

Classification of upper gastrointestinal tract diseases using endoscopic images

Thanh-Hai Tran¹, Van-Tuan Nguyen¹, Viet-Hang Dao^{2,3}, Phuc-Binh Nguyen², Thanh-Tung Nguyen², Hai Vu¹

¹School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam

²Institute of Gastroenterology and Hepatology, Hanoi, Vietnam

³Hanoi Medical University Hospital, Hanoi, Vietnam

Article Info

Article history:

Received Jan 30, 2024

Revised Aug 25, 2024

Accepted Nov 14, 2024

Keywords:

Data augmentation

Deep learning

Endoscopic images

Focal loss

Hard samples

ABSTRACT

Automatic classification and disease detection in medical images, aided by machine learning, provide crucial support to prevent overlooked instances and ensure prompt treatment of diseases. Despite impressive achievements in the field of polyp detection from endoscopic images, classification of other diseases, such as reflux esophagitis, esophageal cancer, gastritis, gastric cancer, and duodenal ulcer, is still subject to significant limitations and remains a challenging area of study because of their different and more challenging characteristics. This paper proposes a method to roughly classify the diseases from the whole images by deep learning. In particular, we focus on identifying hard samples from the training dataset and enriching them with some fundamental augmentation techniques. We then employ a cutting-edge model, specifically ResNet, for the final classification stage. Additionally, we enhance the original ResNet's loss function by incorporating another loss function called focal loss. These modifications play a crucial role in boosting the accuracy of the ResNet model. Our proposed method outputs the disease category and corresponding heat map showing the area of interest. It achieved very promising accuracy (99.55%) for the classification of five lesions on our self-collected dataset. It serves a dual purpose. Firstly, it aids in the training of novice endoscopists, enabling them to gain valuable experience. Secondly, it offers a rapid solution for annotating extensive volumes of endoscopic image data at the label level.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Thanh-Hai Tran

School of Electrical and Electronic Engineering, Hanoi University of Science and Technology

Hanoi, Vietnam

Email: hai.tranthithanh1@hust.edu.vn

1. INTRODUCTION

Gastrointestinal endoscopic screening serves as the most important diagnostic tool for various gastrointestinal diseases. However, diagnostic outcomes from imaging heavily rely on the experience of physicians. Due to overwhelming caseloads, the application of machine learning for automated lesion diagnosis is becoming a common trend in medical image analysis [1]–[6]. Regarding the analysis of endoscopic images captured from the human digestive system, key tasks include: classifying anatomical landmarks [7], [8]; classifying diseases [9], [10]; and detecting and segmenting lesion regions [11]. This paper focuses on the task which identifies the endoscopic images of the upper gastrointestinal tract as normal or having upper gastrointestinal diseases.

Most existing methods for classifying anatomical landmarks and diseases utilize state-of-the-art deep models such as ResNet and VGG [12]–[17]. Particularly, in our previous work, we presented our study about the classification of upper gastrointestinal tract diseases from endoscopic images where all images in training and testing contain at least one disease (positive samples) [10]. However, in a practical application, there exist many images taken during the endoscopy process which do not contain a disease (negative samples). This can lead to a data imbalance problem. Besides, we observe that there are some diseases/landmarks which are very similar in appearance. Naturally, we consider them as hard samples to learn. Accurately classifying and segmenting certain lesions in endoscopic images is known to be challenging due to several factors. Firstly, different types of tissues, such as early gastritis cancer, and normal mucosa, often exhibit similar colors, textures, and shapes in these images. Secondly, lesions can vary widely in size and shape. Additionally, the boundaries between lesion tissues and normal mucosa are often indistinct. These challenges hinder the improvement of convolutional neural network (CNN)-based methods in this domain.

This paper extends our previous work [10] by incorporating anatomical landmarks as an additional category. In addition, we introduce the concept of hardness for each sample in the endoscopic image dataset, allowing us to enrich only the challenging samples in the training dataset selectively. This approach reduces the number of images to be trained, accelerates the convergence of our network, and enhances the classification performance at the same time. In summary, this paper makes three main contributions. First, we identify hard samples in the lesion classification problem from the endoscopic image of the upper gastrointestinal tract, then augment hard samples by brightness and contrast transformation to enrich the training dataset. Secondly, we present a framework for disease classification that incorporates offline hard sample augmentation and focal loss (FL), allowing us to maintain a focus on challenging samples during training. Thirdly, we collected a big dataset of six categories (normal mucosa and five diseases of gastrointestinal tract endoscopy images) and conducted extensive experiments, showing very promising classification accuracy.

2. RELATED WORKS

In this section, we review some existing works on gastrointestinal diseases classification from endoscopic images using deep learning models. We then investigate relevant works on identifying and analysis the impact of hard samples on performance of deep learning models.

