AUTOMATIC CLASSIFICATION OF ESOPHAGOGASTRODUODENOSCOPY SUB-ANATOMICAL REGIONS

Diego Bravo[a], Josué Ruano[a], María Jaramillo[a], Daniel Gallego[c], Martín Gómez[c], Fabio A. González [b], and Eduardo Romero[a]

[a] Computer Imaging and Medical Applications Laboratory (CIM@LAB)
[b] Machine Learning, Perception and Discovery Lab (MindLab)
[c] Hospital Universitario Nacional de Colombia, Unidad de Gastroenterología, Bogotá, Colombia
Universidad Nacional de Colombia

ABSTRACT

Gastric cancer is the fourth most lethal malignancy worldwide. Esophagogastroduodenoscopy is the first choice procedure for diagnosis of upper gastrointestinal lesions, especially early gastric cancer. The success of this procedure depends on endoscopist's skill and the rigorous exploration of the zones with high probability of being affected. It has been documented most gastric neoplasias are lesions already existent at the examination time and unobserved when early detection is possible. For a second reader, automatic strategies must first recognize gastric anatomic regions. The aim of this paper is to assess the performance of convolutional neural networks at classifying anatomical regions. 2.054 raw upper gastrointestinal endoscopic images from 96 patients were collected and labeled as six representative sub-anatomical stomach regions. The networks were trained with transfer learning, data augmentation, and two efficient learning methods: warm-up and fine-tuning. The top-10 macro F1-score rates of the testing dataset were 84% to 87%. These preliminary tests suggest the trained networks showed good performance in recognizing sub-anatomical stomach regions of esophagogastroduodenoscopy images.

***Index Terms*—** Gastric cancer, esophagogastroduodenoscopy, CNN, sub-anatomical regions, computer-assisted.

1. INTRODUCTION

Gastric cancer (GC) is the fourth most common cause of cancer death worldwide and the fifth most common malignancy [1]. Despite the incidence decreasing in some world regions, gastric cancer remains a major clinical challenge since most cases are diagnosed in advanced stages, i.e., poor prognosis and limited treatment options. In recent decades, endoscopic technology has seen advances and is widely used as a screening test for early gastric cancer (EGC) [2]. Esophagogastroduodenoscopy (EGD) is a diagnostic endoscopic procedure that includes visualization of the esophagus, stomach, and proximal duodenum. However, gastroenterologists have documented to miss between 20%–40% for EGC [3]. The Japan Gastroenterological Endoscopy Society developed a guideline for endoscopic diagnosis of EGC [4], mainly focused on the technical skills to examine the upper gastrointestinal tract. During the endoscopy, to avoid blind spots, K. Yao proposed a systematic screening protocol for the stomach (SSS) [5]. Overall, the SSS comprises a series of endoscopic photos of four quadrants of the gastric antrum, body, and middle–upper body. In practice, guidelines to map the entire stomach do exist but they are often partially followed, especially in developing countries [6]. Therefore, it is desirable to develop reliable methods to alert endoscopists about possible EGC lesions and blind spots. In this context, a potential solution is to apply a computer-aided diagnosis system to improve the daily clinical workflow, becoming a "third eye or second reader" for gastroenterologists. In recent years, convolutional neural networks (CNNs) have been broadly applied in the medical domain [7], particularly in endoscopy [8], even though most researchers have focused on detecting lesions, and little attention has been paid to assessing the quality of the endoscopy routine. Therefore, AI algorithms are required to automatically recognize anatomical landmarks of the upper gastrointestinal that can be integrated with the actual exploration procedure.

Takiyama et al. [9] use GoogleNet architecture to recognize four categories (larynx, esophagus, stomach, and duodenum), classifying the anatomical location correctly for 16.632 (97%) out of 17.081 images. Wu et al. [10] divided the gastric locations into 10 and subdivided it into 26 anatomical parts, the CNN correctly identified EGD images with accuracy rates were 90% and $65,9\%$ respectively. Based on CNN and deep reinforcement learning were monitored blind spots with an average accuracy of 90,02% in 107 videos. [11]. Li et al. [12] trained an Inception-V3 and LSTM with 170.297 frames and 3.100 EGD images for testing, the authors reported performance of CNN for recognition of gastric sites in images were $97,18\%$, $99,91\%$, and $99,83\%$ of sensitivity, specificity, and accuracy respectively. Recently, Chang et. al. [13] trained a ResNeSt architecture with 15.305 images and the model was tested with 1.330 frames obtaining an accuracy of $96,64\%$. All these strategies are studies that have presented similar tasks using CNNs on EGD images. Still, a comparison of the performance of different state-of-the-art architectures under similar conditions has not been made.

