This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Automated Auditing of Upper Endoscopy Procedure Times: A Temporal Multiclass Analysis

Diego Bravo[1,2][0000−0003−1957−1615], Josué Ruano[1,2][0000−0003−4801−8229], Martín Gómez[1,3], Fabio A. Gónzalez[1,4], and Eduardo Romero[1,2]

[1] Universidad Nacional de Colombia, Bogotá, Colombia
[2] Computer Imaging and Medical Applications Laboratory (CIM@LAB)
[3] Hospital Universitario Nacional de Colombia, Bogotá, Colombia
[4] Machine Learning, Perception and Discovery Lab (MindLab)
edromero@unal.edu.co

Abstract. Upper endoscopy is the preferred method for detecting early-stage gastrointestinal diseases and plays a crucial role in managing gastric cancer. Quality assessment has been a recurring concern in clinical research, particularly regarding the time specialists spend examining different anatomical sites. While current guidelines emphasize thorough inspection and documentation to minimize blind spots, adherence remains low due to the lack of second readers. State-of-the-art automatic approaches audit single-frame or fixed temporal windows, with limited performance in real applications. This paper introduces the Multi-Scale Sequence Informative (MSSI) module, a Transformer-based attention mechanism that audits video sequences across multiple temporal scales. The proposed approach estimates the time spent exploring different organs and regions of the stomach. The method processes 15 to 196 tokens (1 to 13 seconds) by a sliding window, building up a mosaic of sampled frames. Each frame is encoded with a pre-trained endoscopy embedding which feeds a Vision Transformer to capture short-, mid-, and long-range dependencies. The approach is evaluated with 233 endoscopic procedures (∼1.6 million frames), demonstrating a close alignment between estimated procedural times and expert-validated standards. It achieved 92.03% macro precision in organ classification and 89.34% in distinguishing 23 specific views of different stomach sites, a total of 27 classes to audit, showing real potential to be applicable in real clinical scenarios. Our code is available at https://github.com/Cimalab-unal/EndoAudit.git.

Keywords: Endoscopy Quality · Exam Time · Temporal Analysis.

1 Introduction

Upper endoscopy or esophagogastroduodenoscopy (EGD) is the diagnostic and screening tool to study upper gastrointestinal (UGI) disorders, particularly in high-risk populations [1]. However, despite the wide clinical application, EGD is

limited by cognitive and technical skill requirements [2]. Studies have reported missed detection rates exceeding 20% in Asia and ranging from 7.2% to 14.0% in Western countries [3,4,5,6], all of them concluding in the necessity of standardizing EGD procedures. As a result, guidelines and experts have agreed about quality or auditory standards [7]. In particular, the American Society for Gastrointestinal Endoscopy (ASGE) and the American College of Gastroenterology (ACG) have proposed safety and quality EGD indicators [8]. In 2015, the European Society of Gastrointestinal Endoscopy (ESGE) introduced the first evidence-based EGD indicators [9]. Overall, current guidelines recommend a systematic examination of all gastric regions and a minimum examination time of 7 minutes to optimize lesion detection [10,11]. Non-blind spot EGD has gained attention, prompting standardized protocols for thorough stomach mapping [12]. However, compliance with these standards is suboptimal by a lack of monitoring tools [13]. Stomach photodocumentation is the most studied quality indicator in endoscopy, yet procedure duration is often overlooked. Only 18.2% of endoscopy reports include EGD duration, whereas 51% report photodocumentation [13], despite longer inspection times being associated with higher lesion detection rates [9,10]. Hence integrating temporal data improves accuracy and standardization of EGD quality assessment.

In clinical practice, the need for an appropriate audit positions Artificial Intelligence (AI) as a promising solution to partially or fully automating EGD quality assessment. Pioneering studies of quality in EGD procedures analyzed single-frames, most of them using private datasets to train Convolutional Neural Networks (CNNs) and classify anatomical regions (larynx, esophagus, stomach, and duodenum) [14] or assess photodocumentation guided by specific protocols [15,16]. Yuan et al. [17] trained a MobileNetV3-large with 144,277 images (27 categories) and validated it on 16,031 images, achieving 91.85% of sensitivity. Other methods combined CNN-based single-frame classifications with sliding windows to apply hard-voting-based algorithms and improve stability [18]. Kang et al. [19] used the InceptionResNetV2 model for video analysis, achieving an F1-score of 61.37%. However, single-frame predictions, although incorporating temporal filters in real-time applications, are unstable, likely because they ignore temporal consistency and hardly capture long-range dependencies. To overcome this limitation, Li et al. [20] combined CNN-based single-frame classifications with a Long Short-Term Memory (LSTM) network of 5 frames. Lately, systems like WISENSE [21] and ENDOANGEL [22] used CNNs and reinforcement learning to detect blind spots while tracking the procedure duration. However, multi-frame methods have three limitations: (1) LSTM captures minimal temporal information, (2) the memory window is too short to track organ details, and (3) training relies on private datasets, limiting reproducibility.

