Multimedia Tools and Applications (2023) 82:36577–36598
https://doi.org/10.1007/s11042-023-14982-1

Classification of endoscopic image and video frames using distance metric-based learning with interpolated latent features

Fatemeh Sedighipour Chafjiri[1] · Mohammad Reza Mohebbian[1] · Khan A. Wahid[1] · Paul Babyn[2]

Received: 23 January 2022 / Revised: 16 June 2022 / Accepted: 22 February 2023 /
Published online: 17 March 2023

Abstract

Conventional Endoscopy (CE) and Wireless Capsule Endoscopy (WCE) are well known tools for diagnosing gastrointestinal (GI) tract related disorders. Defining the anatomical location within the GI tract helps clinicians determine appropriate treatment options, which can reduce the need for repetitive endoscopy. Limited research addresses the localization of the anatomical location of WCE and CE images using classification, mainly due to the difficulty in collecting annotated data. In this study, we present a few-shot learning method based on distance metric learning which combines transfer-learning and manifold mixup schemes to localize and classify endoscopic images and video frames. The proposed method allows us to develop a pipeline for endoscopy video sequence localization that can be trained with only a few samples. The use of manifold mixup improves learning by increasing the number of training epochs while reducing overfitting and providing more accurate decision boundaries. A dataset is collected from 10 different anatomical positions of the human GI tract. Two models were trained using only 78 CE and 27 WCE annotated frames to predict the location of 25,700 and 1825 video frames from CE and WCE respectively. We performed subjective evaluation using nine gastroenterologists to validate the need of having such an automated system to localize endoscopic images and video frames. Our method achieved higher accuracy and a higher F1-score when compared with the scores from subjective evaluation. In addition, the results show improved performance with less cross-entropy loss when compared with

Fatemeh Sedighipour Chafjiri and Mohammad Reza Mohebbian contributed equally to this work.

✉ Fatemeh Sedighipour Chafjiri
fas303@usask.ca

[1] Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5A9, Canada

[2] Department of Medical Imaging, University of Saskatchewan and Saskatchewan Health Authority, Saskatoon, SK S7K 0M7, Canada

several existing methods trained on the same datasets. This indicates that the proposed method has the potential to be used in endoscopy image classification.

Keywords Endoscopy · Few shot learning · Manifold mix-up · Siamese neural network · Classification · GI track anatomic locations

1 Introduction

Endoscopy is considered the gold standard for gastrointestinal (GI) examination [33], and is often a key to early mucosal disease identification. All conventional endoscopy (CE) approaches, such as colonoscopy and gastroscopy, are invasive and can cause discomfort and/or harm to the patient [8]; however, they allow real-time examination and visualization of many gastrointestinal abnormalities, including esophagitis, polyposis syndromes, or ulcerative colitis [33]. Wireless Capsule Endoscopy (WCE) offers a non-invasive means for GI inspection and the scanning of areas that are inaccessible, such as the small bowel, using conventional endoscopy techniques. A large number of recorded frames need to be examined by an expert for accurate diagnosis; however, the diagnostic performance achieved from visual inspection is low [50]. For example, the diagnostic accuracy is about 69% for angioectasia, 46% for polyps, and 17% for bleeding lesions [50].

Aside from anomaly detection, localizing the anatomic position of an abnormality accurately within the GI tract is another challenge that remains unsolved [3, 25, 39]. Accurately determining the location of the endoscope's tip in the gastrointestinal tract, and hence the position of an abnormality, is important when further follow-ups, treatment, and/or surgery is needed [37]. Localization also helps to reduce repetitive endoscopy procedures, allowing for targeted drug delivery [30] and automatic endoscopy navigation [41]. Additionally, some diseases characteristically happen at specific locations within the GI tract [14, 18]. For example, dangerous bleeding usually originates from the stomach, small bowel or duodenum [18]. Hence, providing location-based frames for reviewal may reduce examination times and human error in high-risk regions.

Endoscopy frame localization using a single image presents challenges which can be better navigated by using computer-aided intelligent systems. Figure 1 shows an illustration of this challenge. It shows two similar looking frames, one from the proximal part of the stomach (the cardia) and the other from the stomach's distal end (the pylorus). Due to their close resemblance, visual inspection can sometimes be erroneous, leading to an inaccurate diagnosis.

We conducted a survey where nine gastroenterologists were asked to identify the anatomical location of 50 images from the image-based CE dataset, a subset of the dataset used to train the proposed method. Figure 2 shows a screenshot of the questionnaire used in the survey, which is available on the website: https://human-endoscopy-localization.web.app

The CE dataset that we have used (shown in Table 1) contains frames from the esophagus, cardia, stomach's angularis, pylorus, duodenum, ileum, jejunum, colon, rectum, and anus. The gastroenterologists' responses were later analyzed, and performance metrics including F1-score, accuracy, and area under the ROC curve (AUC) were calculated. The results show that performance is poor when human visual inspection is used to identify the GI location from a single image. It will be shown later in the paper that the use of the proposed automated algorithm can improve performance significantly.

There is no publicly available dataset that covers all the significant gastrointestinal anatomical landmarks for WCE and CE. Previous publications have been limited with regard to

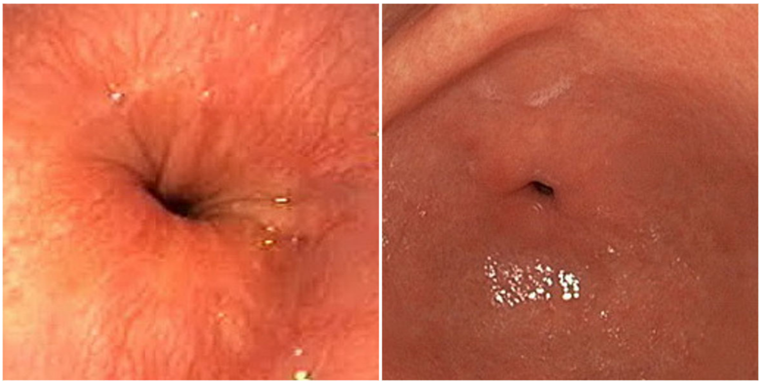

Fig. 1 Difficulty of detecting anatomical location form single image. The left image is of the cardia while the right image is the pylorus

the number of locations used for prediction. Moreover, all previous works are specialized for either WCE or CE. Deep learning methods have yielded great results in image classification [7]. However, the performance of deep learning models is highly dependent on training which typically requires a large number of labeled datasets with a balanced number of samples per class. This is one of the reasons for the limited research available for location classification using deep learning. Moreover, methods that utilized machine learning only focused on the median frame error from a given location or motion estimation due to issues with complexity. Some hardware-based approaches can localize the electronic capsule device but fail to provide information about anatomical locations. Table 1 provides a brief review of existing endoscopic localization techniques.

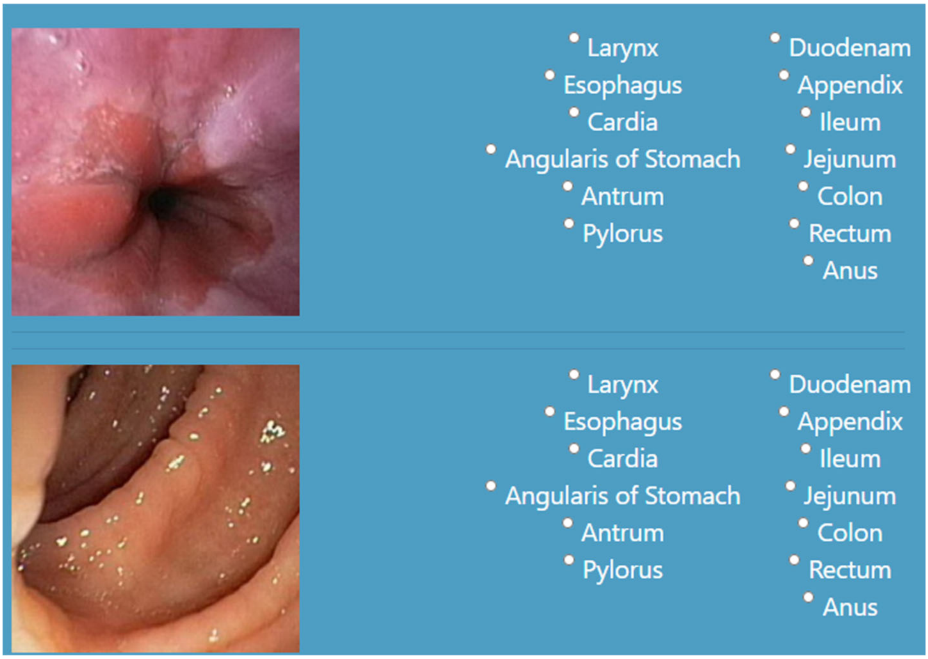

Fig. 2 Two examples from the survey questionnaire used in the subjective evaluation

