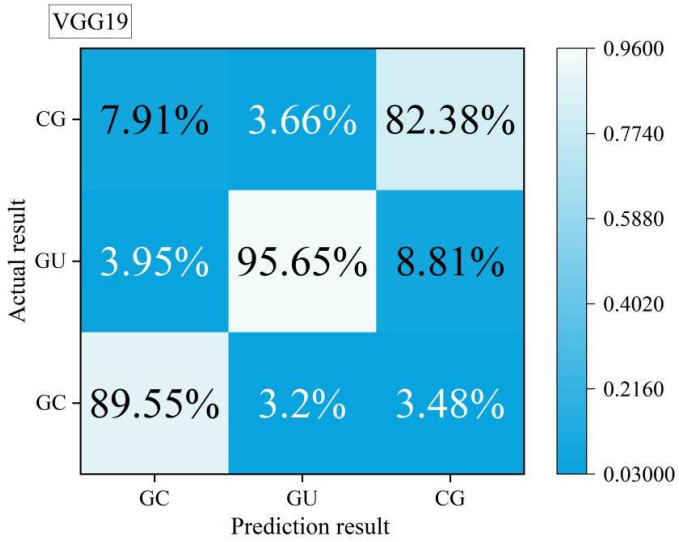

Figure 6: The confusion matrix result of VGG19

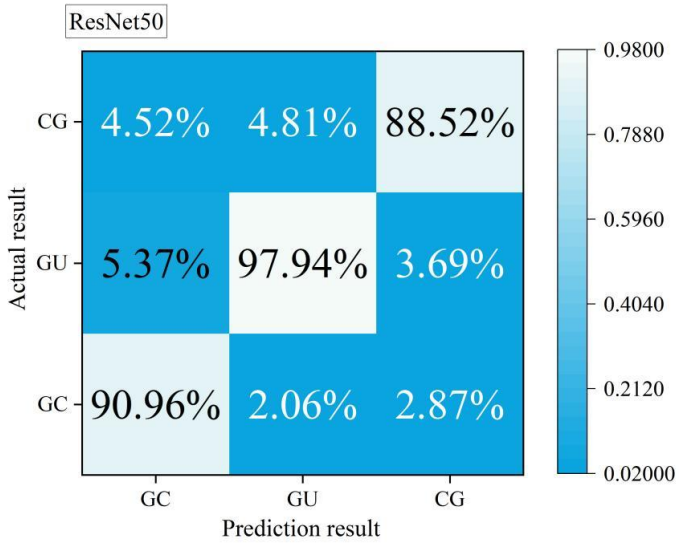

Figure 7: The confusion matrix result of ResNet50

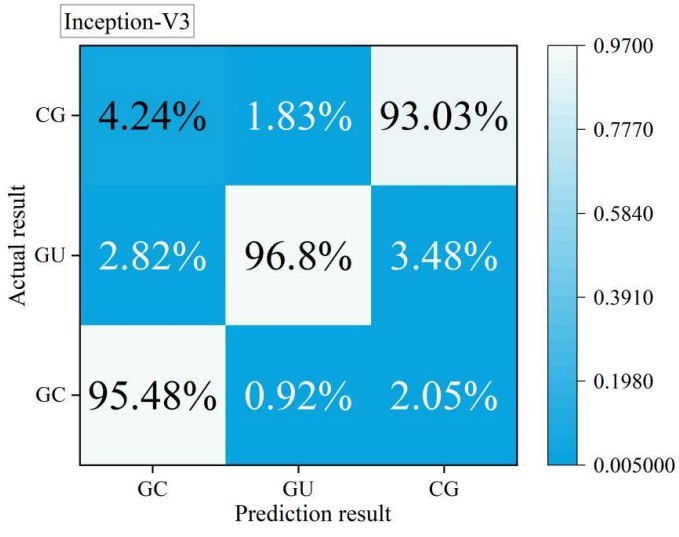

Figure 8: The confusion matrix result of Inception-V3

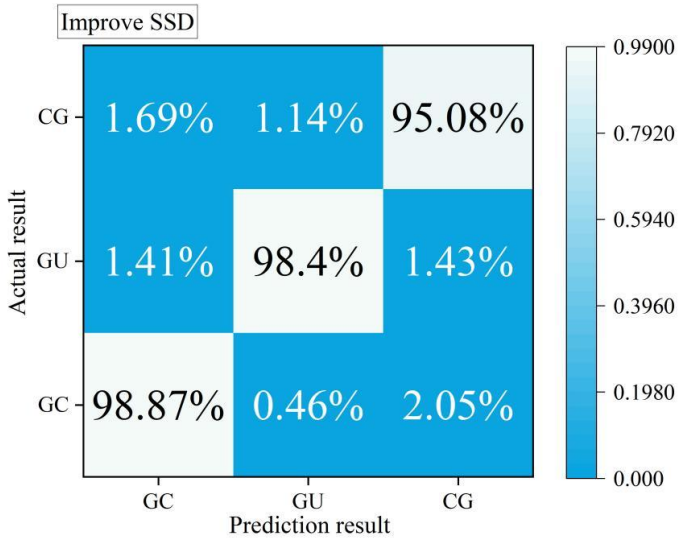

Figure 9: The confusion matrix result of Improve SSD

By comparing the classification performance of VGG19, ResNet50, Inception-V3 and the improved SSD model in three types of diseases, it is found that the improved SSD model is significantly ahead. Taking gastric cancer recognition as an example, the accuracy of the improved SSD reaches 98.87%, and the misclassification rate is only 1.69%, which is much higher than the 89.55% of VGG19 and the 95.48% of Inception-V3. Meanwhile gastric ulcer and chronic gastritis misclassification was reduced. Improved SSD was 98.40% sensitive to gastric ulcer and only 1.14% misclassified as chronic gastritis, while the misclassification rates of VGG19 and ResNet50 were 3.66% and 4.81%, respectively. The accuracy of improved SSD in the identification of chronic gastritis reaches 95.08%, and the misclassification is mainly focused on 1.43% of gastric ulcer, while the misclassification rate of other models is generally higher than 3%.

3.4 Comparative analysis of model and manual classification

To further validate the clinical applicability of the improved SSD model, this section compares the model with the diagnostic results of endoscopists to assess its potential for assisted diagnosis in terms of both accuracy and efficiency.