The main contributions of this paper are two-fold: First, as far as we know, this work is the first study detecting actual stomach regions with a benchmark study of 23 representative deep neural network architectures. Those nets were customized using a transfer learning strategy comprising warmup and fine-tuning phases, and 4 hyperparameter optimization. Second, this benchmark study can direct future research on applying automatic image classification of gastric regions classification.

2. METHODOLOGY

Upper gastrointestinal endoscopy or EGD represents the most common procedure in gastroenterology and therefore the most important to be performed with a minimum of quality. In practice, it has been reported that the exploration protocol and the order of visiting

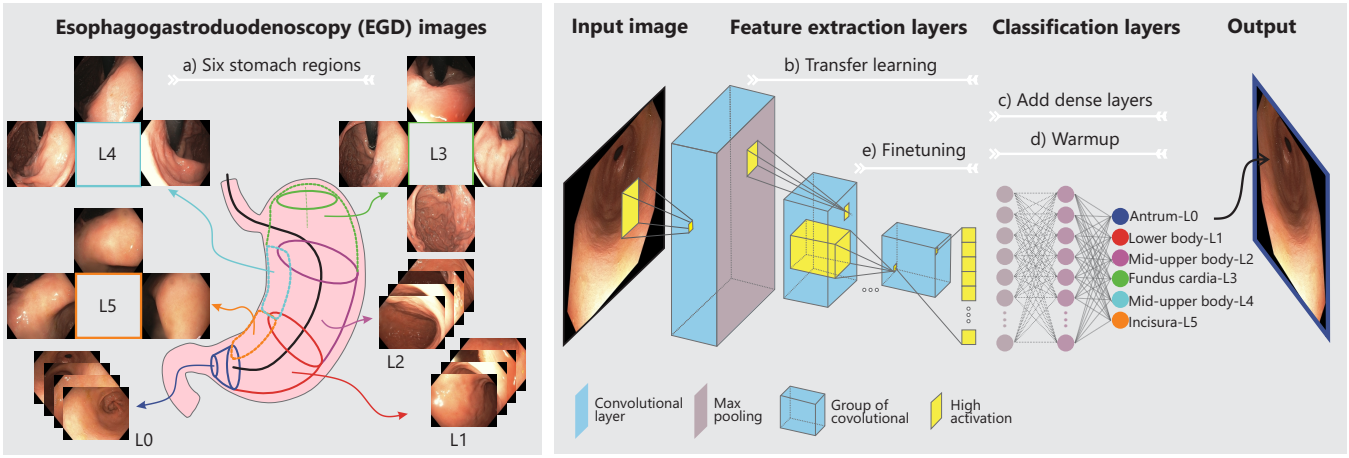

Fig. 1. Pipeline of the proposed approach. (a) anatomical EGD sub-category (see Section 3.1.1). In (b-c) transfer learning and added dense layers, (d-e) warm-up and fine-tuning stage (see Section 3.1.2-3.1.3), and output to classify the stomach regions according to sub-category.

stomach regions may hinder the discovery of incipient lesions [14]. Furthermore, there is no trace of how a specialist carries out the procedure, or even if a minimum of regions were visited or for how long they were observed. In this scenario, a second reader should not only recognize the gastric anatomic regions of interest but also ensure they are followed during a minimal time. As noted by Lee et al. [14], learning the normal anatomical features of the stomach is crucial for gastroenterologists to distinguish between the fundus, body, and antrum during endoscopic procedures. However, mastering these skills can be challenging due to the long learning curve. Furthermore, even with training, distinguishing between these regions can still be difficult due to their similar appearance through the endoscope, which is a serious limitation for this task, as experienced gastroenterologists have noted.