Table 1 Review of some existing endoscopic localization techniques

Ref	Machine learning (Deep learning)	Method	Number of anatomical locations	Anatomic locations included	Performance metrics and results
[17]	No (No)	Variation in HSV intensity in subsequent frames using event correlation	4	Esophagus, stomach (entering stomach), small intestinal (entering duodenal and ileum), and colon	Recall: 76%; Precision: 51%; F1-score:61%
[36]	Yes (Yes)	CNN	6	Larynx, esophagus, stomach (upper, medium, lower), duodenum	AUC: 100% for larynx and esophagus 99% for stomach and duodenum Accuracy: 97%
[32]	Yes (Yes)	CNN	7	The terminal ileum, the cecum, ascending colon to transverse colon, descending colon to sigmoid colon, the rectum, the anus, and indistinguishable parts	AUC: 97% for the terminal ileum; 94% for the cecum; 87% for ascending colon to transverse colon; 84% for descending colon to sigmoid colon; 83% for the rectum; 99% for the anus. Accuracy: 66%
[20]	Yes (No)	multivariate Gaussian classifiers with color, texture, motion features	3	Median error in frame number prediction for detecting esophagogastric junction; pylorus; ileocecal valve	Esophagogastric junction: 8 pylorus: 91 ileocecal valve:285 (frames)
[9]	Yes (No)	SVM with color and texture features	3	Median error in frame number prediction for detecting esophagogastric junction; pylorus; ileocecal valve	esophagogastric junction: 2 pylorus: 287 ileocecal valve: 1057 (frames)
[43]	No (No)	PCA and customized thresholding approach with color features	2	Median error in frame number prediction for detecting pylorus; ileocecal valve	Pylorus:105 ileocecal valve: 319 (frames)
[23]	Yes (No)	SVM with color features	3	Stomach, small intestine, and large intestine	85.2% (overall accuracy)
[20]	No (No)	The probabilistic latent semantic analysis model for unsupervised data clustering with SIFT features	3	Stomach, small intestine, and large intestine	stomach: 99.9% small intestine: 98.3% large intestine: 94.7% Accuracy

Table 1 (continued)

Ref	Machine learning (Deep learning)	Method	Number of anatomical locations	Anatomic locations included	Performance metrics and results
[9]	Yes (No)	kernel SVM with color intensity, motion, and texture features	2	Motion estimation is evaluated based on median error for detecting pylorus and ileocecal valve	92.7% (average accuracy)
[2]	No (No)	Feature Points Matching for capsule speed estimation	–	Speed estimation accuracy and location error	93% accuracy for speed estimation and 2.49 cm for localization error
[35]	Yes (No)	SIFT features matched using random sample consensus and tracked using Kanade-Lucas-Tomasi tracker	–	Robotic-assisted setup provided for evaluation	2.70±1.62 cm localization error
[46]	No (No)	Using RSS, DoA or ToA	–	average RMSE for predicting capsule location	≈100 mm RMSE with 10 sensors on body surface
[34]	No (No)	Adding small magnet in capsule	–	Capsule inside a volume of 380 mm by 270 mm by 240 mm covered by 16 digital magnetic sensors	10 mm RMSE error

DoA Directional of Arrival, *ToA* Time of Arrival, *RSS* Received Signal Strength, *PCA* Principal Component Analysis, *SIFT* Scale Invariant Feature Transform, *SVM* Support Vector Machine, *MLP* Multi-Layer Perceptron, *CNN* Convolutional Neural Network, *HSV* Hue-Saturation-Value, *RMSE* Root Mean Square Error

The lack of sufficient data is a key reason why anatomic localization has suffered. On the other hand, visual inspection by humans can distinguish between new classes with limited labelled instances [31]. Few-shot learning (FSL) has gained attention in computer vision, especially in the medical field where there are limitations with respect to dataset collection. Models should address this issue of how to train a model with little or no labeled data. This technique attempts to distinguish between new visual categories given few labelled samples [44] in an effort to mimic the way humans learn to predict.

There is a relatively small amount of work on few-shot learning in the medical imaging domain. In [29], the authors utilized a few-shot learning-based method for the diagnosis of diseases and conditions from chest x-rays. Li et al. [19] suggested a unique technique for sub-tomogram classification based on few-shot learning. It allows for the categorization of unseen structures in the training data, given limited labelled samples in the test data, by using instance embedding. Because training data for rare diseases is scarce, [47] has conducted a study on using the concept of few-shot learning for detecting rare diseases. They developed an approach based on FSL using GAN-based data augmentation. These previous studies demonstrate that few-shot learning techniques can achieve reliable performance and outperform classical machine learning models when using small training datasets.

Therefore, this category of learning can be used to improve the performance of anatomical location classification when data samples are limited. FSL algorithms can be categorized into three major categories: initialization based, hallucination-based, and distance metric learning-based approaches. In initialization-based methods, the system focuses on learning to fine-tune or by learning an optimizer. The LSTM-based meta-learner, which can replace the stochastic gradient descent optimizer [31], is an example of this category. The hallucination-based approach tries to train a generator to augment data for a new class, and is usually used in combination with other FSL approaches such as distance-based method [49]. By learning to compare inputs, distance metric learning brings a solution to the FSL approach. The hypothesis is that if a model can assess similarities between two images, it can then identify an unknown input image. A distance-based classification model is simple and yet still achieves competitive results with respect to other complex algorithms [7].

Recently, the introduction of manifold mix-up regularization is thought to help models to have better decision boundaries between classes, while reducing the possibility of overfitting due to increased training epochs [42]. In this paper, we used a distance-based classification technique coupled with manifold mixup to train a deep learning model using fewer images than current models for classifying 10 different anatomical locations of the human GI tract. The combination of a manifold mix-up scheme with a few-shot learning model allowed us to increase the number of training epochs, which in turn decreases the possibility of overfitting. Furthermore, the existing temporal information between individual video frames provides additional information useful for further improving the classification accuracy.

2 Materials and methods

2.1 Dataset

The collected dataset is a set of videos captured by endoscopy cameras from 10 different anatomical locations within the GI region of various patients. A set of images is also provided as the supporting material to train the proposed model. The image dataset consists of both CE and WCE frames extracted from the original video, including 78 CE and 27 WCE images, which was taken from approximately 80–110 different patients. The relevant anatomic locations included are depicted in Fig. 3. Images along with their labels were collected from the Gastrolab gallery [38] and a set of Pillcam images. Furthermore, at least 3 images were collected for each class. Since the collected CE and WCE images were initially 256 × 256 and 512 × 512 pixels respectively, all images were resized to 256 × 256 pixels to have the same size. Regarding the anatomical locations, most of them are available in both the CE and WCE image datasets, including the esophagus, cardia, pylorus, duodenum, ileum, jejunum, and colon (transverse, ascending, descending, and sigmoid). However, the rectum, angularis, and anus regions are only available in the CE dataset.

The WCE video and its labels are taken from the Capsule Endoscopy book by Faigal and Cave [11]. The selected video has 1028 seconds and was recorded using a Pillcam with a frame rate of 5 frames per second, thus resulting in a total of 1825 WCE frames. On the other hand, the CE video was taken from a Gastrolab [38] consisting of 1028 seconds with 25 frames per second recording rate. Hence, the CE video contains 25,700 frames in total. Table 2 lists the details about the mentioned dataset used in this research study.