Continuing with the comparative analysis of human-computer classification, 400 samples of each type in the dataset were selected, including 100 each of gastric cancer, gastric ulcer, chronic gastritis, and normal on average. The results of the identification of the real labels of the dataset confirmed by the endoscopists with the model and the 4 endoscopists (replaced by ABCD, where A and B are the senior physicians and C and D are the 2 junior physicians) are shown in Table 5.

The comparison of recognition results and diagnosis time between the model and the endoscopists is shown in Table 5, and in order to show more intuitively the detection of the recognition model versus the manual, the recognition results are presented in a visualization as shown in Fig. 10, in which the bar indicates the accuracy rate, and the folded line indicates the average recognition time for a single image.

Table 5: Comparison of model and manual identification results and diagnosis time

	GC		GU		CG		Normal	
	Time/s	Accuracy%	Time/s	Accuracy%	Time/s	Accuracy%	Time/s	Accuracy%
VGG19	0.38	89.55	0.17	95.65	0.22	82.38	0.13	95.15
ResNet50	0.23	90.96	0.11	97.94	0.16	88.52	0.09	96.74
Inception-V3	0.16	95.48	0.08	96.80	0.18	93.03	0.07	97.71
Improve SSD	0.05	98.87	0.02	98.40	0.03	95.08	0.01	99.13
Doctor A	5.34	91.08	6.07	90.32	4.35	94.53	5.66	95.77
Doctor B	6.95	95.78	5.61	94.98	5.13	92.03	4.64	96.14
Doctor C	7.45	82.38	8.48	85.17	7.38	89.96	6.47	89.93
Doctor D	10.24	77.29	9.43	82.96	8.43	86.65	7.57	88.46