2.1. Classification models

Zhu *et al.* [12] designed a CNN that extracted features from endoscopy image patches. Yang *et al.* [13] proposed a modified version of Inception for the classification of five lesions (normal, bleeding, ileal erosion, colitis, and gastritis) from wireless capsule endoscopy images. Yogapriya *et al.* [18] utilized pre-trained models VGG16, ResNet-18, and GoogLeNet to classify gastrointestinal tract diseases from images obtained using wireless endoscopy (with a dataset comprising 6702 images of 8 classes). Cho *et al.* [19] employed three combined CNN models to classify gastric lesions on grayscale images, distinguishing between advanced gastric cancer, early gastric cancer, high-grade dysplasia, low-grade dysplasia, and non-neoplastic conditions. Wang *et al.* [20] used the VGG classification model to classify endoscopic gastroesophageal reflux disease using a dataset augmented with rotation operations to address imbalance. The dataset comprised classes such as normal, grade AB, and grade CD. Cogan *et al.* [14] proposed a method that pre-processed the images (e.g. edge removal, contrast enhancement, filtering, color mapping and scaling, and gamma correction) and applied deep learning to classify an image into anatomical landmarks (such as pylorus, z-line, and cecum), a diseased state (including esophagitis, ulcerative colitis, and polyps), or a medical procedure (such as dyed lifted polyps such as dyed resection margins).

Liu *et al.* [15] proposed a method for the classification of esophageal lesions in three categories (normal cases, premalignant lesions, and cancerous lesions) using a self-designed CNN of two streams (original images and enhanced images) built upon from Inception-ResNet. In [21], [22], the authors proposed CNN combined with residual long short-term memory (LSTM) model to detect polyps in the early stages. Sharif *et al.* [23] proposed a method that combines deep CNN with geometric features to detect and classify tract diseases using wireless capsule endoscopy images. Thambawita *et al.* [16] studied the impact of resolution on the performance of lesion classification models. They experimented with ResNet and DenseNet with different resolutions. Muruganatham and Balakrishnan [17] combined a CNN model (ResNet) with a self-attention mechanism to improve the performance of wireless capsule endoscopy lesion classification. Wang *et al.* [24] introduced a two-stage convolutional-capsule network for gastrointestinal endoscopy image classification. In conclusion, all of the relevant methods for lesion classification of endoscopic images of the gastrointestinal tract applied existing deep classification models. The problem of confusion and data imbalance is still not carefully considered.

2.2. Hard samples consideration

Some studies have demonstrated that placing emphasis on challenging examples, which are either predicted incorrectly or predicted correctly with low confidence during training, can expedite convergence and enhance learning accuracy [25]–[27]. Intuitively, if a model has already predicted certain examples correctly with high confidence, those samples may not provide substantial information for further improvement of the model. During the training process, numerous samples become “well-learned” after only a few epochs, indicating that not all samples carry equal importance in guiding the training iterations. Wang *et al.* [28] proposed a method for pathological image classification based on hard example-guided CNN (tissue samples as normal, benign, in situ carcinoma, and invasive carcinoma). The authors implemented a mechanism to assign greater weight to hard examples during training. These challenging examples are characterized by a consistently low prediction probability for the ground truth class across multiple iterations. This approach helps direct the model's attention toward samples that are not yet adequately learned. Zhu *et al.* [29] built an easy/hard/noisy (EHN) detection model by using the sample training history for histopathology image classification. In this way, they are able to not only identify hard samples but also correct noisy label samples for better classification. Vandenhende *et al.* [30] introduced a three-player generative adversarial network (GAN) designed to generate challenging samples that enhance the performance of classification networks. In this approach, the generator learns to create augmented data from the training set that is difficult for the classification network to label. These augmentations involve rotations, scaling transformations, and occlusions. In summary, the majority of existing methods for endoscopic image classification are based on the most advanced CNN models. Although several augmentation techniques have been used and proven to enhance model performance, these methods mainly focus on a limited number of diseases. In addition, these methods did not effectively identify or address the challenges posed by hard patterns in the classification task.

3. PROPOSED METHOD

3.1. General framework

In this section, we present our general framework for upper gastrointestinal disease classification from endoscopic images as illustrated in Figure 1. It composes of two stages of training and one stage of testing.

- First training: It takes images from the training set D^{Tr} and trains the deep model with conventional cross-entropy loss. The training is done after k epochs. Then the hardness of each sample is computed. The ten most percentages of hardest samples are augmented to prepare a new training set called D_a^{Tr} .
- Second training: It takes images from the training set D_a^{Tr} and trains the deep model with FL.
- Testing: It takes images from the testing dataset D^{Te} and compute the evaluation metrics (accuracy) using the trained model from the second training step.

In the following, we describe in detail each component in our proposed framework.