Unlike previous works, the main contribution of this study is the automatic auditing of the time spent by gastroenterologist in the pharynx, esophagus, duodenum, stomach, and the 22 stomach sites, as outlined in the Systematic Screening Protocol for the Stomach. The model produces interpretable outputs in the form of time-based quality indicators from sequence classification using a long-

range CNN-Transformer framework, providing practical evidence for endoscopy quality assessment and procedural auditing.

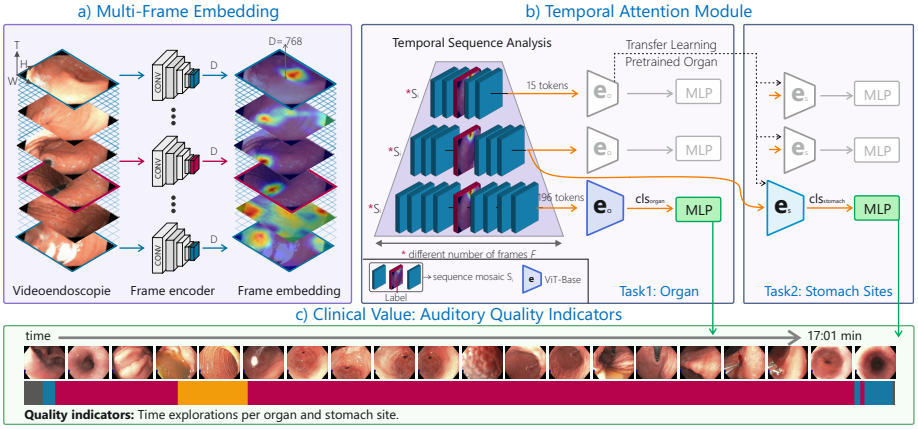

Fig. 1. MSSI applies a Multi-Frame Embedding to each frame and generates rich feature representations for the entire EGD video. The Temporal Module, built upon ViT, captures embeddings from temporal sequences (x_i), with the keyframe positioned at the center. Each sequence embedding (S_i) and its corresponding class token (cls_i) are passed through a multi-layer perceptron (MLP) to classify the organ and stomach site.

2 Method

Multi-Frame Embedding. A Multi-Frame Framework extracts frame-level embeddings to feed token representations of a Temporal Attention Module. Each frame is sequentially processed by a ConvNeXt-Tiny CNN, pretrained with 270 endoscopy cases, which encodes the frame as a D-dimensional feature vector ($D = 768$), reducing redundancy and preserving salient visual details (see Figure 1-a). Formally, let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ denotes the collection of video segments, where N is the total number of segments. Each segment $x_i \in \mathbb{R}^{F\times H\times W\times 3}$ consists of F RGB frames with $H \times W$ spatial resolution. The extracted frame embeddings make an ordered temporal sequence, structured as: $S_i = f_{\text{ConvNeXt}}(x_i) \in \mathbb{R}^{F\times D}$ where F represents the number of frames (tokens) in the temporal window. These S_i sequences (mosaics) serve as input to a Transformer module that learns short- and long-term dependencies, building the analysis upon temporal attention represented by multiple time scales.

Transformer Backbone. The Vision Transformer (ViT) [23] is the core of the video analysis. It is initialized with ImageNet-pretrained weights and fine-tuned with endoscopy videos. The transformer is adapted at modifying the usual

token representation, i.e., rather than applying ViT patch tokenization, a sliding temporal window constructs a mosaic S_i of sampled frames. Each token is derived from CNN-extracted multi-frame embeddings, facilitating the ViT connects high-level feature representations instead of raw pixel patches. This choice simplifies temporal modeling and improves recognition of endoscopic patterns.

