Automated Anatomical Classification and Quality Assessment of Endoscopy by Temporal-Spatial Analysis

Diego Bravo*, Sebastian Medina‡, Josué Ruano*, María Jaramillo*, Martin Gómez†, Fabio A. González‡ and Eduardo Romero*[1]

*Computer Imaging and Medical Applications Laboratory (CIM@LAB)
†Gastroenterology unit, Hospital Universitario Nacional, Bogotá, Colombia
‡Machine Learning, Perception and Discovery Lab (MindLab)
Universidad Nacional de Colombia, Bogotá, Colombia

***Abstract*—Gastric cancer is the fourth deadliest cancer worldwide. Esophagogastroduodenoscopy (EGD) is the preferred method to diagnose upper gastrointestinal lesions, particularly early gastric cancer. The procedure's success relies on the endoscopist's experience and a comprehensive examination by observing a set of anatomical landmarks. Most gastric neoplasias are undetected during early stages, despite being present during examinations, thus, it is essential to evaluate the quality and audit the examination of anatomical regions during the endoscopy procedure. This study assesses the performance of a recurrent neural network and transformer architecture in classifying anatomical and sub-anatomical regions within the gastrointestinal tract. By leveraging temporal information, the study aims to enhance the accuracy of detecting these critical regions. We collected and labeled video endoscopies from 32 patients, organizing them into four organ categories. Additionally, we utilized 565 labeled sequences from six sub-anatomical stomach regions for a separate classification task. The trained networks achieved a macro F1-score of 87.25% for organ classification and an 85.31% in identifying stomach regions. These findings provide substantial evidence supporting that temporal information improves the capabilities of accurately identify upper gastrointestinal regions.**

***Index Terms*—Endoscopy, Sequences, Sub-anatomical region, Classification, GRU, Transformer encoder, Quality indicator.**

I. INTRODUCTION

Gastric cancer (GC) ranks as the fourth leading cause of cancer-related mortality globally and stands as the fifth most prevalent malignancy [1]. Despite decreasing incidence in some regions, gastric cancer remains a clinical challenge as most cases are diagnosed at advanced stages, resulting in poor prognosis and limited treatment options. Endoscopic technology has advanced in recent decades and is now widely used for early gastric cancer screening [2]. Esophagogastroduodenoscopy (EGD) is a diagnostic procedure that visually examines the esophagus, stomach, and proximal duodenum. However, gastroenterologists have reported missing 20%-40% of early gastric cancer (EGC) cases during EGD [3].

Upper gastrointestinal (GI) endoscopy is the most prevalent procedure in gastroenterology and holds paramount importance in terms of ensuring the accurate detection of pre-malignant and malignant lesions. However, acquiring proficiency can be challenging due to the significant learning curve involved. In this context, The Japan Gastroenterological Endoscopy Society (JGES) has developed guidelines specifically focused on the technical skills required for the endoscopic diagnosis of early gastric cancer (EGC) [4]. Furthermore, to enhance procedural efficiency and quality, two important indicators were introduced. Firstly, the time spent on organ exploration serves as a significant factor [5]. Additionally, a systematic screening protocol for the stomach (SSS) was implemented, consisting of a series of endoscopic photos capturing the four quadrants of the gastric antrum, body, and middle-upper body [6]. It is worth noting that while guidelines exist to map the entire stomach, adherence to these guidelines tends to be partial, particularly in developing countries [7].

In this context, integrating computational methods into the medical practice can substantially support gastroenterologists in adhering to the guidelines established for endoscopic procedures such as mandatory anatomical regions to be examined and minimum recommended procedure times [8], [9]. Such methods should act as a navigational second reader, providing clarity on the current anatomical location within the patient's upper gastrointestinal system and quality indicators such as time spent at particular organs or stomach sub-regions, ensuring comprehensive examination, and avoiding blind spots. Incorporating these methods can substantially enhance the efficiency of the endoscopy procedure by reducing the likelihood of overlooked or unexamined areas, reducing the miss-rate of pre-cancerous and malignant lesions, and significantly contributing to early diagnosis and treatment [10].