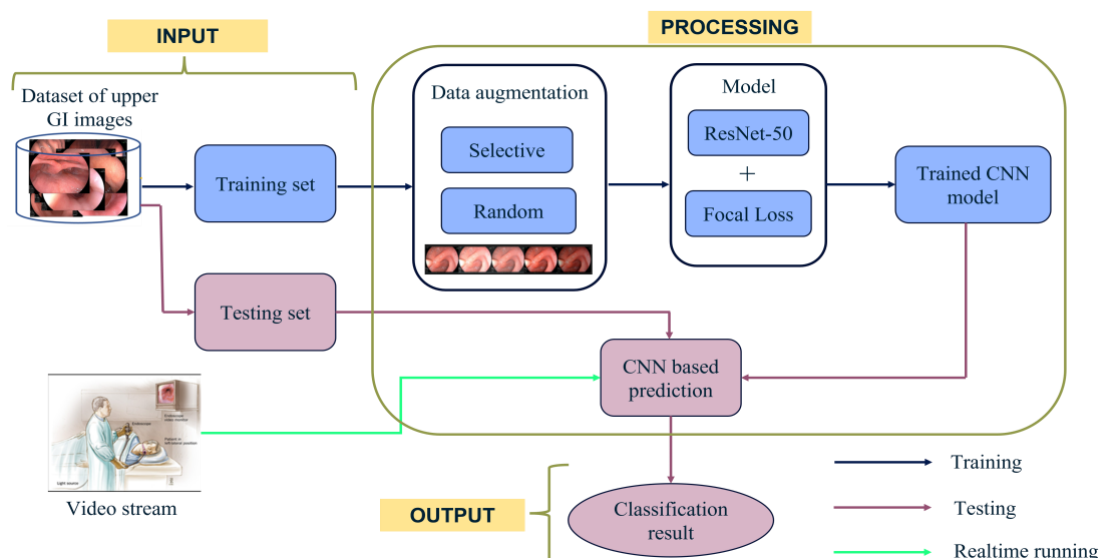


Figure 1. Our proposed framework for disease classification from endoscopic images of the upper gastrointestinal tract

3.2. Residual network for classification

Among many deep models for image classification, we deploy ResNet-50 due to its efficiency on various datasets. The ResNet network is a CNN designed to work with hundreds of layers. A problem that occurs when building a CNN network with many convolutional layers is the vanishing gradient (the phenomenon that makes the model unable to converge) leading to a bad learning process. ResNet overcomes this situation by using a uniform shortcut connection to pass through one or more layers. Each such block is called a residual block. ResNet has various versions, such as ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. In this paper, ResNet-50 is selected because it best trades off the accuracy and the computational time [31].

3.3. Focal loss imbalance data

Cross entropy loss is commonly used in classification tasks. It quantifies the model's degree of uncertainty in predicting the value of the variable. The sum of the entropy of all probability estimates is cross-entropy as in (1).

$$\mathcal{L}_{CE} = -\sum_{i=1}^N t_i \log(p_i) \quad (1)$$

where t_i is the groundtruth and p_i is the prediction probability, N is total number of training samples. The FL is specifically designed to tackle scenarios where there exists a significant imbalance between foreground and background classes during the training process. It provides a solution to mitigate the challenges posed by highly imbalanced datasets. Easily classified negatives comprise the majority of the loss and dominate the gradient. The FL primarily focuses on patterns that lead the model to fail more frequently compared to the easier patterns that the model can confidently predict. By emphasizing these challenging patterns, the FL helps the model improve its performance in handling difficult cases. FL reshapes the loss function to down-weight easy examples and thus focus training on hard negatives. This technique can be implemented by adding a variable adjustment factor to the cross-entropy loss formula as in (2):

$$\mathcal{L}_{focal} = -\sum_{i=1}^N \alpha(1 - p_i)^\lambda \log(p_i) \quad (2)$$

We note two properties of FL: i) when an example is misclassified and the prediction probability p_i is low, the modulating factor remains close to 1, leaving the loss largely unchanged. As $p_i \rightarrow 1$, the modulating factor decreases to 0, thereby reducing the loss for well-classified examples; ii) the focusing parameter λ smoothly adjusts the rate at which easy examples are down-weighted. When $\lambda=0$, the FL function \mathcal{L}_{focal} is equivalent to \mathcal{L}_{CE} . As λ is increased, the impact of the modulating factor becomes more pronounced. In our experiment, we choose $\alpha=0.25$ and $\lambda=2$.

3.4. Hardness and hard sample augmentation

According to Kishida and Nakayama [26], deep neural networks (DNNs) are capable of good generalization despite their large size and ability to memorize all data patterns. The researchers propose that DNNs initially learn from consistently classified simple data samples. However, misclassified samples at this stage can be considered difficult samples. The characteristics of easy and difficult data samples remain underexplored. The study indicates that visually, easy data samples are very similar to each other, while difficult samples are highly diverse in their visual attributes. Although difficult data samples contribute more to the model's generalization ability, removing a large number of easy data samples may impair it. Therefore, easy and difficult data samples are interconnected and constrained in the model training process.

Let us suppose we employ a deep model for our classification problem with the loss function \mathcal{L} and the f produces the prediction probability. The "hardness" of sample x_i in the training dataset $D = \{x_1, x_2, \dots, x_N\}$ with N samples is defined as in (3).

$$\text{Hardness}(x_i^k) = \frac{1}{M} \sum_{l=1}^M \mathcal{L}(t_i, f(x_i, W_l^T)) \quad (3)$$

where M is the number of training epochs, t_i is the ground truth label of the sample x_i , W_l^T is the weight of the deep model after the l^{th} epoch, f is the function that produces the prediction score for the given sample x_i by the deep model.

We define, as in (4), a sample x_i as hard after training k epochs if its Hardness (x_i^k) is larger than a threshold. Otherwise, it is an easy sample.

$$\text{Hard}(x_i) = \begin{cases} 1 & \text{if Hardness}(x_i^k) \geq \text{threshold} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Figure 2 illustrates hard samples identified by our networks after 10 epochs of training. Figure 2(a) presents challenging samples that primarily contain lesions, while Figure 2(b) displays easier samples that are mostly disease-free. We detail our selective data augmentation (DA) strategy for these hard samples. Our classification model is trained for k epochs (in our case, $k=10$) on the training dataset. Retaining the training history, we compute the hardness of each sample and isolate the top 10% hardest ones. Subsequently, we apply a brightness and contrast transformation to the selected hard samples.