Temporal Attention Module. The ViT processes the sequences S_i , applying self-attention to model spatio-temporal dependencies as follows:

$$S_i^{'} = f_{\mathrm{ViT}}(f_{\mathrm{ConvNeXt}}(x_i)) \in \mathbb{R}^{F \times D} \tag{1}$$

where F represents the number of frames (15-196 tokens) in the temporal window (see Figure 1-b), D is the embedding dimension. ViT employs Multi-Head Self-Attention (MSA) to capture short- and long-range dependencies, computing attention weights as [24]:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, V are projections of $S_i^{'}$. MSA enables parallel attention across multiple heads, refining temporal representations for improved sequence modeling and prediction.

2.1 Report Quality Indicators

Automated assessment may ensure endoscopists' compliance with clinical quality standards. This study extracts key EGD quality indicators using two models: UGI organ detection and stomach site identification, both trained with the Multi-Frame + ViT methodology (see Figure 1-a,b). Each model is optimized for its specific classification task. By integrating patient temporal metrics with classification results, this approach provides a standardized evaluation of the procedural performance, measuring organ-specific and stomach sites exploration times, achieving thereby actual EGD monitoring and auditing (see Figure 1-c).

3 Experiments and Results

3.1 Experimental Setup

Datasets. The framework was evaluated with the publicly available GastroHUN UGI video endoscopy dataset [25], focusing on 237 MP4 videos from 233 patients. All videos were normalized to 15 FPS using FFmpeg. To provide procedural time annotations, three expert gastroenterologists manually labeled Pharynx, Esophagus, Stomach, and Duodenum, with annotations publicly available under a CC-BY open license. Assessment of the time spent in 23 stomach sites was done using the labels provided by the database. This research follows the original split (165 training, 33 validation, 35 test patients). As this study uses open-access human subject data, ethical approval was not required.

Implementation Details. The transformer encoder is initialized with a ViT-Base backbone, modified to process variable-length token sequences (15 to 196 frames). Each token is represented by a 768-dimensional feature vector, extracted from a pretrained ConvNeXt-Tiny CNN. The final MLP classification head is composed of 4 output neurons for organ classification and 23 for stomach site detection. The training follows a three-phase strategy. During warm-up, only the classification layers are trained with 5 epochs at a constant learning rate of 0.001 to stabilize predictions. In the unfreezing phase, the last four layers are trained with a scheduled learning rate adjustment, enabling deeper refinement. In fine-tuning, all layers are unfrozen and trained during 15 epochs, with early stopping if no validation F1-macro score improvement is observed after five consecutive epochs. The model is implemented in PyTorch Lightning, using pretrained ImageNet weights [26] for the ViT and learned stomach weights [27] for the initial tokens embeddings, trained with Adam optimization and a weighted cross-entropy loss function to address class imbalance. The learning rate is adjusted every 5 epochs with $\gamma = 0.1$. Training is conducted on dual NVIDIA RTX 4500 GPUs with a batch size of 128. The Temporal Transformer encoder, comprising $\sim$85.5 million parameters, is designed for offline processing, analyzing temporal sequences to model procedural transitions and duration variations.

Ablation Study. An ablation study was conducted to assess the impact of key components. For organ classification, the role of temporal modeling was examined by removing it and evaluate its contribution in different temporal windows (Table 1). For stomach sites classification, the study analyzed the effect of pretrained weights in the Transformer encoder when testing long-term sequences (Table 2).

3.2 Model Validation

Organ. Table 1 presents the overall performance, divided into single-frame and multi-frame analyses. Single-frame spatial embeddings were tested by comparing different feature extraction methods, including ViT's patch-based linear projection (16×16×3), ConvNeXt-Tiny pretrained on ImageNet [26], and EGD data [27]. The best performance was achieved using pretrained stomach weights, where a MLP classifier reached a macro precision of 64.55%. The multi-frame analysis explored temporal modeling, comparing two architectures: one model used a single attention layer without pretrained weights, while the ViT-Base model was trained with and without ImageNet pretrained weights. In both cases, the initial token embeddings were derived from the best-performing single-frame embedding method. The best multi-frame performance was achieved with 135 tokens (9.0 seconds) using a ViT-Base pretrained, reaching a macro precision of 92.03%. Figure 2 illustrates the qualitative and quantitative performance for a particular patient, demonstrating strong temporal correlation and consistent predictions closely aligned with expert annotations.