This paper presents two key contributions: first, we present a method that leverages temporal and spatial information derived from EGD videos to classify upper-gastrointestinal organs and stomach sub-anatomical regions to automatically identify anatomical landmarks and integrate them into the exploration procedure. Second, capitalizing on the temporal information, we quantify the organ and procedure exploration time. This measure is widely accepted as a quality indicator for EGD

[1] Corresponding author, E-mail: edromero@unal.edu.co (Dr. E. Romero)

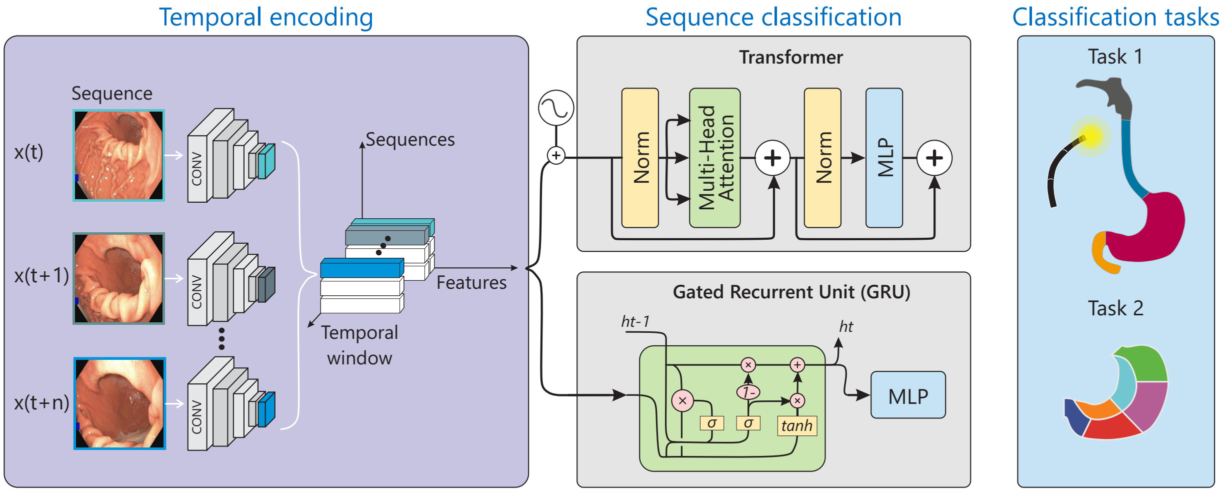

Fig. 1. The proposed pipeline begins with a feature extraction step, represented by the purple box, using a Convolutional Neural Network (CNN) to extract relevant features from sequences of frames (clips) from specific organs or stomach regions. These features, extracted on a frame-by-frame basis, are integrated into a sequence that represents original clips of varying lengths. Following this, two sequence-based methods, depicted in gray boxes, leverage the spatial and temporal information within the clips. These methods involve a Transformer Encoder Block and a Gated Recurrent Unit (GRU), focusing on two main classification tasks: Organ Classification and Stomach Region Classification.

among endoscopists [11]. These contributions could support experts during the procedure, ensuring adherence to high-quality standards and reducing blind spots while exploring the stomach.

In the field of EGD image classification, researchers have explored two main approaches: spatial and temporal. Spatial methods involve frame classification using CNN architectures. For instance, Takiyama et al. [12] utilized a GoogleNet architecture to accurately recognize 4 anatomical locations (larynx, esophagus, stomach, and duodenum), as well as 3 subsequent sub-classifications specifically for stomach images. Wu et al. [13] employed a VGG-16 network to classify gastric locations into ten categories, and further refined the classification into 26 anatomical parts (22 for the stomach, 2 for the esophagus, and 2 for the duodenum). Additionally, Chang et al. [14] trained a ResNet architecture to classify EGD images into 8 anatomical locations, with an additional location for the pharynx. In a temporal analysis, Li et al. [15] leveraged adjacent frames and trained a combination of Inception-V3 and Long short-term memory (LSTM) models to classify EGD images into 31 sites, ranging from the hypopharynx to the duodenum.

