Surgical Endoscopy
https://doi.org/10.1007/s00464-020-08236-6

Development of artificial intelligence system for quality control of photo documentation in esophagogastroduodenoscopy

Seong Ji Choi[1] · Mohammad Azam Khan[2] · Hyuk Soon Choi[3] · Jaegul Choo[2] · Jae Min Lee[3] · Soonwook Kwon[4] · Bora Keum[3] · Hoon Jai Chun[3]

Received: 9 July 2020 / Accepted: 8 December 2020

Abstract

Background Esophagogastroduodenoscopy (EGD) is generally a safe procedure, but adverse events often occur. This highlights the necessity of the quality control of EGD. Complete visualization and photo documentation of upper gastrointestinal (UGI) tracts are important measures in quality control of EGD. To evaluate these measures in large scale, we developed an AI-driven quality control system for EGD through convolutional neural networks (CNNs) using archived endoscopic images.

Methods We retrospectively collected and labeled images from 250 EGD procedures, a total of 2599 images from eight locations of the UGI tract, using the European Society of Gastrointestinal Endoscopy (ESGE) photo documentation methods. The label confirmed by five experts was considered the gold standard. We developed a CNN model for multi-class classification of EGD images to one of the eight locations and binary classification of each EGD procedure based on its completeness.

Results Our CNN model successfully classified the EGD images into one of the eight regions of UGI tracts with 97.58% accuracy, 97.42% sensitivity, 99.66% specificity, 97.50% positive predictive value (PPV), and 99.66% negative predictive value (NPV). Our model classified the completeness of EGD with 89.20% accuracy, 89.20% sensitivity, 100.00% specificity, 100.00% PPV, and 64.94% NPV. We analyzed the credibility of our model using a probability heatmap.

Conclusions We constructed a CNN model that could be used in the quality control of photo documentation in EGD. Our model needs further validation with a large dataset, and we expect our model to help both endoscopists and patients by improving the quality of EGD procedures.

Keywords Endoscopy · Esophagogastroduodenoscopy · Artificial intelligence · Deep learning · Quality control

Upper gastrointestinal endoscopy, also known as esophagogastroduodenoscopy (EGD), plays an important role in the diagnosis and treatment of upper gastrointestinal diseases. Because numerous trials demonstrated the importance of screening for gastric cancer in improving its prognosis, South Korea and Japan, two of the countries with the highest rates of gastric cancer, included endoscopic screening in a national cancer-screening program [1–3]. The demand for EGD, which is a useful gastric cancer-screening method, has increased rapidly every year [4]. Even though EGD is generally a safe procedure, the number of adverse events has increased with the number of procedures, which highlights the necessity of EGD quality control [5]. In addition to the adverse events resulting from EGD, an increasing number of legal actions justifies the need for standard quality control to protect both patients and endoscopists.

The quality of a cancer-screening program should be evaluated by its power to identify lesions [6]. However,

Seong Ji Choi and Mohammad Azam Khan equally contributed to this work as first co-authors.

✉ Hyuk Soon Choi
mdkorea@gmail.com

✉ Jaegul Choo
jchoo@kaist.ac.kr

1 Department of Internal Medicine, Hanyang University College of Medicine, Seoul, Republic of Korea

2 Graduate School of Artificial Intelligence, KAIST, Daejeon, Republic of Korea

3 Division of Gastroenterology and Hepatology, Department of Internal Medicine, Korea University College of Medicine, Seoul, Republic of Korea

4 Department of Anatomy, School of Medicine, Catholic University of Daegu, Daegu, Republic of Korea

Published online: 07 January 2021

this outcome measure cannot be applied to individual EGD procedures because of the low incidence of gastric cancer. Instead, the quality of EGD can be evaluated using other indicators; the completeness of visualization and photo documentation of the upper gastrointestinal (UGI) tracts, from the upper esophageal sphincter to the second portion of the duodenum, is an important measure in rating the performance of each EGD. The first and most widely accepted image documentation guideline was introduced by the European Society of Gastrointestinal Endoscopy (EGSE), recommending the acquisition of images of eight particular UGI landmarks [7]. Other guidelines and recommendations have been proposed since then, but they have not shown much difference from the EGSE method [8, 9]. However, to evaluate the quality of all EGD procedures would require considerable funds, manpower, and time.