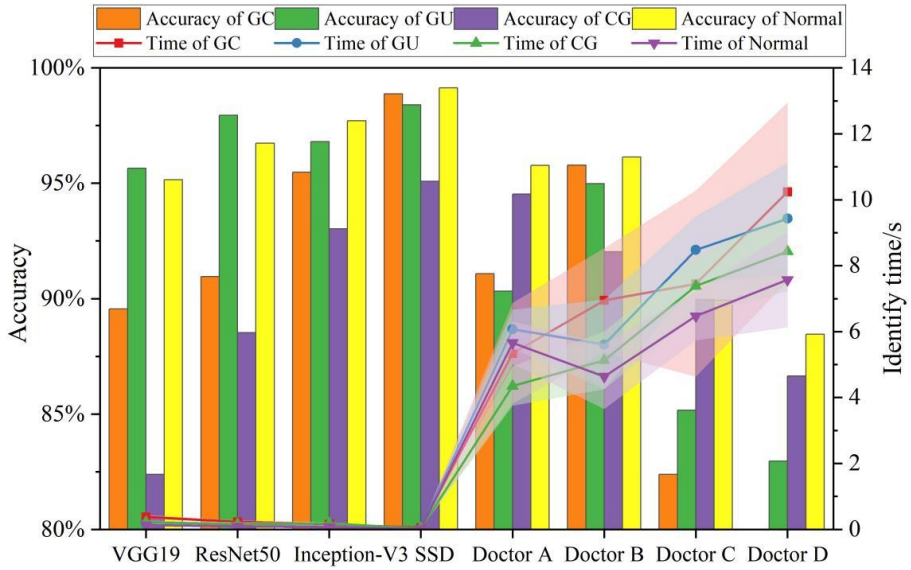

Figure 10: Comparison of model and manual identification results and diagnosis time

The data show that the improved SSD model exhibits significant advantages in all four types of samples. Taking stomach cancer recognition as an example, the accuracy of improved SSD is 98.87%, far exceeding the 89.55% of VGG19 and the 95.48% of Inception-V3, and its average recognition time for a single image is only 0.05s, which is more than three times faster than that of the fastest traditional model, Inception-V3, which is 0.16s.

In human-computer comparisons, the accuracy and efficiency of the improved SSD outperform manual diagnosis. The accuracy of gastric cancer recognition by senior physician B is 95.78% and takes 5.34s, while the accuracy of junior physician D is only 77.29% and takes 10.24s. In comparison, the improved SSD not only has a higher accuracy rate, but also has a speed increase of more than 100 times. In normal sample identification, the accuracy of improved SSD is as high as 99.13%, while the highest accuracy of physicians is 96.14% (Physician B). These results indicate that the improved SSD model has significant potential for application in clinical assisted diagnosis and can assist or partially replace manual diagnosis with higher accuracy and efficiency.

3.5 Model Positioning Accuracy Analysis

Based on the validation of the classification performance, this section focuses on the model's ability to localize the lesion region, and reveals the strengths and limitations of the model in accurate localization through the overlap statistics of lesions with different morphologies.

This dataset contains 350 images identified as early gastric cancer by the model in the classification experiments, and the early gastric cancer lesions are classified into three categories of bulging, flat and depressed according to their morphological features. The number of images with different overlap in the three morphologies and the ratio of overlap in each morphology to the total number of that morphology are counted, and the detailed results of the ratio of different overlap in the three morphologies of early gastric cancer are shown in 6.