Current research has overlooked the potential of extracting temporal context from classification tasks in EGD. This oversight becomes evident when considering the straightforward quantification of time spent at specific organs or stomach subregions through the classification of video sequences. Such temporal data, made readily available through classification, is a critical factor in EGD quality assessment and holds significant value in the medical field. Furthermore, there is a noticeable gap in research comparing the performance of different architectures, particularly those employing video classification approaches. As such, the exploration of sequence classification architectures, especially those incorporating temporal information under equivalent conditions, remains unexplored.

II. METHODOLOGY

In this work, we explore the potential of spatiotemporal information to automatically classify organs and stomach regions explored during video endoscopic procedures. Figure 1 presents an overview of our proposed approach, which consists of two stages: the sequence embedding stage (see Figure 1 purple box), where a sequence tensor is generated by concatenating extracted features from a CNN from contiguous frames of the video endoscopy procedure, and the classification stage (see Figure 1 gray boxes), which employs two sequence-based methods, namely a Recurrent Neural Network (RNN) and Transformer encoder block that leverage the temporal information present in EGD video sequences. The model addresses two tasks (see Figure 1 blue box): the first task focuses on classifying four different upper-GI organs and out-of-body sequences, while the second task involves classifying six distinct key regions of the stomach. By employing these sequence-based methods, we aim to effectively capture temporal dependencies and intricate patterns in the video sequences. This adaptability and flexibility in capturing temporal context is particularly significant for our research, as it allows us to determine an optimal temporal window for contextual information. Our chosen models, such as the GRU and Transformers, excel in learning complex temporal relationships, rendering them particularly well-suited for the analysis of video endoscopies.

A. Feature extraction

Frames are extracted from EGD videos at 30 frames per second. We use a pre-trained ConvNeXT tiny [16] CNN model to extract visual features of 224 by 224-pixel EGD video frames. This CNN model, initially trained for stomach subregion classification, serves as a feature extractor for our classification tasks. We obtain 768 features from the last average pooling layer to represent the spatial information of each frame.

B. Temporal information encoding

To efficiently process and analyze temporal data, the temporal information was encoded by creating tensors with features of adjacent temporal windows of frames (see Figure 1 purple box). This structure enables efficient processing and analysis of sequential data. It consists of three dimensions: sequence dimension for batch processing, temporal windows for capturing dependencies across frames, and features dimension for detailed characteristics within each frame. An effective approach to leveraging the time dimension was established with the GRU and Transformer architectures (see gray boxes Figure 1) to model the relationship between consecutive frames, learn the temporal dynamics within the data, and focus on invariant spatiotemporal information.

C. Organ and Stomach classification tasks

In our research, we set up two distinct classification tasks, each addressed separately. The initial task focuses on identifying upper-GI organs, specifically distinguishing out-of-body, pharynx, esophagus, stomach, and duodenum regions during the procedure. The predictions obtained enable the assessment of the examination time dedicated to each anatomical structure providing a valuable quality indicator. This task is pertinent to the quality evaluation of the procedure, as a recommended minimum examination time is associated with each specific organ.

The following task was structured to pinpoint six specific sub-anatomical regions within the stomach as integral to the classification process. These regions include in the antegrade view: the Antrum, lower body, middle-upper body, and in the retroflex view: fundus-cardia, middle-upper body, and incisura, all of which are essential examination sites during esophagogastroduodenoscopy as dictated by the protocol. The objective is to eradicate blind spots and guarantee a systematic and exhaustive examination of the stomach, which stands as the main organ under focus during EGD.