In our previous work [7], we found that basic augmentation techniques enhance anatomical landmark classification in upper gastrointestinal endoscopic images. We extend this approach to enrich our challenging samples by employing brightness and contrast transformations, as described in (5):

$$I_{bc}(x, y) = cI_{org}(x, y) + bI_{avg} \quad (5)$$

where $I_{org}(x, y)$ is the original image and $I_{bc}(x, y)$ is the modified image by the brightness factor b and the contrast factor c . In our experiments, b, c are chosen as -0.15 and 0.15 , where I_{avg} is the average value of the intensity of the pixel in the original image I_{org} .

The framework, coded in Python, utilizes TensorFlow, Keras, NumPy, Matplotlib, Seaborn, and OpenCV libraries. Key settings include a learning rate of 0.001, batch size of 32, and 80 epochs for training. Pre-trained weights from ImageNet are employed for model initialization. Dropout regularization at 0.5 prevents overfitting and enhances generalization. The Adam optimizer dynamically adjusts learning rates for optimal performance.

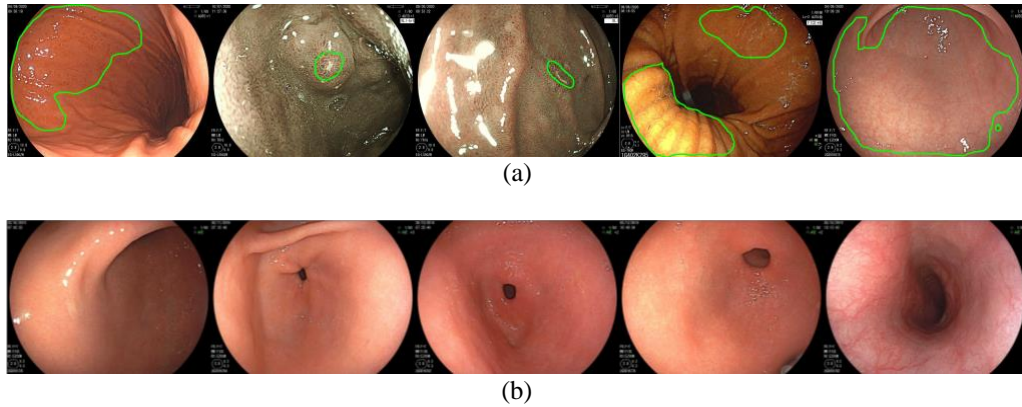


Figure 2. Illustration of hard and easy samples: (a) examples of hard samples (green lines around the lesion in each image); and (b) examples of easy samples (easy samples are usually normal anatomical landmarks)

4. EXPERIMENTS

4.1. Dataset and evaluation metrics

We collected 5546 images with no lesions of 10 anatomical landmarks (larynx, esophagus, cardia, gastric body, fundus, antrum, great curvature, lesser curvature, duodenum bulb, and duodenum) in the upper gastrointestinal tract. The images are of resolution of 1280x960 at four various lighting modes: blue laser imaging (BLI) (1271 images), flexible spectral imaging color enhancement (FICE) (1352 images), linked color imaging (LCI) (1271 images), and white light imaging (WLI) (1652 images). Besides, we collected 3420 images of five lesions: esophagitis (436 images), gastritis (1316 images), duodenal ulcer (596 images), esophageal cancer (538 images), and gastric cancer (534 images). We split the data into training, validation, and testing sets with a ratio of 18:1:1, respectively.

It is noticed that the number of images in these classes is quite imbalanced. Most images had gastritis lesions while others only accounted for 2/3 of the number of images. The inflammation lesions in the images of esophagitis, gastritis, and duodenal ulcers are quite similar. They are all small and flat lesions, so it is challenging to distinguish the lesions from the surrounding normal mucosa. Therefore, inexperienced doctors often have difficulty detecting these lesions. In most cases, the color of the esophagitis lesions is only slightly redder than the normal parts.

4.2. Experimental results

This study investigated the effects of augmentation of hard samples in the training stages and utilization of FL to deal with imbalance issues. Specifically, we evaluate five models:

- ResNet-50: Using only pure ResNet-50 architecture with original data samples.
- ResNet-50+ FL: Using the ResNet-50 architecture with FL technique, we train the model on the original dataset.
- ResNet-50+ DA (random): Performing random (non-selective) DA with 10% of training data samples, trained with the ResNet-50 model.
- ResNet-50+ DA (hard): Performing hard data sample augmentation (selectively) with 10% of the most difficult data samples, trained with the ResNet-50 model.
- ResNet-50+ DA (hard) + FL: Performing hard data sample augmentation (selectively) with 10% of the most difficult data samples, trained with the ResNet-50 model combined with the FL technique.

Table 1 presents the results showing the ability to classify gastrointestinal lesions of the proposed methods with evaluation metrics including average accuracy, precision, recall, and F1-score. We found that overall, all the techniques used have achieved very good performance for the classification model with an impressive accuracy score of over 97%. The effectiveness of the classification model is improved by combining the ResNet-50 model with the proposed techniques. The ResNet-50+ DA (hard) + FL has the best classification results, with accuracy, precision, recall, and F1-score reaching 99.55%. This signifies the exceptional classification performance of this technique, especially in the case of inflammatory lesion datasets, which is considered challenging for classification models. It is observed that the classification model utilizing this method only wrongly predicted two data samples. Moreover, the model's convergence speed, was found to be the fastest among all other methods.