Table 1. Performance comparison of single-frame and multi-frame approaches to classify the organ using deep learning architectures. The models were trained with 165 cases (∼1,182M samples), validated with 33 cases (∼234K samples), and tested with 35 cases (∼242K samples). Metrics include macro precision (Prec.), recall (Rec.), F1-score, and Matthews correlation coefficient (MCC). **Bold**: The best test performance.

Embedding Variations Across Spatial Features												
	Linear Projection				ConvNeXt (ImageNet)				ConvNeXt (Endoscopy)			
time	Prec.	Rec.	F1	MCC	Prec.	Rec.	F1	MCC	Prec.	Rec.	F1	MCC
1 frame	49.74	72.21	54.48	48.86	62.78	85.15	70.37	68.13	**64.55**	**87.06**	**71.98**	**70.38**

Initial Token Embeddings (Endoscopy [27]) for the Temporal Module												
	Attention (Random W.)				ViT (Random Weights)				ViT (ImageNet)			
time	Prec.	Rec.	F1	MCC	Prec.	Rec.	F1	MCC	Prec.	Rec.	F1	MCC
1.0 sec	74.57	85.85	79.02	76.63	70.60	86.10	76.85	74.64	82.24	88.10	84.96	83.08
3.0 sec	82.90	88.90	85.54	84.02	78.54	89.47	83.19	80.80	89.74	89.08	89.14	87.85
5.0 sec	83.94	88.39	85.91	84.29	80.47	89.86	84.53	82.56	91.03	89.91	90.29	89.62
9.0 sec	85.80	86.17	85.16	84.26	77.56	87.52	80.98	78.96	**92.03**	89.89	**90.42**	**89.94**
13.1sec	86.56	84.75	84.54	83.95	77.31	**90.33**	82.72	80.11	89.87	88.53	88.64	88.19

Stomach sites. The stomach site classification models were trained with 159 cases (3,401 samples), validated with 32 cases (654 samples), and tested with 32 cases (674 samples). Table 2 presents a comparison of several multi-frame temporal models using a ViT-Base backbone pretrained on organ classification. The proposed model achieves the highest performance with a 9.0-second input sequence, reaching a macro precision of 89.34±0.30, recall of 88.50±0.34, and F1-score of 87.96±0.33. To evaluate statistical significance, we replicated the evaluation procedure from [27], applying their proposed bootstrapping approach. We computed 95% confidence intervals ($\alpha = 0.05$) for each metric using 100 resamples, confirming the statistical significance of the observed improvements. Using the same dataset and splits, our model surpasses the best results from [27] (Transformer: macro precision of 86.98±0.42, recall of 87.01±0.41, and F1-score of 86.30±0.42; GRU: macro precision of 86.74±0.38, recall of 86.09±0.39, and F1-score of 85.47±0.39), highlighting the effectiveness of temporal information in the classification of stomach regions.

Table 2. Performance comparison of multi-frame approaches for stomach site classification using a pretrained ViT-Base architecture from the organ classification task.

	ViT (Organ 3.0 sec.)				ViT (Organ 9.0 sec.)				ViT (Organ 13.1 sec.)			
time	Prec.	Recall	F1	MCC	Prec.	Recall	F1	MCC	Prec.	Recall	F1	MCC
1.0 sec	83.38	82.66	81.62	82.45	84.20	83.43	82.71	83.46	83.21	81.87	80.97	82.36
3.0 sec	83.87	83.64	82.38	83.22	85.08	84.01	83.02	83.94	86.14	85.21	84.56	85.26
5.0 sec	86.02	86.04	84.96	86.04	87.48	87.18	86.26	87.44	85.61	84.64	83.65	84.69
7.0 sec	87.66	87.30	86.45	87.38	84.90	84.91	83.39	84.71	88.37	87.82	87.03	87.79
9.0 sec	**89.34**	**88.50**	**87.96**	87.90	86.83	86.96	85.61	86.73	86.86	86.81	85.20	86.73
10 sec	87.81	86.83	86.43	87.21	87.82	87.29	86.68	87.27	88.51	88.22	87.53	**87.99**