Table 6: The proportion of different degrees of overlap in 3 forms of gastric cancer

Form	Dataset	The overlap degree is ≥60%		The overlap degree is ≥70%		The overlap degree is ≥80%		The overlap degree is ≥90%	
		Dataset	Proportion	Dataset	Proportion	Dataset	Proportion	Dataset	Proportion
Uplift	175	158	90.29%	135	77.14%	117	66.86%	56	32.00%
Flat	66	53	80.30%	40	60.61%	29	43.94%	14	21.21%
Depression	109	93	85.32%	81	74.31%	69	63.30%	34	31.19%
Total	350	326	93.14%	245	70.00%	184	52.57%	103	29.43%

The results showed that the model was most capable of localizing augmented lesions, with 158 (90.29%) of 175 augmented samples being accurately localized at ≥60% overlap, and 32% of samples remained compliant even when the overlap threshold was raised to ≥90%. For depressed lesions (109 sheets), the model localized 85.32% accurately at ≥60% overlap and maintained 31.19% at high thresholds (≥90%). Flat lesions (66 sheets) showed relatively weak localization performance, with an accuracy of 80.30% at ≥60% overlap but dropping to 21.21% at high thresholds (≥90%).

Overall, the model's ability to localize elevated (93.14%) and depressed (85.32%) types was superior to flat (80.30%) types. This may be related to the lack of significant morphological features in flat-type lesions. In addition, the percentage of all samples with ≥60% overlap was 93.14%, but only 29.43% reached the threshold of ≥90%, indicating that the model still needs to be optimized for precise localization (e.g., lesion edges). For example, the percentage of high-precision localization (≥90%) for flat lesions was only 21.21%, suggesting that the detection ability of such scenes needs to be improved by enhancing the model's sensitivity to subtle texture changes.

4 Conclusion

In this study, an efficient and high-precision model for recognizing and detecting gastrointestinal precancerous lesions is constructed by integrating multi-task learning with an improved SSD algorithm.

(1) The model performs excellently under the global fine-tuning strategy. Taking the r150×3 network as an example, the global fine-tuning resulted in an accuracy of 98.52%, an F1 score of 94.73%, a specificity of 99.44%, and an AUC value of 0.996, which is a significant enhancement over the local fine-tuning, with the AUC rising from 0.986 to 0.996.

(2) Comparing with the traditional model, the improved SSD leads comprehensively in the recognition of three types of lesions, with 98.87% accuracy in gastric cancer recognition, 1.69% misclassification rate, 98.40% sensitivity in gastric ulcer, 95.08% accuracy in chronic gastritis, and the average recognition time of a single image is only 0.05s, which enhances the efficiency more than a hundred times compared with manual diagnosis.

(3) The localization analysis shows that the model has the strongest localization ability for elevated lesions (90.29% accuracy when the overlap degree is ≥60%), followed by the depressed type with 85.32%.

The study verified the potential of improved SSD in clinical auxiliary diagnosis, and its high accuracy and real-time performance can effectively enhance the efficiency of gastrointestinal precancerous lesion screening.

References

[1] Teh, J. L., Shabbir, A., Yuen, S., & So, J. B. Y. (2020). Recent advances in diagnostic upper endoscopy. World journal of gastroenterology, 26(4), 433.

[2] Kawamura, T., Wada, H., Sakiyama, N., Ueda, Y., Shirakawa, A., Okada, Y., ... & Yasuda, K. (2017). Examination time as a quality indicator of screening upper gastrointestinal endoscopy for asymptomatic examinees. Digestive Endoscopy, 29(5), 569-575.

[3] Faulx, A. L., Kothari, S., Acosta, R. D., Agrawal, D., Bruining, D. H., Chandrasekhara, V., ... & DeWitt, J. M. (2017). The role of endoscopy in subepithelial lesions of the GI tract. Gastrointestinal endoscopy, 85(6), 1117-1132.