D. Automatic quality assessment of the procedure

For the organ classification task, the primary objective was to develop an automated system to identify anatomical structures in the upper gastrointestinal tract while providing an indicator that assesses the overall quality of the procedure. This task necessitated pinpointing specific locations within the digestive system with accuracy. A quality indicator was formulated, based on the time exploration of each organ, to audit and measure how the procedure correlates with the goals of a correct upper-GI examination, guaranteeing a minimum required exploration time per organ. Figure 2 provides a visual depiction of the exploration time quality indicator derived from upper-GI organ classification.

To compute this exploration time, the inference is performed over the entire procedure. Independent of the temporal windows chosen during training, our model provides a classification of each frame of the procedure. The time spent on each organ can be effectively calculated by coupling the per-frame predicted organ with the frame rate.

E. Dataset

The database consists of 32 patients who underwent EGD. The recorded video was in white light and Narrow Band Imaging (NBI) and captured at 30 frames per second, all videos were manually labeled by a resident in gastroenterology. Additionally, 565 anatomical sequences were manually labeled into six anatomical locations (see Table I columns 3-4) by a resident according to the systematic stomach screening protocol [6]. Each video was captured at a spatial resolution of $1,920 \times 1,080$ and $1,200 \times 720$ pixels.

TABLE I
DISTRIBUTION OF EGD DATABASE

Category Organ	**Sequences (n=32)**	**Category Stomach**	**Sequences (n=26)**
Out body	7,662	Antrum (a.v)	88
Pharynx	10,536	Lower body (a.v)	101
Esophagus	47,531	Middle-upper body (a.v)	100
Stomach	442,379	Fundus-cardia (r.v)	103
Duodenum	41,128	Middle-upper body (r.v)	79
Total	549,236	Incisura (r.v)	94

Abbreviations - n: patient, a.v : antegrade view, r.v: retroflex view

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).

III. EVALUATION AND RESULTS

A. Experimental Setup

The organ classification task involved sampling the entire video with a frame step size of 1 where the center frame represents the sequence label, thereby aligning with $sequencesVideo = framesVideo - (temporalWindow - 1)$. This experiment was conducted using a 90-10 split for training and validation comprising 29 cases (∼465,959 sequences) and 3 patients (∼82,829 sequences) with a complete procedure for testing. For sub-anatomical stomach region classification, a 70-30 evaluation scheme was utilized. This involved selecting 70% of the overall patients, amounting to 20 cases (422 sequences), for training and validation. The remaining 30% of patients, representing 6 cases (143 sequences) were reserved for testing purposes.

Two distinct architectures were employed. The GRU model was configured with 128 hidden units and a dense layer for classification. Simultaneously, the Transformer model was designed with 8 attention heads and 2 transformer blocks, with input feature vectors linearly projected down to a dimension of 512. A dropout layer with a probability of $P = 0.45$ was incorporated to mitigate overfitting. The optimal configurations for both models were determined through automated hyperparameter optimization using Optuna, over the course of 100 trials [17]. The general configuration shared by both models includes the Adam optimizer, with a learning rate of 0.0001, and utilizes cosine annealing with warm restarts with 10 iterations for the first restart, followed by a doubling