Meanwhile, the recent advancement of convolutional neural networks (CNNs) has made it possible to apply artificial intelligence (AI) to almost every industry; its applications were successful in many areas of the medical field [10], especially in image recognition and classification. In the field of gastrointestinal endoscopy, AI has been actively applied and shown promising results in the identification and classification of colon adenoma using colonoscopic images [11–13]. Furthermore, several studies showed the success of AI applied to EGD images in the diagnosis of Barrett's esophagus, esophageal cancer, stomach cancer, and helicobacter pylori infection [14–17]. Based on these accomplishments of AI applications in endoscopic images, we developed an AI-driven quality control system for EGD using CNNs with endoscopic images.

Materials and methods

Study design and methods

A total of 13 endoscopists performed EGD at Korea University Anam Hospital, Seoul, Korea in January 2019. Images of EGD performed for screening are included in this study. Images were taken using standard upper endoscopy (GIF-H260, GIF-Q260, GIF-H290, GIF-HQ290; Olympus Medical Systems, Co. Ltd., Tokyo, Japan) and a standard endoscopic system (EVIS LUCERA ELITE CV-290/CLV-290SL; Olympus Medical Systems, Co. Ltd., Tokyo, Japan).

For the analysis, we collected 2599 gastroendoscopic images of 250 patients from the hospital Picture Archiving and Communication System. Prior to the image collection, we reviewed the endoscopic reports and excluded cases with mucosal abnormalities. The images were labeled based on the eight UGI landmarks according to the EGSE recommendation (Fig. 1) [7]. Five endoscopists, who had more than three years of experience in endoscopy and were blinded to the study, labeled each image in a separate room, and differences were resolved by majority decision. We also extended our study to validate inspection completeness of EGD procedures at the patient-level. Patient-level image statistics are shown in Table 1. To reduce selection bias, we collected