Diseases and abnormalities may affect the classification result by introducing new structures, which in turn results in different features being extracted. Therefore, the data is supplemented with other data containing numerous diseases. Half of the images in the WCE and CE image-based dataset contain some form of pathology, such as polyps, vascular anomalies, cancer, and inflammation. This is done to determine the efficiency of the proposed method in real world conditions. On the other hand, the video-based dataset has approximately 6500 and 600 frames containing abnormalities for CE and WCE categories respectively. While the evaluation set has a significant number of frames, the image-based dataset for training purposes is quite small, consisting of only 3–10 images per category. The motivation behind having such a size discrepancy between the training and evaluation datasets is to demonstrate

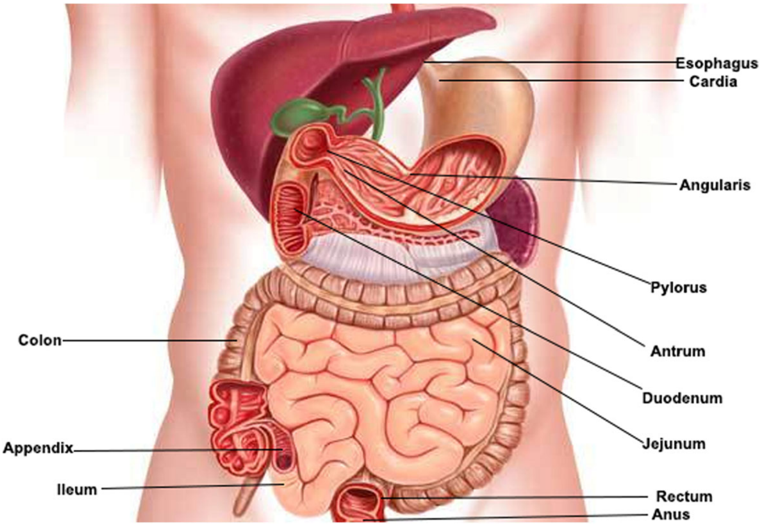

Fig. 3 The anatomic positions of the images included in the dataset for the human GI tract

Table 2 Description of the data set used for training and testing

Position		Images (support set)		Video frames	
Index	Name	CE	WCE	CE	WCE
1	Esophagus	6	3	3075	260
2	Cardia	6	3	2450	20
3	Angularis	8	0	500	0
4	Pylorus	5	3	2500	280
5	Duodenum	16	5	2700	130
6	Jejunum	5	3	1500	380
7	Ileum	11	5	475	280
8	Colon	11	5	5400	475
9	Rectum	5	0	5100	0
10	Anus	5	0	2000	0
Total (Frame)		78	27	25,700	1825
Total (Second)		–	–	1028	365

the efficiency of the proposed few-shot learning model, and to demonstrate that a model trained on only a few images is capable of producing promising results for multiclass disease classification.

2.2 Proposed method

2.2.1 Problem formulation

In this section, we introduce the few-shot classification problem. Considering a dataset, D, with K classes, we have $\sum_{k=1}^{K} N_k$ samples as the supporting set (S), where N_k is the number of images in the kth class. The few-shot image classification problem can be defined as classifying a query set (Q) where the unseen data based on the model is trained on the supporting set samples. The goal is to map the similarity between different images into a metric space, allowing samples from the same category to be as near together as feasible and examples from other categories to be as far apart as possible. This will be done using the Siamese Neural Network (SNN) structure and will be discussed in more details in the following sections.

2.2.2 Siamese network model

SNN is a successful example of distance metric-based methods and was first presented by Bromley et al. [5] in order to detect forged signatures. In that study, by comparing two signatures, the SNN was able to demonstrate whether two signatures were original or whether one was fake. Recently, SNN has been successful with many FSL tasks, including clinical endoscopy image classification [1], COVID-19 diagnosis from x-ray images [15], and medical image classification [6].

The SNN used here employs the DenseNet121 architecture and is combined with a manifold mix-up scheme in order to have more training samples and better decision boundaries. The block diagram of the SNN is shown in Fig. 4. Suppose that two of our supporting set samples are $S_{n_1}\ k_1$ and $S_{n_2}\ k_2$. The result of the network is a feature vector (latent vector) for

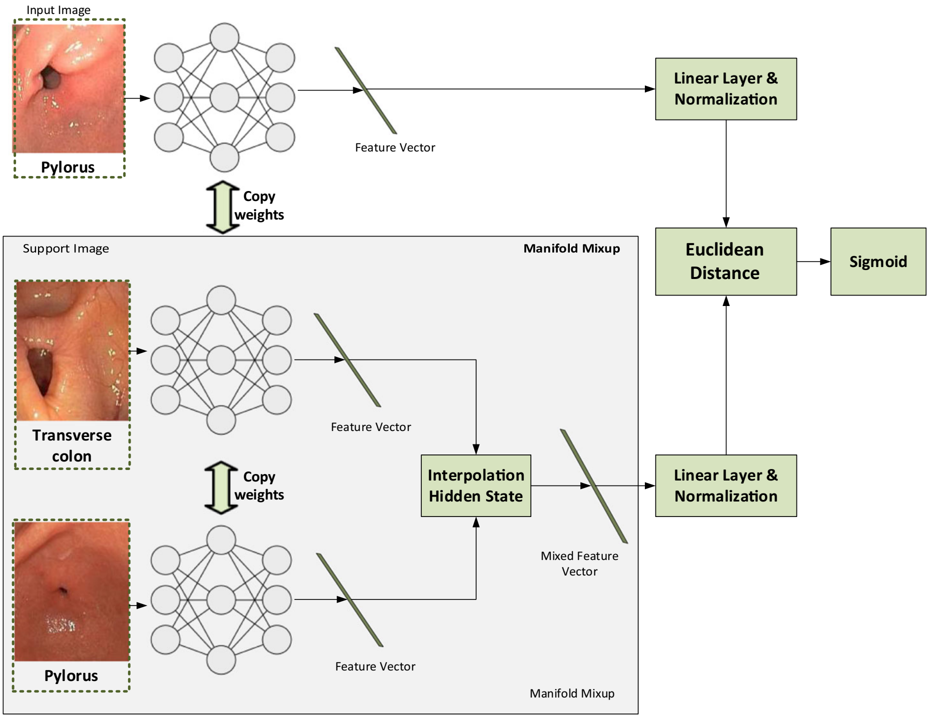

Fig. 4 The block diagram of the training SNN. Instead of using support set image directly, the mixing up of latent features is used for training

each image, which is usually a dense layer, before applying the last activation function, and can be defined as F(x). Various deep learning approaches can be used for feature extraction. We tried different learning architectures that were all pre-trained using ImageNet [28], including DenseNet121, GoogleNet, AlexNet, Resnet50 and VGG16. DenseNet121 was selected for the baseline model since it displayed the highest accuracy. However, the other architectures can replace DenseNet121 without incurring any major performance differences. The next step involves calculating the Euclidean distance between the two extracted feature vectors as shown below:

$$E_w = D_w\left(S_{n_1k_1}, S_{n_2k_2}\right) = \left|F\left(S_{n_1k_1}\right) - F\left(S_{\mathrm{n}_2k_2}\right)\right| \tag{1}$$

If $k_1 = k_2$, which means that the input samples are from the same class, the model learns to extract features that have less distance. On the other hand, if the two images come from separate groups, then the algorithm aims to obtain features such that the distance is greater. The sigmoid function is used to map the distance between 0 to 1. This helps when comparing distances and helps manifold mix-up to have confined values [26].

Contrastive loss is used to train the network [13]. The map for converting an image to its latent vector should preserve neighboring relationships and should be generalizable on unseen data. The loss is defined by Eq. 2 below as:

$$L(D_w, Y) = (1-Y)D_w^2 + Y\{max(0, 1-D_w)\}^2 \tag{2}$$

Where, Y is 0 when $k_1 = k_2$ are similar and is 1 when they are different; D_w is the Euclidean distance. The loss function is optimized using an RMSprop optimizer [51].