2.1. Stomach region classification pipeline

In this work, we explore the feasibility of automatically classifying different stomach regions, which should be visited during endoscopic procedures, using state-of-the-art deep image classification methods. Figure 1.a shows the location of regions of interest in the stomach along with sample images. Figure 1.b-d illustrates the pipeline proposed in this work. The strategy used is transfer learning, which leverages the ability of pre-trained deep models to capture low-level visual features and use them as input for a classification module. The first stage of the pipeline, feature extraction, corresponds to a CNN that has been pre-trained with a large set of natural images. The second stage, classification, corresponds to a set of dense layers which are trained using the problem-specific set of training images. The output of the model is a softmax layer with six neurons corresponding to the six different stomach region classes.

For the feature extraction stage, different state-of-the-art CNN architectures were considered: AlexNet, the family of VGG architectures (VGG-11-13-16) without batch normalization layers, Inception-V3, GoogleNet, ResNet-18-34-50-101-152, EfficientNet-B0-V2L, DenseNet-121-161-169-201, SqueezeNet-1_0-1_1, MobilNet-V3, MnasNet-0_5-1_3 and ConvNext_tiny[15].

2.2. Dataset

The database consists of 96 patients who underwent EGD procedures. From the recorded video in white light, 2.054 anatomical frames were obtained and manually labeled into six anatomical locations (see Table 1 columns 1-2) by an expert according to the systematic stomach screening protocol [5]. Each frame was captured at a spatial resolution of 1.350×1.080 pixels. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of the Hospital Universitario Nacional (approval number: CEI-2019-06-10).

2.3. Model training

The model is trained in two phases using a standard transfer learning and fine-tuning approach. In the first phase, warmup, only the classification layers are modified by the training process and the feature extraction layers are frozen. This allows the model to quickly learn the transformation of the features that better capture the patterns of the region images. In the second, phase, fine-tuning, the feature extraction layers are unfrozen, this permits a further adaptation of the visual features to the problem.

3. EVALUATION AND RESULTS

We introduce in this section the assessment and performance comparison of 23 architectures in the proposed classification tasks.

3.1. Experimental Setup

3.1.1. Dataset

The models were challenged under a $70 - 30$ evaluation scheme of the overall patients: 70% (67 cases - 1.383 frames) selected for training-validation and 30% for testing (29 cases - 671 frames) as presented in the Table 1.

L	Sub Category	Training (n=58)	Validation (n=9)	Testing (n=29)
L0	Antrum	197	26	129
L1	Lower body	230	45	118
L2	Middle-upper body	221	31	119
L3	Fundus-cardia	220	35	122
L4	Middle-upper body	171	29	89
L5	Incisura	146	32	94

Table 1. Distribution of EGD database for the validation scheme (n: patient, L: label, L3 to L5 correspond to retroflex view).

The input of the architectures were RGB images with shape $3 \times H \times W$, where H and W are 299 pixels for Inception-V3 and 224 pixels for all the other models considered. Additionally, the training and validation sets were balanced by the number of frames at twice the predominant class, and random data augmentation was applied, including vertical or horizontal flips and rotations ($\pm 5°$) were used to modulate capture variability. The number of frames per class was set to 460 for the training set and 90 for the validation set. The unbalanced proportion for the testing data set was preserved.

3.1.2. CNN configuration

In order to perform a direct comparison, the CNNs were trained in two stages: (a) a warmup of the classification layers with a constant learning rate during 10 epochs, and subsequently, (b) a fine-tuning to the last 20% features layers during 40 epochs. Details of CNNs and training configuration are presented below:

- **Pre-trainned weights:** ImageNet.
- **Optimizer**: Adam
- **Loss function:** Cross entropy.
- **Dense layers:** for ResNet, DenseNet, EfficientNet, MnasNet, Inception-V3, and GoogleNet were added extra-layers: a dense layer, then a batch Normalisation layer then a dropout layer, and finally two dense layers with the output of 6 sub-anatomical categories presented in the in Table 1.