Table 1. Comparison of five models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-50	97.52	97.59	97.52	97.52
ResNet-50+FL	98.20	98.23	98.20	98.19
ResNet-50+DA (random)	98.42	98.46	98.42	98.40
ResNet-50+DA (hard)	99.32	99.33	99.32	99.32
ResNet-50+DA (hard)+FL (ours)	99.55	99.55	99.55	99.55

The remaining proposed techniques also contribute to achieving good classification results for the model. The training method of the classification model with a selectively augmented dataset (ResNet-50+DA (hard)) also yielded impressive results, with the average evaluation metrics reaching 99.32%. This technique effectively prevented the model from making errors in predicting gastritis lesions, which are evaluated as difficult and prone to confusion with esophagitis lesions. However, the model exhibited misclassification errors in predicting esophageal cancer lesions, misclassifying one image as esophagitis. Moreover, two images of esophagitis were misclassified as esophageal cancer. On the other hand, when using random augmentation (ResNet-50+ DA (random)), the overall classification performance decreased by approximately 1% compared to the selectively augmented approach. With this method, the classification model misclassified three images of esophagitis as esophageal cancer and gastric cancer.

The technique of using only ResNet-50 with the original dataset yielded less satisfactory results compared to the other methods. However, the evaluation metrics for the model's classification ability still reached a relatively high level, approximately 97.5%. In difficult lesion classes such as esophagitis, gastritis, duodenal ulcer, and gastric cancer, there were images that were misclassified. Only the class of anatomical location (normal class) achieved perfect classification results. When combining the ResNet-50 architecture with FL, the evaluation metrics for the model's classification ability increased by an average of 0.7%. However, the model's convergence speed was highly impressive.

When utilizing DA techniques (selective or random), the convergence speed of the classification model improves compared to using only the ResNet-50 model with the original dataset. Additionally, combining FL with ResNet-50 also enhances the convergence speed, as the model approaches convergence within half of the training time. Consequently, when employing the ResNet-50+DA (hard)+FL technique, the classification model achieves the fastest convergence speed, thereby demonstrating improved generalization ability.

Figure 3 shows misclassified examples using the original ResNet-50. Figure 3(a) illustrates two incorrect predictions, where gastric cancer lesions were misclassified as esophageal cancer lesions. Figure 3(b) displays two actual esophageal cancer lesions. It can be observed that visually distinguishing between the two types of cancer is challenging for the naked eye. However, using the ResNet-50+DA (hard)+FL technique, these samples are no longer misclassified.

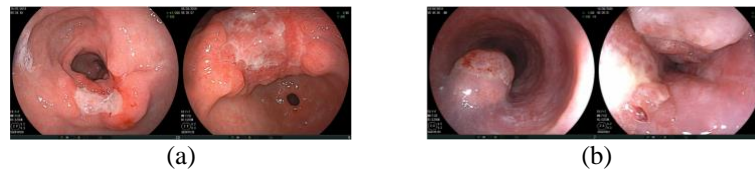


Figure 3. Two data samples were misclassified when using original ResNet-50: (a) gastric cancer lesion misclassified as esophageal cancer; and (b) examples of real esophageal cancer for reference

In addition, we employed the t-distributed stochastic neighbor embedding (t-SNE) visualization algorithm to visually represent the test data in a three-dimensional space as shown in Figure 4. As observed, before training the classification model, the data points from the six classes (duodenal ulcer, gastritis, esophagitis, gastric cancer, esophageal cancer, and normal anatomical landmarks) were mixed without clear separation into distinct clusters as shown in Figure 4(a). However, after applying the ResNet-50+DA (Hard)+FL technique, it can be seen that the data points representing the six classes are now separated into distinct clusters as shown in Figure 4(b). This indicates that our classification model achieves good classification effectiveness. Furthermore, we utilized our classification model to generate gradient-weighted class activation mapping (Grad-CAM) visualizations for the test dataset. By visualizing the Grad-CAM heatmaps, we can identify the regions of the images that the model focuses on the most before making predictions. Figure 5 presents Grad-CAM heatmaps for several examples of the five types of gastrointestinal lesions. This helps us gain a better understanding of the important features associated with these gastrointestinal lesions.