2.2.3 Manifold mixup

Deep learning networks usually perform appropriately on the data distribution they were trained on; however, they may provide incorrect (and sometimes very confident) answers when evaluated on points from outside the training distribution. The adversarial examples presented in [12] are an example of this issue. Manifold mix-up, introduced by Verma et al. [42], uses a regularization that solves this problem by training the classifier with interpolated latent features, which allows it to be less confident with points outside of the distribution. It also enhances the latent representations and decision boundaries of neural networks. We suppose that the extracted features from one location are unique to that location. As a result, combining latent features from two locations generates a new feature that is close to both locations. The degree of resemblance is then determined by mixing the weights.

Suppose $\check{x} = g(x)$ is the neural network function that maps one support image x to its latent feature $\check{x}$. We assume two support images x_1 and x_2, and proceed to mix two latent features $\check{x}_1$ and $\check{x}_2$. The mixing function is defined by the following equation:

$$Mix_\lambda(\check{x}_1, \check{x}_2) = \lambda\check{x}_1 + (1-\lambda)\check{x}_2 \tag{3}$$

Where, λ is defined based on the $Betta(\alpha, \ \alpha)$ distribution [48] . The value of α is set to 2 because the original paper achieved the best results with this value. A bigger λ means that the latent feature is more like x_1. Similarly, the labels of two support images x_1 and x_2, which are defined as y_1 and y_2, are mixed:

$$Mix_\lambda(\check{y}_1, \check{y}_2) = \lambda\check{y}_1 + (1-\lambda)\check{y}_2 \tag{4}$$

If two support images are from different locations than the SNN network's input image, the output does not change. Therefore, one of the images should be from same location as the input SNN image. For each pair, 50 different mixed latent features and labels are generated.

2.2.4 Applying the model to a single frame and a sequence of frames

Figure 5 shows the method used to apply a single image to the trained model. When a new image is fed to the trained model, a feature vector is calculated. The Euclidean distance between the obtained feature vectors and other classes are calculated. It is observed that the minimum median distance between each group indicates the group that the new image is classified as. If the median distance from all group members is above the threshold of 0.5, a

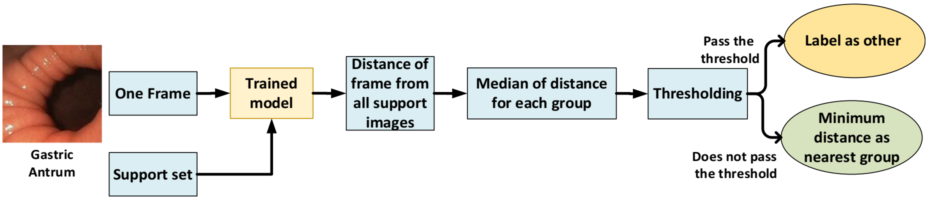
Gastric Antrum
One Frame
Support set
Trained model
Distance of frame from all support images
Median of distance for each group
Thresholding
Pass the threshold
Label as other
Does not pass the threshold
Minimum distance as nearest group

Fig. 5 The pipeline of applying a single frame to SNN for detecting location

new category is generated for the image, and subsequently labeled as "Other". According to the block diagram of the proposed method (Fig. 4), the sigmoid function is used as the last layer to convert the distance of two classes to a value between 0 and 1. This approach is like a calibration which is famous as the sigmoid method. The sigmoid method assumes the calibration curve can be corrected by applying a sigmoid function to the raw predictions. This assumption has been empirically justified on various benchmark datasets [4]. As a result, the calibration curve also referred to as the reliability diagram [45] shows a characteristic sigmoid shape, indicating that the classifier returns probabilities closer to 0 or 1 typically. Therefore, 0.5 is used as a discriminator for whether an image belongs to a specific class or not. We used the median, instead of the average, as it makes the algorithm more robust against noise [10].

In the next stage, we develop our model using video sequences. Video is understood to be a series of individual images (or frames) in sequence that can help further improve the predictions using the temporal information.

In order to take advantage of the correlation between successive frames, extra steps were implemented. Figure 6 represents a block diagram that shows the application of the model to a video sequence. Each video is segmented into 1 sec windows with 0.5 sec overlaps. Since the anatomic changes in adjacent video frames are not usually high, frames inside a window can be assigned to a single location instead of assigning a location to each frame. Therefore, the error incurred from applying the model to a single frame can be reduced by taking advantage of the temporal information. In this regard, each frame is applied to the single frame model. Then, the statistical mode from 1 second worth of frame locations is used as the predicted label for that second. It is worth noting that WCE and CE videos contain 5 and 25 frames per second respectively.

It is assumed that the frame positions are in anatomical order, and that the anatomic order is preserved throughout the processing of a video sequence. For example, it is not possible for "colon" to precede "cardia". Hence, if the predicted label for a sliding window was not ordered according to their anatomical positions, the label with a higher average distance from its group is set to "Other". The algorithm is presented below.

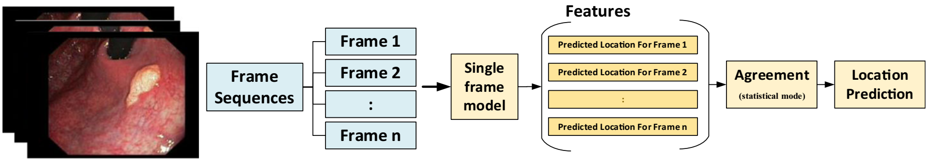

Fig. 6 The pipeline of applying a video sequence to SNN for detecting location

Algorithm 1 Applying a video sequence to the model.

The pseudo code for applying a video sequence to the model.

```
Input: Batch size N of video sequence frames, support set S, trained SNN, and the threshold $\Theta$ for assigning samples to the "Other" class
Output: batchLabel
$\Theta = 0.5$
prediction_list = []
for each frame in batch:
    features = extractFeatures (SNN, frame)
    distance_list = euclideanDistance (S, features)
    prediction = median (distance_list)
    if prediction >= $\Theta$:
        predicted_label = "Others"
    else:
        predicted_label = getClosestLabel (prediction)
    end if
    update prediction_list by adding predicted_label to it.
batchLabel = statisticalMode (prediction_list)
```

2.3 Performance evaluation

Two stages of validation were applied to the proposed method. Firstly, we applied SNN for evaluating single frame model performance, which was tested on all frames from the video-based dataset. For validating the entirety of the proposed system, including SNN and postprocessing, the test dataset was 50% of the video-based dataset, which included 13,762 endoscopy video frames.

For the multiclass problem, the macro-average of Accuracy, AUC and F1-score are reported. AUC is a performance measurement to check and visualize the multi-class classification, and additionally specifies the degree of separability. The higher the AUC, the better the model is at distinguishing between classes. It is worth mentioning that the micro-average is not sensitive to individual group predictive results and can be misleading when data is imbalanced [22]. For the multiclass problem, the overall accuracy is reported, which is the average accuracy for all of classes.

All algorithms ran in Python 3.6 on a system with a Core-i9 CPU, 16 GB of RAM, and 6 GB NVIDIA GeForce GTX 1060 Graphic Cards. The training and inference time of classifying the locations are about 8.07 s and 0.06 s, respectively.

3 Results

3.1 Subjective evaluation using a survey questionnaire

The results of the subjective evaluation by the gastroenterologists are shown in Fig. 7. It shows the macro-average F1-score, AUC and overall accuracy to be 55%, 78% and 60% respectively. The numbers are lower which indicates that GI tract localization through visual inspection is difficult. This is partly to do with the fact that there are many similarities between different locations which can lead to erroneous classification by humans.