3.1.3. Training and hyper-parameter optimization

The warmup and fine-tuning stages were included in a hyperparameter optimization across 40 trials monitoring the F-measure, to find an optimal batch size, initial learning rate, and learning rate schedule (gamma and step size). The values during optimization were the following:

- **Learning rate for warmup:** $0,001$ with gamma $= 0,1$.
- **Range of hyperparameter values during optimization:** batch size [5-30], gamma [0, 1-0, 5], step size [5-10], and learning rate [$1e^{-3}$ to $1e^{-5}$].

3.2. Results

For each CNN, the model with the highest validation f1-score across the trials was challenged with the testing set. The results were provided in four scenarios ($A - D$) using the same classification metric.

A. General CNN Results

The results of the proposed approach transfer learning, warmup, and fine-tuning are listed in a top 10 macro-f1 score (see Table 2). Clearly, ConvNext_tiny and the family ResNet and VGG yield consistently better results than other CNNs.

CNN [%]	Acc.	Prec.	Recall	F1
ConvNext_tiny	**87,332**	**87,251**	**87,332**	**87,256**
ResNet152	86,438	86,669	86,438	86,294
VGG13	85,544	85,496	85,544	85,390
VGG11	85,395	85,253	85,395	85,227
VGG16	84,948	85,004	84,948	84,861
AlexNet	84,501	84,692	84,501	84,454
DenseNet201	84,650	84,735	84,650	84,409
ResNet34	84,650	84,958	84,650	84,250
ResNet101	84,352	84,679	84,352	84,250
SqueezeNet1_1	84,352	84,464	84,352	84,181

Table 2. Top-10 CNNs performance macro F1-score (Acc: Accuracy, Prec: macro-Precision, Recall: micro-Recall, F1: macro-F1).

B. Stomach region (sub-anatomical) results

ConvNext_tiny architecture achieved the best $TP = 586$ and $TN = 3.270$, also with the smallest $FP, FN = 85$ compared to all other models. However, we observe a decrease in the performance between the lower body vs the mid-upper body (L1-L2 in antegrade view), and fundus-cardia vs middle-upper body (L3-L4 as presented in Figure 2), these stomach regions are captured in retroflex view and at the inverted axis where the endoscope is observed in the frame as presented in Figure 3.

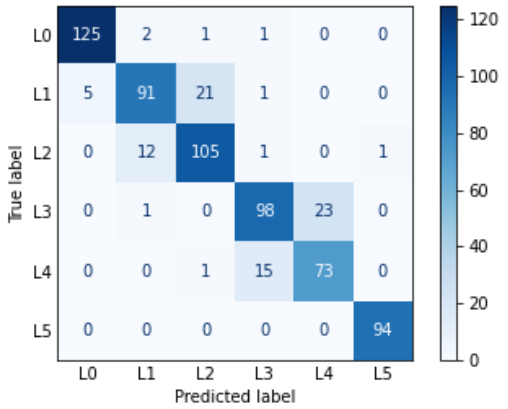

Fig. 2. Confusion Matrix ConvNext tiny.

C. Qualitative Results

Figure 3 displays prediction cases for L3 and L4 labels. True positives (L3) are explored with a retroflex view from fundus-cardia capturing the lesser curvature (P14) and anterior wall (P53) views. In addition, True negative (L4) examples photo documented the middle-upper body in retroflex view observing the posterior wall (P42) and anterior wall (P95). In contrast, False negative examples include frames with lesser curvature (P45-P52) and False positives with the anterior wall (P33-P87). Also, misclassified frames occur in images that are similar like P53 (TP) vs P87 (FP) and P52 (FN) vs P95 (TN).