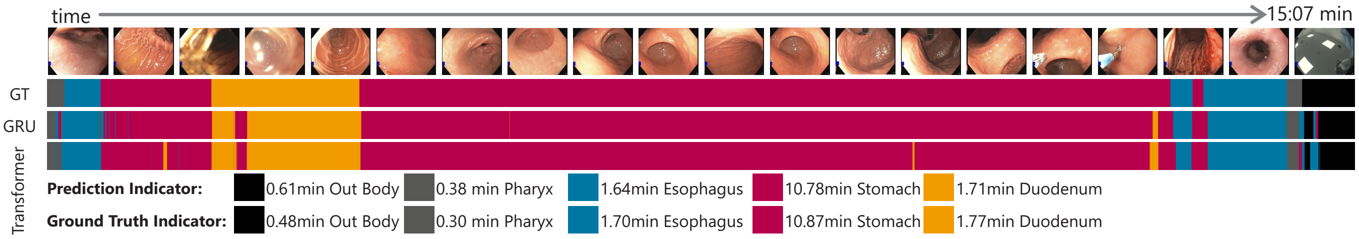

Fig. 2. Presenting qualitative outcomes displaying predictions within a single video endoscopy session, derived from both the GRU and Transformer models. These predictions were generated using a 5-second temporal window (150 frames) and are accompanied by time estimation indicators (GT: GroundTruth, and GRU: Gated Recurrent Unit).

factor for subsequent restarts. The Cross-Entropy loss function, augmented with class weights, was employed as the loss metric. The training was conducted for 30 epochs or stopped if the validation loss ceased to improve for 5 consecutive epochs.

B. *Results*

Both models were evaluated using the testing set with the best model obtained in the validation set, representing different patient's complete EDG procedures. The results were analyzed and presented in different scenarios, mainly differing in the temporal window for sequences used during training. The evaluation of the proposed approaches is based on the weighted-F1, macro-F1, and accuracy scores calculated per frame, considering its contextual information for the organ task and by sequence for the identification of the subatomic region of the stomach, as shown in Table II.

1) Organ and stomach region classification performance: In the organ task, the transformer architecture demonstrated superior performance when considering a wide temporal window. However, there were misclassifications mainly observed in the pharynx, as shown in Figure 2 (see gray and yellow of GT row compared to transformer predictions). These regions presented challenges primarily due to limited sample availability and a particularly large amount of noise during the endoscope pass through the pharynx.

For the second task of stomach sub-region classification, performance correlated with temporal windows ranging between 1 and 3 seconds. This observation is attributed to the inspection process in which experts capture photos of specific landmarks within those zones. The use of large temporal windows could result in intersecting with other sub-regions and potentially lead to less accurate classification. The best result for this experimental setup was achieved using the sequence transformer architecture with 50 frames like temporal windows.

2) Organ quality indicator: On the test set, the video endoscopy procedure was labeled per frame using predictions from the GRU model, enabling the extraction of an organ exploration time indicator. Figure 2 illustrates the application of predictions to a procedure within the test set, demonstrating how the quantification of time per organ can be used as an inspection time indicator. The mean square error (MSE) between labels and prediction was 0.04, 0.08, and 0.10 for Esophagus, Stomach, and Duodenum respectively.

TABLE II
QUANTITATIVE PERFORMANCE WITH DIFFERENT TEMPORAL WINDOWS

	Organ task			Stomach task		
Metric	**Time**	**GRU**	**Transf.**	**Time**	**GRU**	**Transf.**
Accuracy	0.50	86.48	88.19	0.33	83.22	83.22
Weighted F1		86.91	88.20		83.12	82.87
Macro F1		78.20	78.89		82.54	82.35
Accuracy	1.00	85.89	88.30	0.50	83.22	83.22
Weighted F1		86.29	88.42		82.96	82.73
Macro F1		77.46	79.83		82.41	82.34
Accuracy	2.00	86.59	88.43	1.00	84.62	82.52
Weighted F1		87.10	88.60		84.50	82.13
Macro F1		80.86	80.72		83.83	81.52
Accuracy	3.00	88.45	89.12	1.33	83.22	81.82
Weighted F1		88.52	89.13		82.67	81.34
Macro F1		80.44	81.34		81.99	80.76
Accuracy	4.00	87.15	89.82	1.66	82.52	**85.31**
Weighted F1		87.22	89.91		82.36	**85.00**
Macro F1		77.19	82.89		81.71	**85.31**
Accuracy	5.00	89.24	**91.24**	2.00	81.12	83.92
Weighted F1		89.33	**91.42**		80.62	83.78
Macro F1		82.51	**87.25**		79.88	83.35
Accuracy	10.00	89.17	89.06	3.00	83.92	80.42
Weighted F1		88.35	88.33		83.68	80.26
Macro F1		71.56	70.83		83.22	79.84