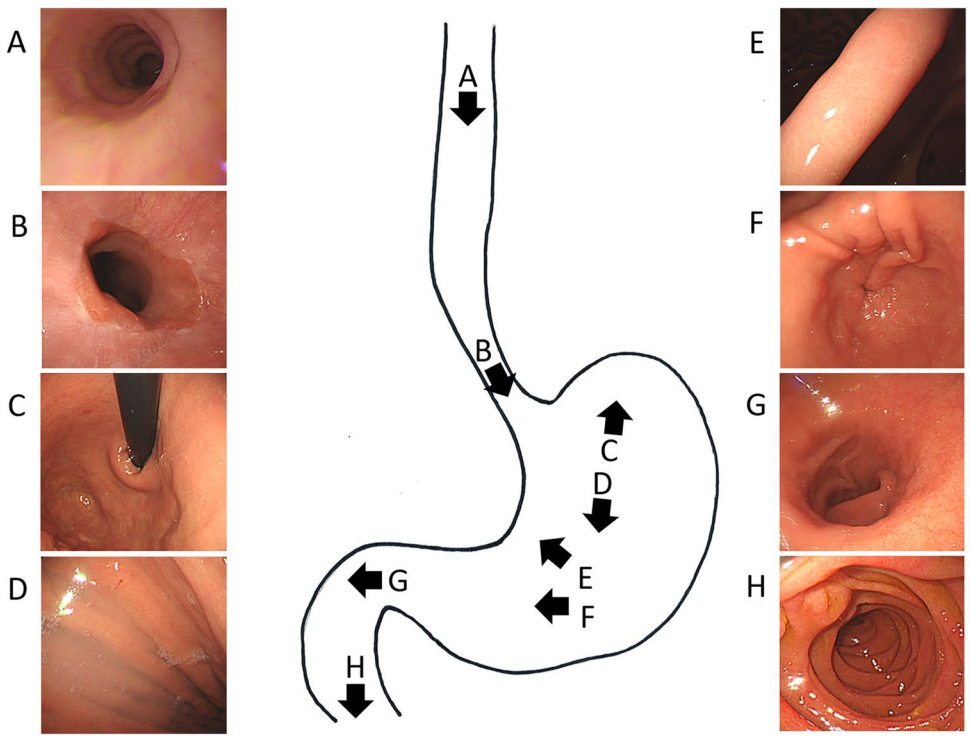

Fig. 1 Upper gastrointestinal (UGI) landmarks (**A**–**H**) according to the European Society of Gastrointestinal Endoscopy EGSE recommendation [7]. **A** proximal esophagus; **B** GEJ or Z-line; **C** stomach cardia and fundus on retroflexed view; **D** stomach body; **E** stomach angle; **F** stomach antrum; **G** duodenal bulb; **H** the second part of duodenum

Table 1 Dataset composition (image-level)

Endoscopic landmarks	Number of images
A	345
B	348
C	379
D	406
E	273
F	302
G	250
H	296
Total	2599

EGD images of 200 consecutive patients with complete visualization and 50 consecutive patients with incomplete visualization. The images were then labeled, fed into the CNN model, and classified into one of the eight locations. We utilized this classification result to check whether an inspection contains all the eight positional images.

An Intel® Xeon® CPU E5-2650 v4 @ 2.20 GHz machine equipped with 128 GB RAM was used to conduct all the experiments in this study. A total of eight NVIDIA TITAN Xp GPUs have been configured with the machine. However, we utilized only one GPU for the entire experiment. The proposed model was developed with PyTorch 1.1.0 using Python 3.6.8 in an Ubuntu 16.04.5 LTS environment [18].

Training and testing image sets preparation

We divided the entire dataset into ten cross-validation folds so that all images of one patient were placed together in either a training or validation/test set. Therefore, the sets were constructed in a manner that images from the same patient did not overlap among the training and validation/test sets. The statistics for different folds are shown in Table 2.

Preprocessing

Images captured during the EGD procedure include a black region that is left after removing the patient information. For deep learning models, such noise may hinder the training procedure and can impede the development of a robust model. Thus, we removed the unnecessary black portion from each image of the dataset by using automatic cropping, to minimize human intervention and maximize the utility of AI. The black portion of the image was removed effectively based on pixel statistics.

Obtaining a region of interest (ROI) from an image can be achieved using various techniques. We utilize the fact that most of the pixel values in the black region are close to 0 (on the scale from 0 to 255). Our cropping strategy is to calculate the mean pixel value of each row of the image

Table 2 Statistics of different folds

Fold	Number of patients/cases		Number of images	
	Complete inspection	Incomplete	Complete inspection	Incomplete
1	20	5	226	56
2	20	5	233	58
3	20	5	212	37
4	20	5	213	44
5	20	5	233	35
6	20	5	207	46
7	20	5	216	39
8	20	5	201	37
9	20	5	203	54
10	20	5	202	47
Total	200	50	2146	453

to differentiate between almost black rows and those rows containing useful RGB values. At the edge of our target ROI, the mean pixel value is significantly different from the mean pixel value of the black region. This phenomenon holds for the column-wise mean pixel value as well. Based on these values, we derive a rectangular bounding box with a left top point ($x1$, $y1$) and bottom right point ($x2$, $y2$) for cropping the informative region of the image. This process is illustrated in Fig. 2. Through this preprocessing step, we produced a more pertinent image than the original one and were able to train the CNN with more relevant information.

We applied the following data augmentation methods to the images: changing the aspect ratio and scaling after random cropping, using random rotations, and horizontal flips. Then, the input images were resized to 224×224 pixels in order to be compatible with the proposed network. Before being fed to the training network, an input image is normalized by channel-wise mean subtraction and divided by the standard deviation using ImageNet mean and standard deviation values, respectively [19].