[4] Li, H., Hou, X., Lin, R., Fan, M., Pang, S., Jiang, L., ... & Fu, L. (2019). Advanced endoscopic methods in gastrointestinal diseases: a systematic review. Quantitative Imaging in Medicine and Surgery, 9(5), 905.

[5] Yao, K., Uedo, N., Kamada, T., Hirasawa, T., Nagahama, T., Yoshinaga, S., ... & Tajiri, H. (2020). Guidelines for endoscopic diagnosis of early gastric cancer. Digestive Endoscopy, 32(5), 663-698.

[6] Yamamoto, H., Ogata, H., Matsumoto, T., Ohmiya, N., Ohtsuka, K., Watanabe, K., ... & Fujimoto, K. (2017). Clinical practice guideline for enteroscopy. Digestive Endoscopy, 29(5), 519-546.

[7] Terao, S., Suzuki, S., Yaita, H., Kurahara, K., Shunto, J., Furuta, T., ... & Haruma, K. (2020). Multicenter study of autoimmune gastritis in Japan: clinical and endoscopic characteristics. Digestive Endoscopy, 32(3), 364-372.

[8] Mabe, K., Inoue, K., Kamada, T., Kato, K., Kato, M., & Haruma, K. (2022). Endoscopic screening for gastric cancer in Japan: Current status and future perspectives. Digestive Endoscopy, 34(3), 412-419.

[9] Kaji, K., Hashiba, A., Uotani, C., Yamaguchi, Y., Ueno, T., Ohno, K., ... & Yasuda, K. (2019). Grading of atrophic gastritis is useful for risk stratification in endoscopic screening for gastric cancer. Official journal of the American College of Gastroenterology| ACG, 114(1), 71-79.

[10] Jacobs, M. F., Dust, H., Koeppe, E., Wong, S., Mulholland, M., Choi, E. Y., ... & Stoffel, E. M. (2019). Outcomes of endoscopic surveillance in individuals with genetic predisposition to hereditary diffuse gastric cancer. Gastroenterology, 157(1), 87-96.

[11] Ono, H., Yao, K., Fujishiro, M., Oda, I., Uedo, N., Nimura, S., ... & Fujimoto, K. (2021). Guidelines for endoscopic submucosal dissection and endoscopic mucosal resection for early gastric cancer. Digestive Endoscopy, 33(1), 4-20.

[12] Kim, G. H. (2019). Endoscopic Treatment of GI Subepithelial Tumors. Therapeutic Gastrointestinal Endoscopy: A Comprehensive Atlas, 233-254.

[13] Spiceland, C. M., & Lodhia, N. (2018). Endoscopy in inflammatory bowel disease: Role in diagnosis, management, and treatment. World journal of gastroenterology, 24(35), 4014.

[14] Lee, J. H., Kim, Y. J., Kim, Y. W., Park, S., Choi, Y. I., Kim, Y. J., ... & Chung, J. W. (2019). Spotting malignancies from gastric endoscopic images using deep learning. Surgical endoscopy, 33, 3790-3797.

[15] He, Q., Bano, S., Ahmad, O. F., Yang, B., Chen, X., Valdastri, P., ... & Zuo, S. (2020). Deep learning-based anatomical site classification for upper gastrointestinal endoscopy. International journal of computer assisted radiology and surgery, 15, 1085-1094.

[16] Zhu, Y., Wang, Q. C., Xu, M. D., Zhang, Z., Cheng, J., Zhong, Y. S., ... & Li, Q. L. (2019). Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointestinal endoscopy, 89(4), 806-815.

[17] Du, W., Rao, N., Liu, D., Jiang, H., Luo, C., Li, Z., ... & Zeng, B. (2019). Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images. Ieee Access, 7, 142053-142069.

[18] Sharma, A., Kumar, R., & Garg, P. (2023). Deep learning-based prediction model for diagnosing gastrointestinal diseases using endoscopy images. International Journal of Medical Informatics, 177, 105142.