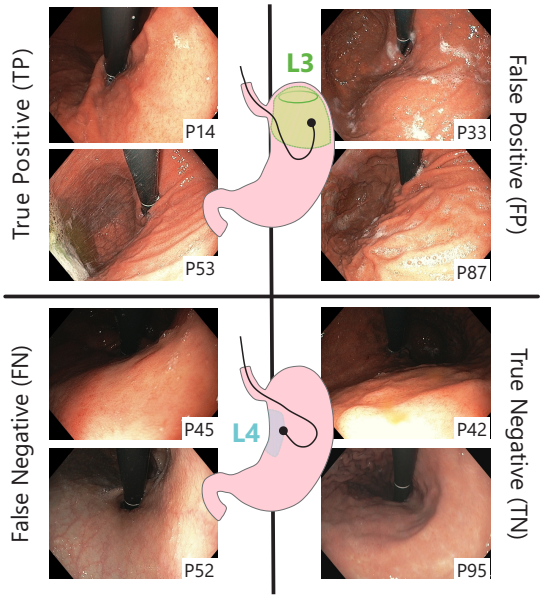

Fig. 3. Examples of different classification outcomes for images of classes L3 and L4 (ConvNext tiny). The captions in the images represent the testing patient (i.e P95: Patient 95).

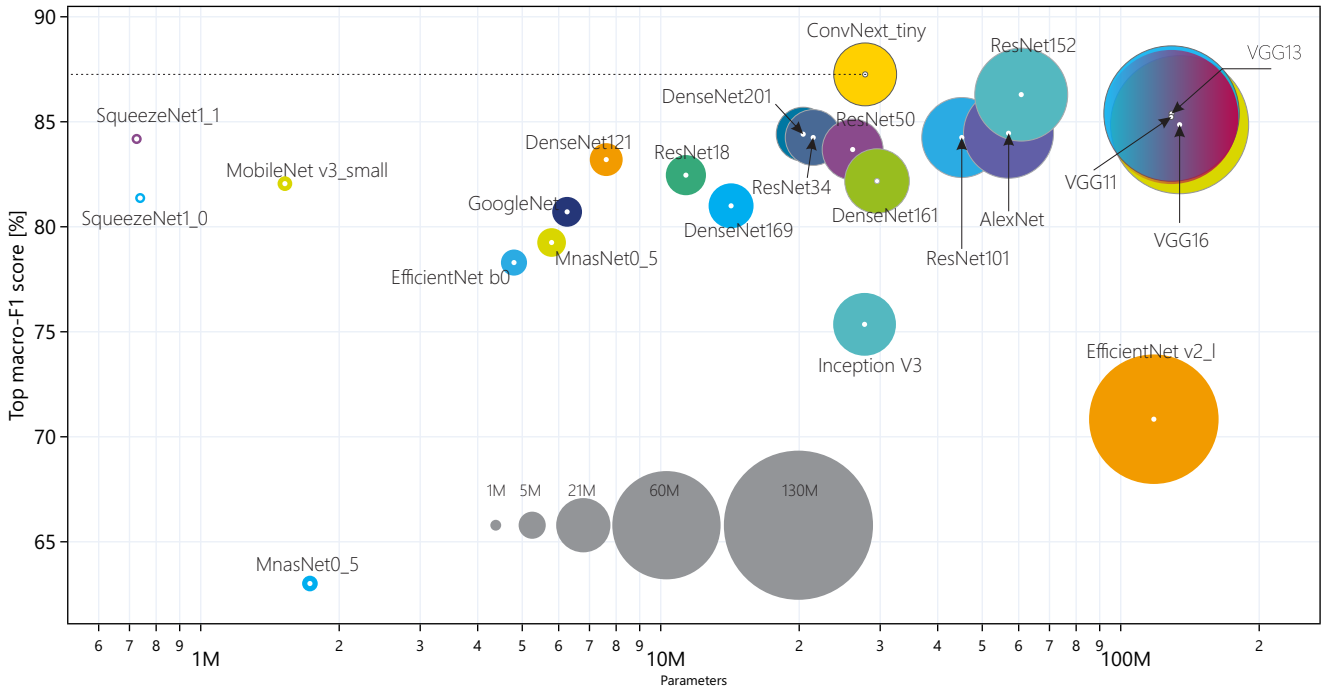

Fig. 4. Ball chart reporting the top macro F1-score vs total parameters (feature extraction layers + classification layers).