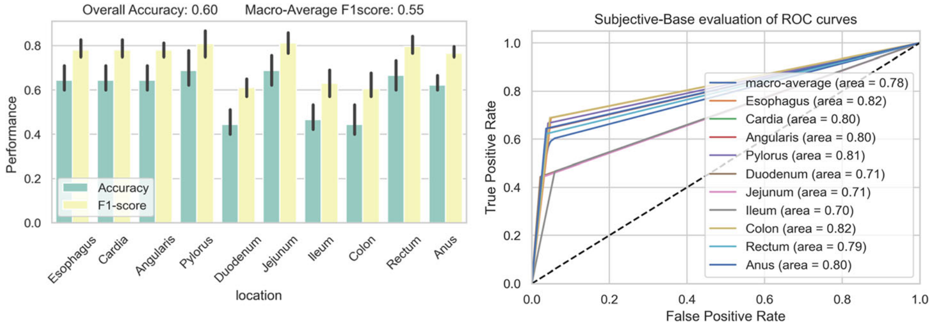

Fig. 7 Performance of predicted location by nine gastroenterologists using the CE dataset. The ROC curve for each location along with the macro-average (right) and the F1-score and accuracy (left) are provided

3.2 Proposed method performance

The proposed SNN, which is trained on an image-based dataset, was applied to the video-based dataset for CE and WCE images without considering frame sequence. The results are provided in Fig. 8. Put concisely, the proposed SNN method, using DenseNet121 while being trained on 78 CE images, achieved a macro-average F1-score, AUC, and overall accuracy of 78%, 90%, and 83% respectively for CE. Similarly, the model trained on 27 WCE images achieved 78%, 90%, and 84% for the F1-score, AUC and overall accuracy, respectively.

Figure 9 shows the effect of applying agreement (statistical mode) to a sequence of frames. In order to use the information from neighboring frames, an agreement of 25 and 5 frames

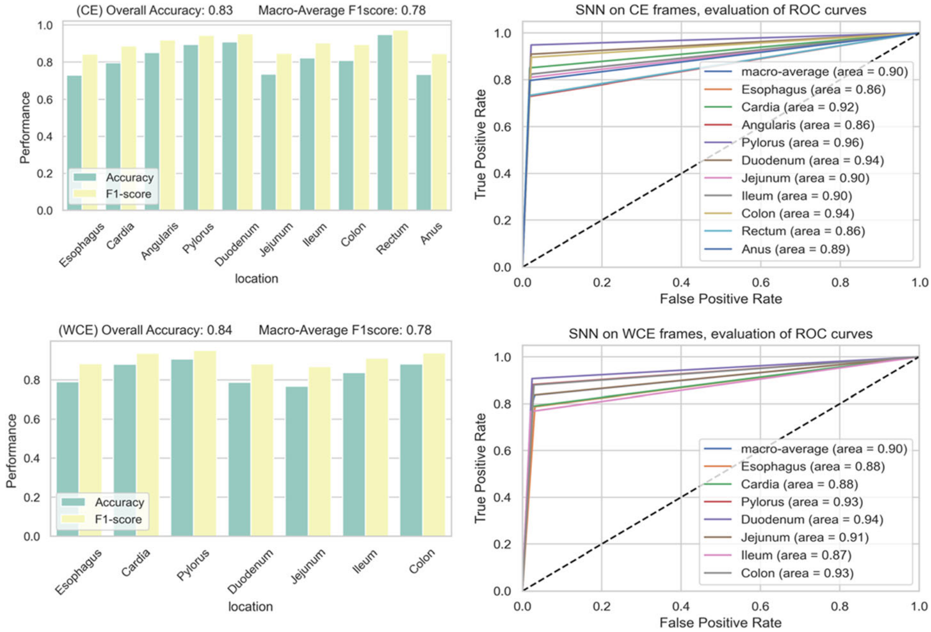

Fig. 8 Results of the proposed method on single frames from the CE (top) and WCE (bottom) video-based datasets

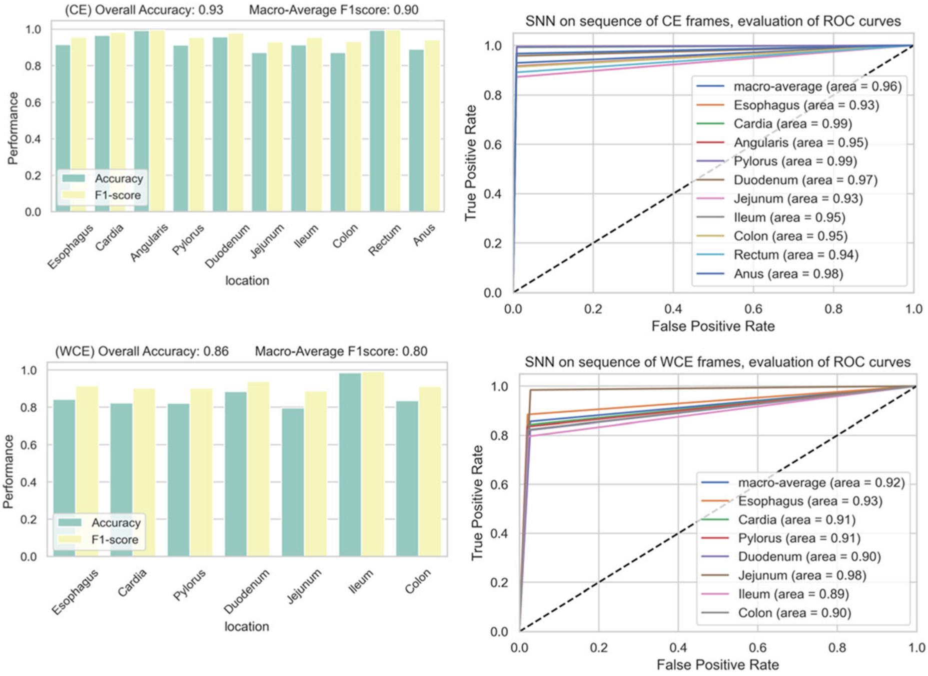

Fig. 9 Results of the proposed method on a sequence of frames from the CE (top) and WCE (bottom) video-based datasets

respectively was selected for CE and WCE location labels. The proposed method, based on the agreement of frame sequence predictions, achieved a macro-average F1-score, AUC, and overall accuracy of 90%, 96%, and 93% for CE and 80%, 92%, and 86% for WCE respectively.

To illustrate the performance of the proposed method better, we have provided an example in Fig. 10 for processing a 34-second conventional endoscopy video. While the endoscope is in the esophagus, there are times at which the proposed method (without agreement) cannot detect the correct location. The presence of different artifacts such as bubbles, instrument noise, blurring, contrast issues, color saturation, or simply the frame belonging to a location that was not in the training set (such as the antrum) are examples of false predictions. The agreement of locations in a time frame can reduce error. For instance, after detecting the

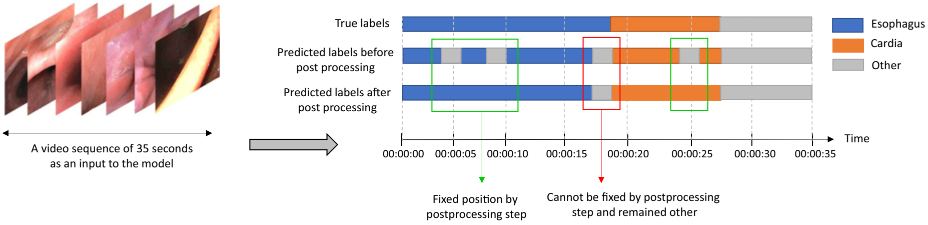
A video sequence of 35 seconds as an input to the model
True labels
Predicted labels before post processing
Predicted labels after post processing
Esophagus
Cardia
Other
Time
00:00:00
00:00:05
00:00:10
00:00:15
00:00:20
00:00:25
00:00:30
00:00:35
Fixed position by postprocessing step
Cannot be fixed by postprocessing step and remained other

Fig. 10 An overview of the outputs of the system and the error correction mechanism applied by our postprocessing step. The "Other" label is mainly because of the inability of SNN to detect correct location because of artifacts and noise, or it is a location that was not in the training set such as the Antrum. Blue boxes show erroneous predictions corrected using the proposed postprocessing step

esophagus' position, the next position, in this case the cardia, is expected to be predicted. Therefore, if irrelevant positions are detected, the agreement process fixes the incorrectly predicted frames, and thereby improves the performance.