Abbreviations - Time: in seconds, Transf: Transformer encoder

IV. CONCLUSIONS AND DISCUSSION

Quantifying examination time and auditing the EGD procedure are vital steps toward enhancing the early detection of upper GI neoplasia. Automatically quantifying procedure times not only contributes to improving patient outcomes but also serves as a quality indicator of endoscopy centers worldwide. This paper presents an automated methodology for encoding spatiotemporal information to efficiently audit the upper endoscopy workflow. The comparison of various temporal methodologies employed within consistent conditions in this study demonstrates the potential benefits of incorporating temporal information into the classification tasks of anatomical and sub-anatomical regions during EGD. Furthermore, it highlights the potential to derive procedural quality indicators exclusively from these classifications, marking a significant advancement in the field. Looking ahead, future research will aim to confront challenges such as the substantial noise prevalent in EGD procedures, particularly in regions like the pharynx. The intention is to enhance the classification of specific regions and anatomical transitions, combine the organ and stomach classification tasks with the quality indicator in a cohesive pipeline, and release an open database of EGD videos, thereby contributing further to the field.

V. Acknowledgments

This work was supported by project with code 110192092354 and entitle “Program for the Early Detection of Premalignant Lesions and Gastric Cancer in urban, rural and dispersed areas in the Department of Nariño” of call No. 920 of 2022 of MinCiencias.

References

[1] H. Sung, J. 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] V. Pasechnikov, S. Chukov, *et al.*, “Gastric cancer: prevention, screening and early diagnosis,” *World journal of gastroenterology: WJG*, vol. 20, no. 38, p. 13842, 2014.

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

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

[5] P. W. Y. Chiu, N. Uedo, R. Singh, T. Gotoda, E. K. W. Ng, K. Yao, T. L. Ang, S. H. Ho, D. Kikuchi, F. Yao, *et al.*, “An asian consensus on standards of diagnostic upper endoscopy for neoplasia,” *Gut*, vol. 68, no. 2, pp. 186–197, 2019.

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

[7] G. H. Kim, S. J. 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, p. 747, 2015.

[8] A. M. Veitch, N. Uedo, K. Yao, and J. E. East, “Optimizing early upper gastrointestinal cancer detection at endoscopy,” *Nature reviews Gastroenterology & hepatology*, vol. 12, no. 11, pp. 660–667, 2015.

[9] W. Januszewicz and M. F. Kaminski, “Quality indicators in diagnostic upper gastrointestinal endoscopy,” *Therapeutic Advances in Gastroenterology*, vol. 13, p. 1756284820916693, 2020.

[10] J. K. Min, M. S. Kwak, and J. M. Cha, “Overview of deep learning in gastrointestinal endoscopy,” *Gut and liver*, vol. 13, no. 4, p. 388, 2019.

[11] S. Y. Kim and J. M. Park, “Quality indicators in esophagogastroduodenoscopy,” *Clin Endosc*, vol. 55, pp. 319–331, May 2022.

[12] H. Takiyama, T. Ozawa, *et al.*, “Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks,” *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[13] L. Wu, W. 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.

[14] Y.-Y. Chang, P.-C. Li, R.-F. Chang, C.-D. Yao, Y.-Y. Chen, W.-Y. Chang, and H.-H. Yen, “Deep learning-based endoscopic anatomy classification: an accelerated approach for data preparation and model validation,” *Surgical Endoscopy*, pp. 1–11, 2021.

[15] Y.-D. Li, S.-W. 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.

[16] D. Bravo, J. Ruano, M. Jaramillo, D. Gallego, M. Gómez, F. A. González, and E. Romero, “Automatic classification of esophagogastroduodenoscopy sub-anatomical regions,” in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2023.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.