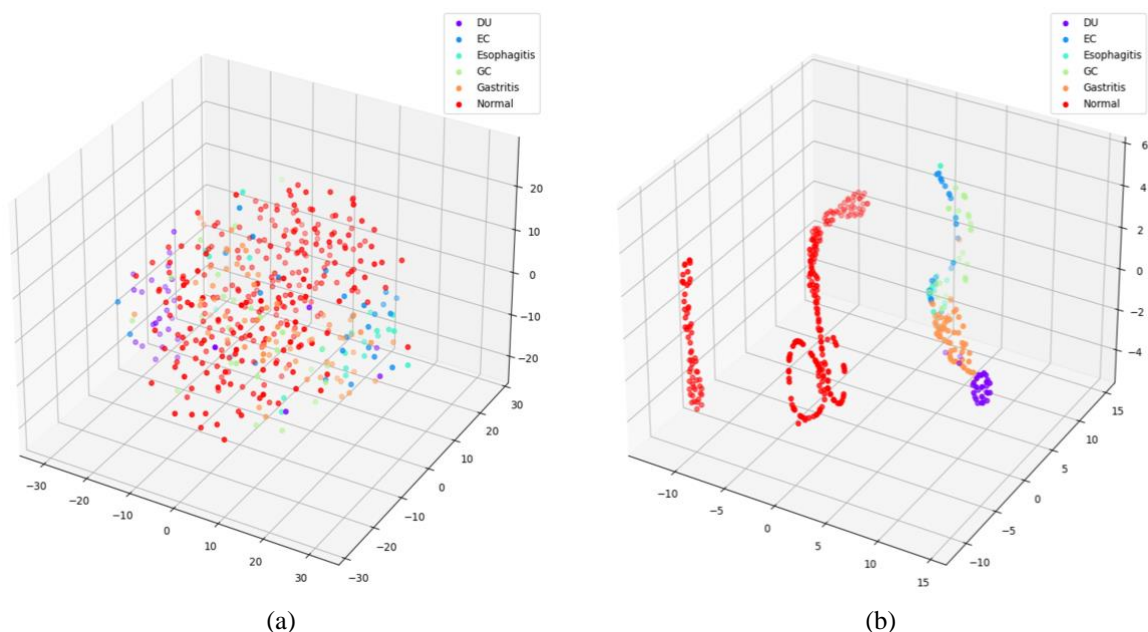


Figure 4. t-SNE visualization of our test data: (a) original testing data samples; and (b) samples with features extracted using ResNet-50+DA (hard)+FL technique

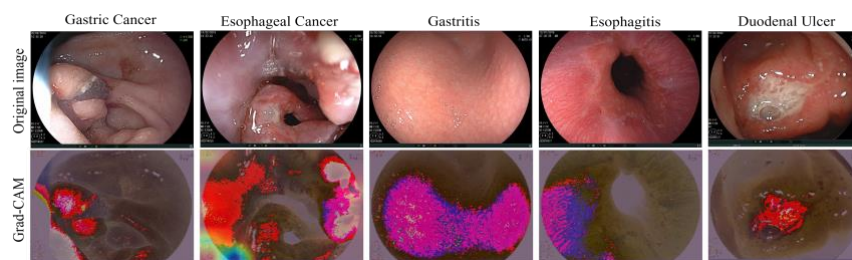


Figure 5. Grad-CAM visualization for some examples when using ResNet-50+DA (hard)+FL technique

Tran *et al.* [7] utilized the ResNet-18 and MobileNet-V2 models to perform anatomical localization classification on continuous endoscopic video streams. While ResNet-18 operated at 10 FPS and exhibited unstable recognition, it achieved high accuracy. In our work, we employed the best-performing model to analyze endoscopic videos of gastric cancer lesions, utilizing Nvidia GeForce RTX 3090 with 24 GB of RAM. Due to the larger size of ResNet-50, video processing speed remains relatively slow, at 25 FPS.

In summary, this study explored a comprehensive of classification model. The DA technique for hard samples (selective augmentation) combined with the ResNet-50 network architecture and FL technique not only addressed the data imbalance issue but also improved the generalization ability of the classification model, with an average accuracy of 99.55%. Further and in-depth studies may be needed to confirm its robustness on bigger datasets.

4.3. Discussion on hard samples

We invited two endoscopic doctors to participate in an evaluation of hard samples generated by artificial intelligence (AI) models. Our objective is to answer the question of whether the AI models and doctors have the same point of view. To this end, the doctors will carefully check by neck-eye 811 images generated by the 10% hardest samples by AI model. According to the doctors, out of the 811 images rated as difficult by the AI, 89 were also deemed difficult by the endoscopists to identify lesions or their boundaries. Esophageal and gastric cancer displayed the highest rates of difficulty in imaging, at 26.1% and 19.9%, respectively. It reflects that there is a correlation between the AI model's conclusions with human doctors.

5. CONCLUSION

We presented a framework for classifying diseases of upper gastrointestinal endoscopic images. Our framework was built upon a backbone ResNet-50 with a FL function to deal better with data imbalance. In addition, we introduced a concept of hardness and generated more artificial images on the hardest samples to enrich our training set. Our framework converged more quickly than the original model (ResNet-50) while producing higher accuracy. The finding indicates that the proposed framework can distinguish the five main diseases very well with the normal cases captured from ten anatomical landmarks. This primitive result can help inexperienced doctors to classify diseases quickly. The visualization using Grad-CAM also helps them to focus on regions of interest with the highest heat map value. In the future, we first enrich the subset of hard samples assessed again by human doctors to investigate the improvement in the performance of the proposed framework. We then integrate the proposed model on edge devices and test it with real continuous endoscopy images. We think of combining a tracking algorithm to smooth the classification result.

ACKNOWLEDGEMENTS

This research is funded by the Vietnam Ministry of Science and Technology under grant number KC-4.0-17/19-25 "Research and Develop Intelligent Diagnostic Assistance System for Upper GastroIntestinal Endoscopy Images".