3.3 Effect of distance metric-based and manifold mix-up

Figure 11 shows the comparison results among SNN with manifold mix-up (proposed method), SNN without manifold mix-up, simple CNN, SVM with Scale Invariant Feature Transform (SIFT) features, SVM with color and texture features, GoogleNet, AlexNet, Resnet50 and VGG16. The advantage of this experiment is to apply the same dataset to the base methods of the mentioned state-of-the-arts architectures in Table 3 and compare the results with the proposed method. According to this table [32, 36] have employed CNN, [23] classified the locations using SVM, and [35] used unsupervised data clustering with SIFT. It is worth mentioning that the proposed method with manifold mix-up is trained on limited data, while the others (even SNN without manifold mix-up) are trained on 50% of the frames from the video-based dataset. The proposed method outperforms other models, although it is trained on only 78 CE and 27 WCE images whereas the other models are trained on 12,850 CE and 912 WCE images. For CE, the VGG16 achieved the best score after the proposed method with a macro-average F1-score of 77.1%. On the other hand, Resnet50 achieved the best score for WCE after the proposed method with a macro average F1-score of 73.7%. Additional information about VGG16 and Resnet50 is provided in the supplementary material.

Figure 12 shows the latent vector visualization for CE and WCE images based on DenseNet121 on two dimensions using t-SNE. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique that is ideally suited for the visualization of high-dimensional data [40]. The extracted latent feature from the model is visualized here using t-SNE for better interpretation of the trained model. All test samples are fed into the base model and the t-SNE of the latent features are calculated and depicted with and without manifold mix-up. It is worth noting that since t-SNE holds probabilities rather than distances, calculating any error between Euclidean distances in high-D and low-D is pointless. Continuous lines in the 2D plot also shows that there is a time series behavior in features, which is

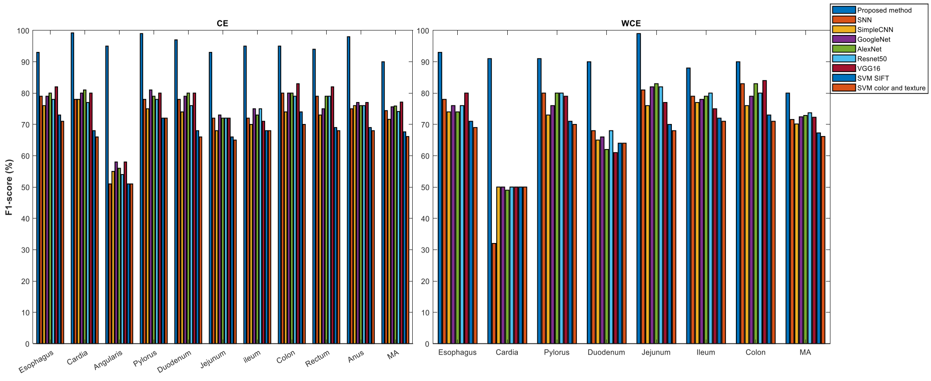

Fig. 11 Comparing the F1-score of the proposed method using manifold mix-up with other methods including SNN without manifold mix-up, other architectures, and hand-crafted features with machine learning

Table 3 Comparing the performance and properties of the proposed method with other methods that used image classification as the localization method

Study	Method	Endoscopy Images	Number of locations	Locations	Training sample Size	Validation strategy (Test size)	Best result
[17]	Variation in HSV intensity in subsequent frames using event correlation	WCE	5	esophagus, stomach (entering stomach), small intestinal (entering duodenal and ileum), and colon	NA	External Dataset (10 videos, number of frames is NA)	Recall: 76%; Precision: 51%; 61% F1-score
[36]	Convolutional Neural Network	CE	6	Larynx, Esophagus, Stomach (Upper, Medium, Lower), Duodenum	27,335	13,048	97% Accuracy
[32]	Convolutional Neural Network	CE	6	the terminal ileum, the cecum, ascending colon to transverse colon, descending colon to sigmoid colon, the rectum, the anus	4100	1025	66% Accuracy
[23]	SVM with color features	WCE	3	stomach, small intestine, and large intestine	26,469	10-fold cross validation	85% Accuracy
[35]	The probabilistic latent semantic analysis model for unsupervised data clustering with Scale Invariant Feature Transform (SIFT) features	WCE	3	stomach, small intestine, and large intestine	50,000	10-fold cross validation	97.6% Accuracy
Proposed method	Attention-based SNN with Manifold mixup	WCE and CE	10 for CE 7 for WCE	Esophagus, Cardia, Angularis, Pylorus, Duodenum, Jejunum, Ileum, Colon, Rectum, Anus	78 CE 27 WCE	External Dataset (2570 CE, 1825 WCE)	CE: 93% Accuracy WCE: 86% F1-score

NA Not Available.

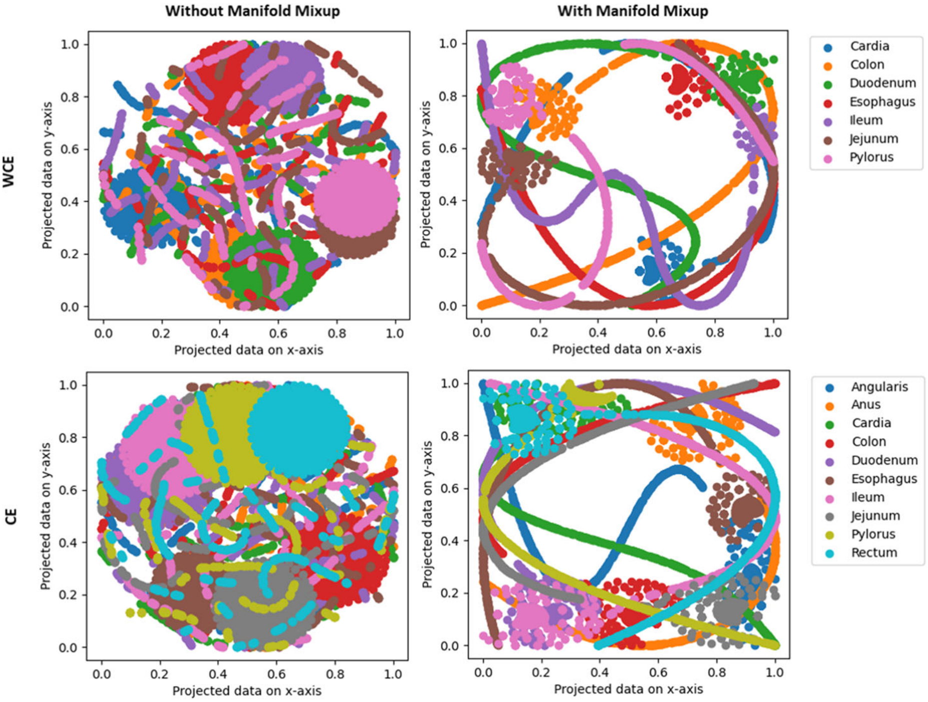

Fig. 12 The visualization of latent features extracted from CE and WCE video-based datasets using t-SNE with a perplexity of 50 based on the proposed method with and without manifold mix-up. The latent features with manifold mix-up have better discrimination and decision boundaries, whereas the latent features extracted without manifold mix-up have more overlaps

because of video frames. Moreover, the 2D plot shows that the complexity of manifold without the manifold mix-up scheme is higher (clusters are more correlated), and that the manifold mix-up can help to better find similarities between frame sequences (more rigid lines).

4 Discussion

In this paper the problem of classifying the anatomic origin of endoscopic images from the GI track has been investigated. Our subjective evaluation shows that, due to the resemblance in locations, achieving a high accuracy is not possible by gastroenterologists. Therefore, this paper aimed to investigate the use of deep learning-based technique to classify the GI tract locations.