3.3 Quality Indicators

Table 3 presents the mean and standard deviation of the exploration times of complete procedures, namely Pharynx, Esophagus, Stomach, and Duodenum. The average procedure duration is 9:22±4:17, complying with the recommended minimum of 7 minutes of a thorough examination [9,10]. Regarding the second quality indicator, stomach site inspection duration across 22 stomach sites [12]. Certain regions, such as the Lesser Curvature in the Antrum and the Greater Curvature in the Antrum, Lower Body, and Middle-upper Body (L1, G1, G2, and G3), require larger than 24 seconds for inspection, corresponding to biopsy extraction sites in the Sydney protocol. Interestingly, posterior walls in Lower Body and Middle-upper Body (P2 and P3), widely acknowledged as regions with high-risk of missing lesions, show an inspection time of barely 13 seconds, a major concern if one considers a higher time should be devoted to these areas after clinic protocols [5]. Overall, current stomach examination standards specify which sites to visit but do not mandate a minimum exploration time for each site. However, multiple studies have shown that longer inspection durations improve gastric lesion detection [10].

Table 3. Quality indicators of complete procedures. L: lesser curvature, A: anterior wall, G: greater curvature, P: posterior wall, and SSS: systematic screening protocol for the stomach.

Indicator 1: Organ-Specific Exploration Time (Protocol: [9])					
Patients	Procedure	Pharynx	Esophagus	Stomach	Duodenum
15	9:22±4:17	0:13±0:17	0:54±0:38	7:17±2:54	0:56±1:19

Indicator 2: Stomach Sites Duration (Protocol SSS: [12])					
Region	Site	Time	Region	Site	Time
Antrum Antegrade	A1	0:21±0:10	Lower Body Antegrade	A2	0:11±0:06
	L1	0:29±0:27		L2	0:11±0:06
	P1	0:19±0:13		P2	0:15±0:12
	G1	0:36±0:19		G2	0:34±0:36
Middle Body Antegrade	A3	0:08±0:06	Fundus Cardia Retroflex	A4	0:05±0:04
	L3	0:07±0:06		L4	0:06±0:04
	P3	0:11±0:08		P4	0:06±0:05
	G3	0:24±0:17		G4	0:09±0:07
Middle Body Retroflex	A5	0:05±0:05	Incisura Retroflex	A6	0:11±0:09
	L5	0:10±0:08		L6	0:11±0:11
	P5	0:05±0:03		P6	0:10±0:09

Figure 2 presents the quality report of a test patient (# 124), summarizing both organ-specific exploration time and stomach site inspection duration. The predicted organ sequence (using a pretrained ViT with a 9.0-second temporal window) is compared against the ground truth. Similarly, stomach site predictions (using a 9.0-second temporal window) are visualized. The total procedure duration (17:01 min) largely exceeds the recommended minimum of 7 minutes.

However, while all 22 stomach sites were inspected, some regions (A2, L3, P3, A4, P4, G4, and P5) were examined in less than 10 seconds (warning icons), supporting the need for automated quality auditing. This also highlights the importance of further discussion and validation of minimum exploration times, which should be standardized in clinical auditing protocols. Additionally, Figure 2 in frame 2 illustrates a challenging sample where the method makes a correct prediction, despite the frame being non-informative due to motion blur and low visibility in the anatomical transition between the pharynx and esophagus (frames 1–3). A longer temporal window provides both local and global context to stabilize predictions. This mechanism resembles how physicians use temporal continuity to interpret ambiguous or low-quality scenes. Regarding stomach site classification, the model accurately predicted all frames between 5 and 9 in accordance with the labels provided in the dataset. In regions such as the antrum, Grad-CAM visualizations as shown in Figure 2 suggest that the model consistently attends to distinctive anatomical landmarks, these landmarks, such as the pylorus, gastric folds, or the incisura, appear to generate high-gradient responses, likely due to their structural prominence (e.g., orifices, folds, or protrusions). This focus is supported by the alignment of predictions in frames 5 (L1), 6 (L2), 7 (G3), 8 (A6), and 9 (G4). These frames serve as crucial reference points, enabling the model to differentiate between neighboring stomach sites.