REFERENCES

- [1] J. Ahmad, K. Muhammad, M. Y. Lee, and S. W. Baik, "Endoscopic image classification and retrieval using clustered convolutional features," *Journal of Medical Systems*, vol. 41, no. 12, 2017, doi: 10.1007/s10916-017-0836-y.
- [2] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
- [3] G. Wimmer, A. Vecsei, M. Häfner, and A. Uhl, "Fisher encoding of convolutional neural network features for endoscopic image classification," *Journal of Medical Imaging*, vol. 5, no. 3, 2018, doi: 10.1117/1.jmi.5.3.034504.
- [4] X. Zhang *et al.*, "Real-time gastric polyp detection using convolutional neural networks," *Plos One*, vol. 14, no. 3, 2019, doi: 10.1371/journal.pone.0214133.
- [5] M. Rana and M. Bhushan, "Machine learning and deep learning approach for medical image analysis: diagnosis to detection," *Multimedia Tools and Applications*, vol. 82, no. 17, pp. 26731–26769, 2023, doi: 10.1007/s11042-022-14305-w.
- [6] R. Hashimoto *et al.*, "Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in barrett's esophagus (with video)," *Gastrointestinal Endoscopy*, vol. 91, no. 6, pp. 1264–1271, 2020, doi: 10.1016/j.gie.2019.12.049.
- [7] T.-H. Tran *et al.*, "Classification of anatomical landmarks from upper gastrointestinal endoscopic images*," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, IEEE, 2021, pp. 278–283, doi: 10.1109/NICS54270.2021.9701513.
- [8] M. Q. Le, Q. T. Nguyen, V. H. Dao, and T. H. Tran, "CNN quantization for anatomical landmarks classification from upper gastrointestinal endoscopic images on edge devices," in *ICCE 2022 - 2022 IEEE 9th International Conference on Communications and Electronics*, 2022, pp. 389–394, doi: 10.1109/ICCE55644.2022.9852098.
- [9] P.-T. Nguyen, T.-H. Tran, V.-H. Dao, and H. Vu, "Improving gastroesophageal reflux diseases classification diagnosis from endoscopic images using stylegan2-ada," in *Artificial Intelligence in Data and Big Data Processing*, 2022, pp. 381–393, doi:





- 10.1007/978-3-030-97610-1_30.
- [10] P. T. Nguyen, M. Q. Le, Q. T. Dao, V. A. Tran, V. H. Dao, and T. H. Tran, "Automatic classification of upper gastrointestinal tract diseases from endoscopic images," in *2022 11th International Conference on Control, Automation and Information Sciences, ICCAIS 2022*, 2022, pp. 442–447, doi: 10.1109/ICCAIS56082.2022.9990445.
 - [11] T. H. Tran *et al.*, "DCS-unet: dual-path framework for segmentation of reflux esophagitis lesions from endoscopic images with u-net-based segmentation and color/texture analysis," *Vietnam Journal of Computer Science*, vol. 10, no. 2, pp. 217–242, 2023, doi: 10.1142/S2196888822500385.
 - [12] R. Zhu, R. Zhang, and D. Xue, "Lesion detection of endoscopy images based on convolutional neural network features," in *2015 8th International Congress on Image and Signal Processing (CISP)*, IEEE, 2015, pp. 372–376, doi: 10.1109/CISP.2015.7407907.
 - [13] W. Yang, Y. Cao, Q. Zhao, Y. Ren, and Q. Liao, "Lesion classification of wireless capsule endoscopy images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019, pp. 1238–1242, doi: 10.1109/ISBI.2019.8759577.
 - [14] T. Cogan, M. Cogan, and L. Tamil, "MAPGI: accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning," *Computers in Biology and Medicine*, vol. 111, 2019, doi: 10.1016/j.compbiomed.2019.103351.
 - [15] G. Liu *et al.*, "Automatic classification of esophageal lesions in endoscopic images using a convolutional neural network," *Annals of Translational Medicine*, vol. 8, no. 7, pp. 486–486, 2020, doi: 10.21037/atm.2020.03.24.
 - [16] V. Thambawita, I. Strümke, S. A. Hicks, P. Halvorsen, S. Parasa, and M. A. Riegler, "Impact of image resolution on deep learning performance in endoscopy image classification: an experimental study using a large dataset of endoscopic images," *Diagnostics*, vol. 11, no. 12, 2021, doi: 10.3390/diagnostics1112183.
 - [17] P. Muruganatham and S. M. Balakrishnan, "Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization," *Journal of Medical and Biological Engineering*, vol. 42, no. 2, pp. 157–168, 2022, doi: 10.1007/s40846-022-00686-8.
 - [18] J. Yogapriya, V. Chandran, M. G. Sumithra, P. Anitha, P. Jenopaul, and C. S. G. Dhas, "Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021, doi: 10.1155/2021/5940433.
 - [19] B. J. Cho *et al.*, "Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network," *Endoscopy*, vol. 51, no. 12, pp. 1121–1129, 2019, doi: 10.1055/a-0981-6133.
 - [20] C.-C. Wang, Y.-C. Chiu, W.-L. Chen, T.-W. Yang, M.-C. Tsai, and M.-H. Tseng, "A deep learning model for classification of endoscopic gastroesophageal reflux disease," *International Journal of Environmental Research and Public Health*, vol. 18, no. 5, 2021, doi: 10.3390/ijerph18052428.
 - [21] Ş. Öztürk and U. Özkaya, "Residual LSTM layered CNN for classification of gastrointestinal tract diseases," *Journal of Biomedical Informatics*, vol. 113, 2021, doi: 10.1016/j.jbi.2020.103638.
 - [22] Ş. Öztürk and U. Özkaya, "Gastrointestinal tract classification using improved LSTM based CNN," *Multimedia Tools and Applications*, vol. 79, no. 39–40, pp. 28825–28840, 2020, doi: 10.1007/s11042-020-09468-3.
 - [23] M. Sharif, M. A. Khan, M. Rashid, M. Yasmin, F. Afza, and U. J. Tanik, "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 33, no. 4, pp. 577–599, 2021, doi: 10.1080/0952813X.2019.1572657.
 - [24] W. Wang, X. Yang, X. Li, and J. Tang, "Convolutional-capsule network for gastrointestinal endoscopy image classification," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 5796–5815, 2022, doi: 10.1002/int.22815.
 - [25] H. S. Chang, E. Learned-Miller, and A. McCallum, "Active bias: training more accurate neural networks by emphasizing high variance samples," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1003–1013.
 - [26] I. Kishida and H. Nakayama, "Empirical study of easy and hard examples in cnn training," in *Communications in Computer and Information Science*, 2019, pp. 179–188, doi: 10.1007/978-3-030-36808-1_20.
 - [27] R. J. N. Baldock, H. Maennel, and B. Neyshabur, "Deep learning through the lens of example difficulty," in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021, pp. 10876–10889.
 - [28] Y. Wang *et al.*, "Pathological image classification based on hard example guided cnn," *IEEE Access*, vol. 8, pp. 114249–114258, 2020, doi: 10.1109/ACCESS.2020.3003070.
 - [29] C. Zhu, W. Chen, T. Peng, Y. Wang, and M. Jin, "Hard sample aware noise robust learning for histopathology image classification," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 881–894, 2022, doi: 10.1109/TMI.2021.3125459.
 - [30] S. Vandenhende, B. D. Brabandere, D. Neven, and L. V. Gool, "A three-player gan: generating hard samples to improve classification networks," in *2019 16th International Conference on Machine Vision Applications (MVA)*, IEEE, 2019, pp. 1–6, doi: 10.23919/MVA.2019.8757893.
 - [31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