Two approaches to overcome this challenge are to either collect more samples to train the model or to take advantage of neural network architectures that are able to learn from few samples and predict based on them. Since we cannot always rely on getting more data, the proposed method is based on two SNN models that are trained using manifold mix-up for classifying anatomical locations of the GI track given 78 CE and 27 WCE images, separately.

In Table 1, different endoscopic localization techniques are presented that vary based on various features including their type of output. For example, the output of [17, 32, 36] are predicted locations while [9, 20, 43] have estimated the median error. Table 3 compares our results with the existing studies that perform the same task as the proposed method, namely the classification of the anatomic locations of the GI track and label prediction.

To summarize, Lee et al. [17] designed a system to detect the esophagus, stomach, duodenal, ileum and colon (5 locations), based on color changes observed in consecutive video frames, and achieved a 61% F1-score; however, they did not utilize any machine learning or deep learning approaches. The proposed approach outperforms their result by 25% with respect to F1-score. Marques et al. [23] used color features and SVM for the stomach, small intestine, and large intestine (3 locations) for classification on WCE frames and achieved an overall accuracy of 85.2%. Shen et al. [35] used SIFT local feature extraction on WCE images and unsupervised learning-based on clustering for localization of the stomach, small intestine, and large intestine (3 locations), and achieved an overall accuracy of 97.6%. For the first time, Takiyama et al. [36] used standard endoscopy images for training a CNN to classify input images as either the larynx, esophagus, stomach (upper, medium, and lower part) or duodenum (6 locations). They achieved 97% accuracy with an AUC > 99%.

While the last two methods have higher accuracy, their classification task is less complex than what we have presented. Having more classes makes the problem more complicated. Increasing the number of classes is also investigated in other fields such as anomaly detection. For instance, Mohammed et al. [24] showed that increasing the number of classes makes the problem more complicated and causes a drop in performance. On the other hand, having more locations for prediction makes the classification more precise. Saito et al. [32]. Applied a CNN to standard colonoscopy images, such as the terminal ileum, cecum, ascending colon, transverse colon, descending colon, sigmoid colon, rectum, and the anus (8 locations), and achieved an overall accuracy of 66%.

All methods are applied to a limited number of locations, on either WCE or CE datasets. However, in this research both WCE and CE localization are investigated with a wide range of locations from the esophagus to the anus. Furthermore, the number of images that we used for training is significantly lower than that used in the other methods. In this study, the number of samples have been reduced by over 99% and 98% for WCE and CE datasets respectively.

Trained models could be used for the location prediction of a single frame or a sequence of frames. Using the proposed postprocessing step, which is an agreement of predicted neighbor labels, the method took advantage of the temporal dependencies that exist in the frame sequences. Consequently, the error rate decreased.

This study has suggested that few-shot learning methods have great potential in medicine. To this end, a SNN with manifold mix-up has been presented. In order to generalize this notion, two other well-known approaches for few-shot learning, namely matching [27] and prototypical networks [16], have been chosen. In addition, the same dataset is used to train them with and without manifold mix-up. The results are presented in Fig. 13. Comparing the methods without manifold mix-up with those that utilize it shows that this technique improves few-shot learning. [21] has also investigated the role of Manifold mix-up on few-shot tasks and achieved increased performance; however, they used self-supervised learning and the efficiency was not shown for supervised approaches.

Matching and prototypical networks have high performance, close to that achieved by the SNN. Although, there are a few studies [35, 36] in Table 3 that have achieved slightly higher accuracy. Few-shot learning-based methods have acceptable results when trained on much fewer samples while covering more classes. This proves the effectiveness of few-shot learning-based methods in CE and WCE areas and motivates further exploration of these techniques and their application in the field.

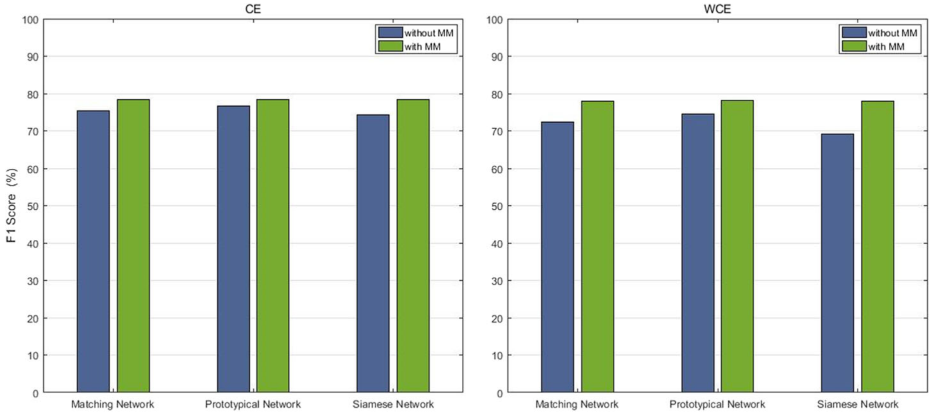

Fig. 13 Comparing three few-shot learning-based methods with and without manifold mix-up. Siamese network with manifold mix-up is the proposed method in this study. The implemented backbone network is Denenet121

5 Conclusion

To classify WCE and CE images based on their anatomical locations, a few-shot learning strategy based on the Siamese Neural Network and Manifold Mix-up is used in this research. Only 78 CE and 27 WCE images were used to train the proposed method. The number of training pairings is significantly enhanced when employing the distance metric-based technique and manifold mix-up (regularization and augmentation), which reduces the potential of overfitting. On an external dataset of 25,700 CE and 1825 WCE video frames, the suggested technique obtains a macro-average F1-score, AUC, and overall accuracy of 90%, 96%, and 93% for CE and 80%, 92%, and 86% for WCE respectively. The need for this research and investigating methods for automatic classification of gastrointestinal anatomical locations has also been validated using a subjective evaluation using nine gastroenterologists. According to this evaluation, the proposed method had higher performance.

Various studies have been carried out to demonstrate the significance of each part of the proposed method. The results showed that in the proposed method, other architectures can also be used instead of DenseNet121 without major changes in performance. Furthermore, it was demonstrated that distance metric-based approaches with manifold mix-up, which are trained on poorly sampled data, have the ability to outperform models trained using categorical cross-entropy loss. For instance, the proposed method outperformed other techniques, including a support vector machine with hand-crafted features, a convolutional neural network, and transfer learning-based methods, which are trained on categorical cross-entropy loss. The proposed classification technique has been used for classifying the GI track location in video frame sequences. The defined agreement stage improved the prediction result compared to not using this postprocessing step by taking advantage of the temporal information. The visual inspection performed by nine experts on images also showed that an AI system can outperform visual inspections and it can help to improve diagnosis performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11042-023-14982-1.

Acknowledgments We thank all physicians who completed the survey. Also, we would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting this work.