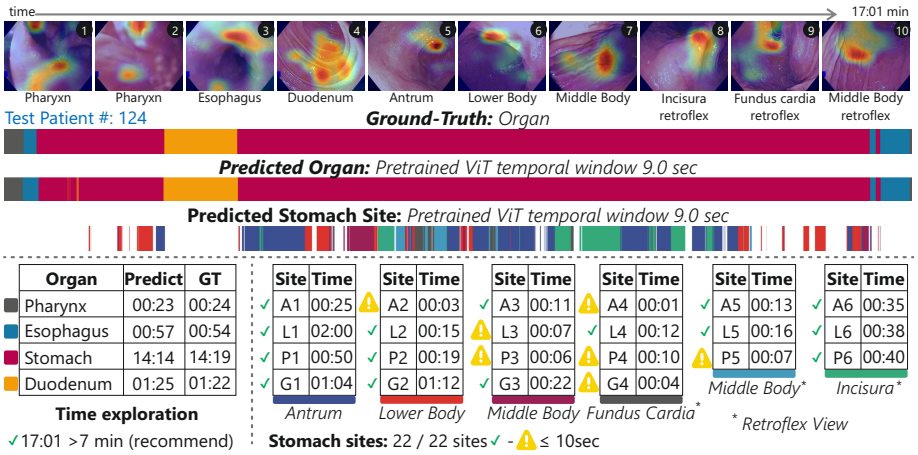

Fig. 2. Quality Indicator Report. The top panel displays organ and stomach site classification timelines. The bottom panel shows inspection times.

4 Conclusions

This study evaluates two classification tasks to construct a detailed and objective audit of procedural quality, incorporating time as a key assessment factor.

The method is tested on a public dataset at both metric and patient levels. Transformer-based multi-frame detection enables quantitative monitoring of organ exploration time and stomach sites inspection for endoscopic quality assessment. However, since this study was conducted as a controlled offline experiment using a limited dataset from single-brand Olympus endoscopes, its generalizability may be affected. Adapting the algorithm to other endoscope systems will require transfer learning or fine-tuning. Future work will focus on clinical validation, refining spatiotemporal representations, exploring interpretability and explainability, and enabling real-time evaluation.

Acknowledgments. This work was partially supported by project 110192092354 "Program for the Early Detection of Premalignant Lesions and Gastric Cancer in urban, rural and dispersed areas in the Department of Nariño" of MinCiencias.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Hamashima, C., Ogoshi, K., Narisawa, R., Kishi, T., Kato, T., Fujita, K., Sano, M., Tsukioka, S.: Impact of endoscopic screening on mortality reduction from gastric cancer. World journal of gastroenterology: WJG **21**(8), 2460 (2015)
2. Menon, S., Trudgill, N.: How commonly is upper gastrointestinal cancer missed at endoscopy? a meta-analysis. Endoscopy international open **2**(02), E46–E50 (2014)
3. Ren, W., Yu, J., Zhang, Z.M., Song, Y.K., Li, Y.H., Wang, L.: Missed diagnosis of early gastric cancer or high-grade intraepithelial neoplasia. World Journal of Gastroenterology: WJG **19**(13), 2092 (2013)
4. Chadwick, G., Groene, O., Hoare, J., Hardwick, R.H., Riley, S., Crosby, T.D., Hanna, G.B., Cromwell, D.A.: A population-based, retrospective, cohort study of esophageal cancer missed at endoscopy. Endoscopy **46**(07), 553–560 (2014)
5. Pimenta-Melo, A.R., Monteiro-Soares, M., Libânio, D., Dinis-Ribeiro, M.: Missing rate for gastric cancer during upper gastrointestinal endoscopy: a systematic review and meta-analysis. European journal of gastroenterology & hepatology **28**(9), 1041–1049 (2016)
6. Beck, M., Bringeland, E.A., Qvigstad, G., Fossmark, R.: Gastric cancers missed at upper endoscopy in central norway 2007 to 2016—a population-based study. Cancers **13**(22), 5628 (2021)
7. Chiu, P.W.Y., Uedo, N., Singh, R., Gotoda, T., Ng, E.K.W., Yao, K., Ang, T.L., Ho, S.H., Kikuchi, D., Yao, F., et al.: An asian consensus on standards of diagnostic upper endoscopy for neoplasia. Gut **68**(2), 186–197 (2019)
8. Rizk, M.K., Sawhney, M.S., Cohen, J., Pike, I.M., Adler, D.G., Dominitz, J.A., Lieb, J.G., Lieberman, D.A., Park, W.G., Shaheen, N.J., et al.: Quality indicators common to all gi endoscopic procedures. Official journal of the American College of Gastroenterology| ACG **110**(1), 48–59 (2015)
9. Bisschops, R., Areia, M., Coron, E., Dobru, D., Kaskas, B., Kuvaev, R., Pech, O., Ragunath, K., Weusten, B., Familiari, P., et al.: Performance measures for upper gastrointestinal endoscopy: a european society of gastrointestinal endoscopy (esge) quality improvement initiative. Endoscopy **48**(09), 843–864 (2016)