BIOGRAPHIES OF AUTHORS







Thanh-Hai Tran    graduated with an engineer's degree in Information Technology from Hanoi University of Science and Technology (HUST) in 2001. She holds an M.S. degree and a Ph.D. degree in Imagery Vision Robotics from Grenoble INP, France, in 2002 and 2006 respectively. Currently, she is a lecturer/researcher at the School of Electronics and Telecommunications and Computer Vision Department, International Research Institute in Multimedia, Information, Communication and Application, HUST. Her main research interests are visual object recognition, video understanding, human-robot interaction, and text detection for applications in computer vision. She can be contacted at email: hai.tranthithanh1@hust.edu.vn.







Van-Tuan Nguyen     received a degree of engineering in computer engineering from Hanoi University of Science and Technology in May 2023. Since his third year of university, he has been actively involved in research within the research group at the International Research Institute Multimedia, Information, Communication, and Applications (MICA). Currently, he holds the position of a software engineer at FPT Software. His research interests encompass the areas of detection, classification, and segmentation of lesions in gastrointestinal endoscopy images. He can be contacted at email: tuannv618329@gmail.com.







Viet-Hang Dao     graduated from the Hanoi Medical University in 2011 and got her Ph.D. degree in 2016. She is an experienced gastroenterologist and hepatologist, a clinical researcher and a lecturer at the Hanoi Medical University. She is now the Vice General Secretariat of Vietnam Association of Gastroenterology (VNAGE) and Vietnam Association for the Study of Liver Diseases (VASLD), Vice Director of Endoscopic Centre, Hanoi Medical University Hospital, and Adjunct Assistant Professor of Johns Hopkins University School of Medicine. She is interested in application of innovative technologies, such as image-enhanced endoscopy, smart apps, artificial intelligence in endoscopy, viral hepatitis, liver cancer treatment, microbiome, and gastrointestinal motility. She can be contacted at email: hangdao.fsh@gmail.com.







Phuc-Binh Nguyen     graduated as a general practitioner from Hanoi Medical University (HMU) in 2018. He is currently working as an endoscopist and researcher at the Institute of Gastroenterology and Hepatology (IGH) in Hanoi, Vietnam. In IGH, he has been involved in several projects that use artificial intelligence (AI) to enhance endoscopy procedures, such as automatic detection of colorectal polyps or digestive cancers. His research interests are about applying new technologies, especially AI, in improving patients' experience and quality of endoscopy. He can be contacted at email: binhnguyen.fsh@gmail.com.



Thanh-Tung Nguyen     graduated from the Vietnam Military Medical University in 2017. He got his Master's Degree in 2021 at Hanoi Medical University, Vietnam. Currently, he is working as an endoscopist and researcher at the Endoscopy Centre-Hoang Long Clinic, and the Institute of Gastroenterology and Hepatology in Hanoi, Vietnam. His research area of interest is gastrointestinal endoscopy, particularly artificial intelligence in endoscopy. He can be contacted at email at: nguyentung.uh@gmail.com.



Hai Vu     received his B.E. in Electronic and Telecommunications in 1999 and M.E. in Information Processing and Communication in 2002, both from the Hanoi University of Science and Technology (HUST). He received Ph.D. in Computer Science from Osaka University, Japan, in 2009. He has been a lecturer and a researcher at the Department of Computer Vision, MICA International Research Institute (HUST-Grenoble INP), since 2012. His current research interests are in computer vision, pattern recognition, particularly, applying these techniques in agricultural engineering, medical imaging, and human-computer interactions. He can be contacted at email: hai.vu@hust.edu.vn.