Data availability The data supporting the findings of this study are openly available at https://bit.ly/GITrackImageClassification.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ali S, Bhattarai B, Kim T-K, Rittscher J (2020) Additive angular margin for few shot learning to classify clinical endoscopy images. In: International Workshop on Machine Learning in Medical Imaging, Springer, pp 494–503
2. Bao G, Pahlavai K (2013) Motion estimation of the endoscopy capsule using region-based kernel SVM classifier. In: IEEE international conference on electro-information technology, EIT 2013, pp 1–5. https://doi.org/10.1109/EIT.2013.6632652
3. Bernal J, Sánchez J, Vilariño F (2012) Towards automatic polyp detection with a polyp appearance model. Pattern Recogn 45:3166–3182. https://doi.org/10.1016/j.patcog.2012.03.002
4. Böken B (2021) On the appropriateness of Platt scaling in classifier calibration. Inf Syst 95:101641
5. Bromley J, Bentz JW, Bottou L, Guyon I, LeCun Y, Moore C, Säckinger E, Shah R (1993) Signature verification using a "siamese" time delay neural network. Int J Pattern Recognit Artif Intell 7:669–688
6. Cai A, Hu W, Zheng J (2020) Few-shot learning for medical image classification. In: International Conference on Artificial Neural Networks, Springer, pp 441–452
7. Chen W-Y, Liu Y-C, Kira Z, Wang Y-CF, Huang J-B (2019) A closer look at few-shot classification, ArXiv Preprint ArXiv:1904.04232
8. Ciuti G, Menciassi A, Dario P (2011) Capsule endoscopy: from current achievements to open challenges. IEEE Rev Biomed Eng 4:59–72
9. Cunha JPS, Coimbra M, Campos P, Soares JM (2008) Automated topographic segmentation and transit time estimation in endoscopic capsule exams. IEEE Trans Med Imaging 27:19–27. https://doi.org/10.1109/TMI.2007.901430
10. Doerr B, Sutton AM (2019) When resampling to cope with noise, use median, not mean. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp 242–248
11. Douglas DRC, Faigel O (2007) Capsule Endoscopy, SAUNDERS ELSEVIER
12. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples, ArXiv Preprint ArXiv:1412.6572
13. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), IEEE, pp 1735–1742
14. Hosoe N, Hayashi Y, Ogata H (2020) Colon capsule endoscopy for inflammatory bowel disease, clinical endoscopy
15. Jadon S (2021) COVID-19 detection from scarce chest x-ray image data using few-shot deep learning approach. In: Medical imaging 2021: Imaging informatics for healthcare, research, and applications, international society for optics and photonics, p 116010X
16. Jake S, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning, ArXiv preprint ArXiv:1703.05175
17. Lee J, Oh J, Shah SK, Yuan X, Tang SJ (2007) Automatic classification of digestive organs in wireless capsule endoscopy videos. In: Proceedings of the 2007 ACM Symposium on Applied Computing, pp 1041–1045
18. Lewis BS (2000) Small intestinal bleeding. Gastroenterol Clin N Am 29:67–95
19. Li R, Yu L, Zhou B, Zeng X, Wang Z, Yang X, Zhang J, Gao X, Jiang R, Xu M (2020) Few-shot learning for classification of novel macromolecular structures in cryo-electron tomograms. PLOS Comput Biol 16(2):e1008227
20. Mackiewicz M, Berens J, Fisher M (2008) Wireless capsule endoscopy color video segmentation. IEEE Trans Med Imaging 27:1769–1781

21. Mangla P, Kumar Singh M, Sinha A, Kumari N, Balasubramanian V, Krishnamurthy B (2020) Charting the right manifold: manifold mixup for few-shot learning. IEEE Winter Conference on Applications of Computer Vision (WACV), pp 2207–2216
22. Mansourian M, Marateb HR, Mansourian M, Mohebbian MR, Binder H, Mañanas MÁ (2020) Rigorous performance assessment of computer-aided medical diagnosis and prognosis systems: a biostatistical perspective on data mining. Model Anal Active Biopotential Signals Healthcare, 2 (2020) 17–1 to 17–24. https://doi.org/10.1088/978-0-7503-3411-2ch17
23. Marques N, Dias E, Cunha JPS, Coimbra M (2011) Compressed domain topographic classification for capsule endoscopy, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 6631–6634. https://doi.org/10.1109/IEMBS.2011.6091635
24. Mohammed A, Farup I, Pedersen M, Yildirim S, Hovde Ø (2020) PS-DeVCEM: pathology-sensitive deep learning model for video capsule endoscopy based on weakly labeled data. Comput Vis Image Underst 201: 103062
25. Mohebbian MR, Sohag MHA, Vedaei SS, Wahid KA (2020) Automated detection of bleeding in capsule endoscopy using on-chip multispectral imaging sensors. IEEE Sensors J 1–1. https://doi.org/10.1109/JSEN.2020.3034831
26. Mourgias-Alexandris G, Tsakyridis A, Passalis N, Tefas A, Vyrsokinos K, Pleros N (2019) An all-optical neuron with sigmoid activation function. Opt Express 27:9620–9630
27. Oriol V, Blundell C, Lillicrap T, Wierstra D (2016) Matching networks for one shot learning. Adv Neural Inf Process Syst 29:3630–3638
28. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
29. Paul A, Tang Y, Shen T, Summers R (2021) Discriminative ensemble learning for few-shot chest x-ray diagnosis. Med Image Anal 68:101911
30. Pedersen PB, Bar-Shalom D, Baldursdottir S, Vilmann P, Müllertz A (2014) Feasibility of capsule endoscopy for direct imaging of drug delivery systems in the fasted upper-gastrointestinal tract. Pharm Res 31:2044–2053
31. Ravi S, Larochelle H (2016) Optimization as a model for few-shot learning
32. Saito H, Tanimoto T, Ozawa T, Ishihara S, Fujishiro M, Shichijo S, Hirasawa D, Matsuda T, Endo Y, Tada T (2020) Automatic anatomical classification of colonoscopic images using deep convolutional neural networks. Gastroenterol Rep
33. Schwartz DA, Wiersema MJ, Dudiak KM, Fletcher JG, Clain JE, Tremaine WJ, Zinsmeister AR, Norton ID, Boardman LA, Devine RM (2001) A comparison of endoscopic ultrasound, magnetic resonance imaging, and exam under anesthesia for evaluation of Crohn's perianal fistulas. Gastroenterology 121: 1064–1072
34. Shao G, Tang Y, Tang L, Dai Q, Guo Y-X (2019) A novel passive magnetic localization wearable system for wireless capsule endoscopy. IEEE Sensors J 19:3462–3472
35. Shen Y, Guturu P, Buckles BP (2012) Wireless capsule endoscopy video segmentation using an unsupervised learning approach based on probabilistic latent semantic analysis with scale invariant features. IEEE Trans Inf Technol Biomed 16:98–105. https://doi.org/10.1109/TITB.2011.2171977
36. Takiyama H, Ozawa T, Ishihara S, Fujishiro M, Shichijo S, Nomura S, Miura M, Tada T (2018) Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. Sci Rep 8:1–8
37. Than TD, Alici G, Zhou H, Li W (2012) A review of localization systems for robotic endoscopic capsules. IEEE Trans Biomed Eng 59:2387–2399
38. The Gastrointestinal Image Site, Gastrolab (n.d.). http://www.gastrolab.net/ (accessed November 17, 2020)
39. Turkoz M, Kim S, Son Y, Jeong MK, Elsayed EA (2020) Generalized support vector data description for anomaly detection. Pattern Recogn 100:107119
40. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9
41. van der Stap N, van der Heijden F, Broeders IA (2013) Towards automated visual flexible endoscope navigation. Surg Endosc 27:3539–3547
42. Verma V, Lamb A, Beckham C, Najafi A, Mitliagkas I, Lopez-Paz D, Bengio Y (2019) Manifold mixup: Better representations by interpolating hidden states. In: International Conference on Machine Learning, PMLR, pp 6438–6447
43. Vu H, Yagi Y, Echigo T, Shiba M, Higuchi K, Arakawa T, Yagi K (2010) Color analysis for segmenting digestive organs in VCE. In: 2010 20th International Conference on Pattern Recognition, pp 2468–2471. https://doi.org/10.1109/ICPR.2010.604
44. Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surveys (CSUR) 53:1–34
45. Wilks D (1990) On the combination of forecast probabilities for consecutive precipitation periods. Weather Forecast 5(4):640–650

46. Ye Y, Swar P, Pahlavan K, Ghaboosi K (2012) Accuracy of RSS-based RF localization in multi-capsule endoscopy. Int J Wireless Inf Networks 19:229–238
47. Yoo T, Choi J, Kim H (2021) Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. Med Biol Eng Comput 59(2):401–415
48. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: Beyond empirical risk minimization, ArXiv Preprint ArXiv:1710.09412
49. Zhang H, Zhang J, Koniusz P (2019) Few-shot learning via saliency-guided hallucination of samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2770–2779
50. Zheng Y, Hawkins L, Wolff J, Goloubeva O, Goldberg E (2012) Detection of lesions during capsule endoscopy: physician performance is disappointing. Am J Gastroenterol 107:554–560
51. Zou F, Shen L, Jie Z, Zhang W, Liu W (2019) A sufficient condition for convergences of adam and rmsprop. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 11127–11135