10. Teh, J.L., Tan, J.R., Lau, L.J.F., Saxena, N., Salim, A., Tay, A., Shabbir, A., Chung, S., Hartman, M., So, J.B.Y.: Longer examination time improves detection of gastric cancer during diagnostic upper gastrointestinal endoscopy. Clinical Gastroenterology and Hepatology **13**(3), 480–487 (2015)
11. Veitch, A.M., Uedo, N., Yao, K., East, J.E.: Optimizing early upper gastrointestinal cancer detection at endoscopy. Nature reviews Gastroenterology & hepatology **12**(11), 660–667 (2015)
12. Yao, K.: The endoscopic diagnosis of early gastric cancer. Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology **26**(1), 11 (2013)
13. Zagari, R.M., Frazzoni, L., Fuccio, L., Bertani, H., Crinò, S.F., Magarotto, A., Dajti, E., Tringali, A., Da Massa Carrara, P., Cengia, G., et al.: Adherence to european society of gastrointestinal endoscopy quality performance measures for upper and lower gastrointestinal endoscopy: a nationwide survey from the italian society of digestive endoscopy. Frontiers in Medicine **9**, 868449 (2022)
14. Takiyama, H., Ozawa, T., Ishihara, S., Fujishiro, M., Shichijo, S., Nomura, S., Miura, M., Tada, T.: Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. Scientific reports **8**(1), 7497 (2018)
15. Xu, Z., Tao, Y., Wenfang, Z., Ne, L., Zhengxing, H., Jiquan, L., Weiling, H., Huilong, D., Jianmin, S.: Upper gastrointestinal anatomy detection with multitask convolutional neural networks. Healthcare technology letters **6**(6), 176–180 (2019)
16. He, Q., Bano, S., Ahmad, O.F., Yang, B., Chen, X., Valdastri, P., Lovat, L.B., Stoyanov, D., Zuo, S.: Deep learning-based anatomical site classification for upper gastrointestinal endoscopy. International journal of computer assisted radiology and surgery **15**, 1085–1094 (2020)
17. Yuan, P., Ma, Z.H., Yan, Y., Li, S.J., Wang, J., Wu, Q.: Artificial intelligence-based classification of anatomical sites in esophagogastroduodenoscopy images. International Journal of General Medicine pp. 6127–6138 (2024)
18. Ding, A., Li, Y., Chen, Q., Cao, Y., Liu, B., Chen, S., Liu, X.: Gastric location classification during esophagogastroduodenoscopy using deep neural networks. In: 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE). pp. 1–8. IEEE (2021)
19. Kang, S.M., Lee, G.P., Kim, Y.J., Kim, K.O., Kim, K.G.: Deep learning models for anatomical location classification in esophagogastroduodenoscopy images and videos: A quantitative evaluation with clinical data. Diagnostics **14**(21), 2360 (2024)
20. Li, Y.D., Zhu, S.W., Yu, J.P., Ruan, R.W., Cui, Z., Li, Y.T., Lv, M.C., Wang, H.G., Chen, M., Jin, C.H., et al.: Intelligent detection endoscopic assistant: An artificial intelligence-based system for monitoring blind spots during esophagogastroduodenoscopy in real-time. Digestive and Liver Disease **53**(2), 216–223 (2021)
21. Wu, L., Zhang, J., Zhou, W., An, P., Shen, L., Liu, J., Jiang, X., Huang, X., Mu, G., Wan, X., et al.: Randomised controlled trial of wisense, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut **68**(12), 2161–2169 (2019)
22. Wu, L., He, X., Liu, M., Xie, H., An, P., Zhang, J., Zhang, H., Ai, Y., Tong, Q., Guo, M., et al.: Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial. Endoscopy **53**(12), 1199–1207 (2021)

23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
25. Bravo, D., Gómez, M., Frías, J., Martínez, C., Polanía, F.V., González, F.A., Romero, E., Naranjo, J.: GastroHun: an endoscopy dataset of complete systematic screening protocol for the stomach. Figshare. `https://doi.org/10.6084/m9.figshare.27308133` (2025)
26. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
27. Bravo, D., Frias, J., Vera, F., Trejos, J., Martínez, C., Gómez, M., González, F., Romero, E.: Gastrohun an endoscopy dataset of complete systematic screening protocol for the stomach. Scientific Data **12**(1), 102 (2025)