CNN model construction

In a conventional image classification task, CNNs take advantage of multiple consecutive convolutional layers to extract important latent features from a given image. The outputs predict the likelihood across different classes through the fully connected layers. In this study, we leveraged a state-of-the-art CNN model called squeeze-and-excitation network (SENet) to validate our target EGD inspection task [20]. We utilized pre-trained models with weight initialization technique to improve the performance and convergence speed [21, 22]. The proposed CNN was fine-tuned by training for 200 epochs using Adam optimizer [23]. The initial learning rate was set to 0.001, and then an

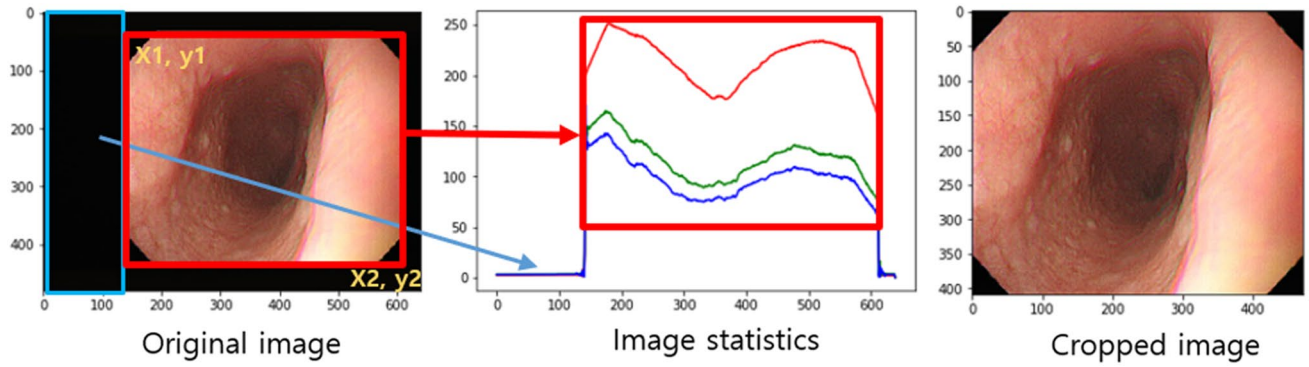

Fig. 2 Removing the black region of an image using image statistics. The color lines in the middle sub-figure indicate the column-wise mean value of RGB channels with color code red, green, and blue, respectively. Similarly, the row-wise mean value can be applied for removing the black region as well (Color figure online)

exponential-like learning rate scheduling was applied to train our model.

For the conventional multi-class classification task, a neural network is usually trained by using the cross-entropy loss ($\mathcal{L}_{\text{CE}}$) function. The loss function provides a gradient for the output layer, and the gradient is then backpropagated to the previous layers to create an updated direction for the weights. However, our study is slightly different from the classical multi-class context as all of the eight positional EGD images are closely related and share similar features. In this context, developing a robust classifier by using only a cross-entropy loss would be challenging, so we introduced an additional loss function called a positional loss ($\mathcal{L}_{\text{Pos}}$).

Due to anatomical structure, after observing a certain location during EGD examination, the next observing location is inevitably near the previous site. Although not always applicable, we applied this regression problem setting in our model to improve the outcome. We added a regression head in addition to the classification head just after CNN feature extractor. We penalized the model by calculating mean square error (MSE) between ground-truth value (1 to 8) and the predicted value using the regression head and added this loss value to the cross-entropy loss as well. Intuitively, we forced the network to learn both an image class and its positional value concurrently. Hence, we trained the network by optimizing the two loss functions, ($\mathcal{L}_{\text{CE}}$) and ($\mathcal{L}_{\text{Pos}}$), simultaneously. A brief overview of the training procedure is shown in Fig. 3.

Additionally, expanding on the results obtained from the multi-class classification task, we defined inspection completeness as a binary classification task: an inspection is evaluated as complete or incomplete. For an individual patient, the assessment is considered complete when the EGD test covers all the eight positional images. Otherwise, the EGD test is counted as incomplete.