D. Macro-F1 score and Model parameters

In Figure 4 we analyze the relationship between the model parameters (model complexity) and the macro f1-score metric. The comparison among the performance of CNNs show the importance of specific configurations, as relevant differences in the number of parameters in order to learn the sub-anatomical class. It has to be noted that larger parameters more time to predict. In general, architectures with a relatively low number of parameters, such as the ConvNext_tiny achieve a higher f1 score than the VGG family (see Figure 4). Additionally, there is not a linear relationship between model parameters and f1 score metric.

4. CONCLUSIONS AND DISCUSSION

Comparative performance of different CNNs under similar conditions was the main interest of this paper. However, different training strategies and/or hyperparameter settings should be evaluated depending on the architecture.

The lack of a common validation framework is a frequent problem in endoscopy image analysis and this has limited the comparison between existing approaches, it is difficult to determine which of them could have the actual advantage in clinical use. Then, the next step is to construct a public database from patients of Hospital Universitario Nacional de Colombia to establish a baseline for a comparative study of 22 anatomical stations (complete guidelines by K. Yao [5]).

5. ACKNOWLEDGMENT

This work was partially funded by project 92384 “Design of an audit system for the automatic esophagogastroduodenoscopy (EGD) procedure” from MinCiencias resources.

6. REFERENCES

[1] Hyuna Sung, Jacques Ferlay, et al., “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] Victor Pasechnikov, Sergej Chukov, et al., “Gastric cancer: prevention, screening and early diagnosis,” *World journal of gastroenterology: WJG*, vol. 20, no. 38, pp. 13842, 2014.

[3] Mitsuru Kaise, “Advanced endoscopic imaging for early gastric cancer,” *Best Practice & Research Clinical Gastroenterology*, vol. 29, no. 4, pp. 575–587, 2015.

[4] Kenshi Yao, Noriya Uedo, et al., “Guidelines for endoscopic diagnosis of early gastric cancer,” *Digestive Endoscopy*, vol. 32, no. 5, pp. 663–698, 2020.

[5] Kenshi Yao, “The endoscopic diagnosis of early gastric cancer,” *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, vol. 26, no. 1, pp. 11, 2013.

[6] Gwang Ha Kim, Sung Jo Bang, et al., “Is screening and surveillance for early detection of gastric cancer needed in korean americans?,” *The Korean Journal of Internal Medicine*, vol. 30, no. 6, pp. 747, 2015.

[7] Geert Litjens, Thijs Kooi, et al., “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[8] Dexin Gong, Lianlian Wu, et al., “Detection of colorectal adenomas with a real-time computer-aided system (endoangel): a randomised controlled study,” *The Lancet Gastroenterology & Hepatology*, vol. 5, no. 4, pp. 352–361, 2020.

[9] Hirotoshi Takiyama, Tsuyoshi Ozawa, et al., "Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[10] Lianlian Wu, Wei Zhou, et al., "A deep neural network improves endoscopic detection of early gastric cancer without blind spots," *Endoscopy*, vol. 51, no. 06, pp. 522–531, 2019.

[11] Lianlian Wu, Jun Zhang, Wei Zhou, Ping An, Lei Shen, Jun Liu, Xiaoda Jiang, Xu Huang, Ganggang Mu, Xinyue Wan, et al., "Randomised controlled trial of wisense, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy," *Gut*, vol. 68, no. 12, pp. 2161–2169, 2019.

[12] Yan-Dong Li, Shu-Wen Zhu, et al., "Intelligent detection endoscopic assistant: An artificial intelligence-based system for monitoring blind spots during esophagogastroduodenoscopy in real-time," *Digestive and Liver Disease*, vol. 53, no. 2, pp. 216–223, 2021.

[13] Yuan-Yen Chang, Hsu-Heng Yen, et al., "Upper endoscopy photodocumentation quality evaluation with novel deep learning system," *Digestive Endoscopy*, vol. 34, no. 5, pp. 994–1001, 2022.

[14] Seung-Hwa Lee, Young-Kyu Park, et al., "Technical skills and training of upper gastrointestinal endoscopy for new beginners," *World Journal of Gastroenterology: WJG*, vol. 21, no. 3, pp. 759, 2015.

[15] Simone Bianco, Remi Cadene, et al., "Benchmark analysis of representative deep neural network architectures," *IEEE access*, vol. 6, pp. 64270–64277, 2018.