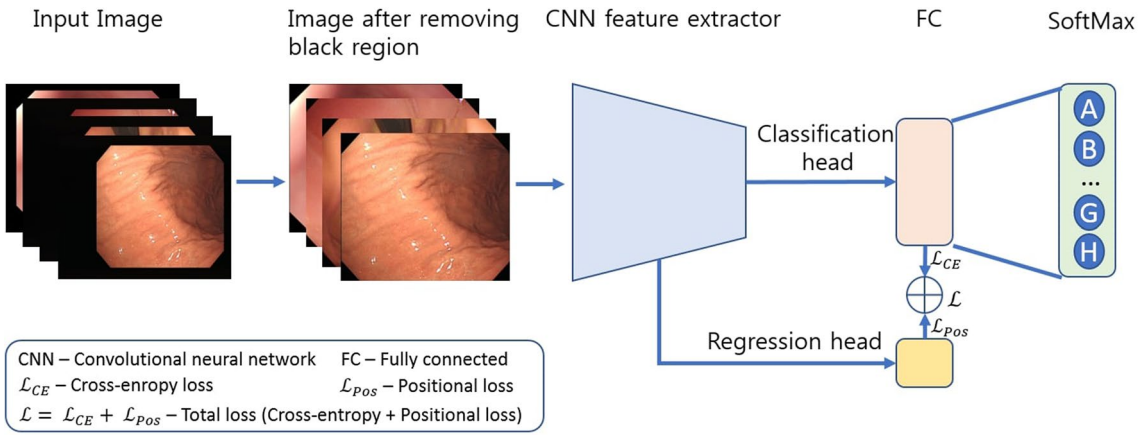

Fig. 3 The overall classification pipeline of the proposed model. An input image is first preprocessed by removing the black region from it and then normalized before being fed to the network to train. Afterwards, the model is trained by jointly optimizing the loss functions—a cross-entropy loss ($\mathcal{L}_{\text{CE}}$) and a positional loss ($\mathcal{L}_{\text{Pos}}$). Finally, the softmax layer predicts the likelihood through fully-connected (FC) layers across eight positions (**A**–**H**)

Finally, we visualized the class activation map using gradient-weighted class activation mapping (Grad-CAM) to investigate whether the model's decision was made based on the desired area of the image. This technique uses class-specific gradient knowledge that flows into the final convolutional layer of the CNN to generate a localization map of the important regions in the given image [24].

Outcome measures

The main outcome measure of the study was classification accuracy, but we also measured sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The following are the definitions we used to calculate the evaluation metrics, expressed as percentages.

- Accuracy is the proportion of true-positive, and true-negative images among the total images examined.
- Sensitivity is the proportion of true-positive images among true-positive and false-negative images.
- Specificity is the proportion of true-negative images among true-negative and false-positive images.
- PPV is the proportion of true-positive images among true-positive and false-positive images.
- NPV is the proportion of true-negative images among true-negative and false-negative images.

Results

Image collection

Table 1 shows the composition of the dataset according to the UGI landmarks. A total of 2599 images were collected and annotated as eight different UGI locations for the subsequent multi-class classification. Then, our dataset was re-organized to evaluate the completeness of the EGD procedures. In the 2599 images from the 250 EGD cases, 200 complete cases and 50 incomplete cases were included. Table 2 shows the statistics of the patients and the number of images for the different folds.

Multi-class classification

Our experiments focused on two tasks: multi-class classification and validation of inspection completeness. In multi-class classification, the aim was to categorize each image into one of the eight locations of the UGI tracts as shown in Fig. 1. Our proposed CNN model classified the corresponding location of an image with 97.58% accuracy, 97.42% sensitivity, 99.66% specificity, 97.50% PPV, and 99.66% NPV. These performance results were consistent across the ten different folds. Table 3 shows the model performance for the multi-class classification task, and Fig. 4 shows the confusion matrix.

Inspection completeness

In this task, we aimed to validate inspection completeness by expanding the results obtained from the multi-class classification task. The main goal of the task was to assess whether the proposed model can evaluate the completeness of an EGD procedure, an image set consisting of eight or more positional photos. Our proposed CNN model classified the completeness of the visualization of EGD procedure with 89.20% accuracy, 89.20% sensitivity, 100.00% specificity, 100.00% PPV, and 64.94% NPV. Figure 5 shows the confusion matrix describing the performance of the model in this task. In incorrectly classified sets, the model failed to identify a complete inspection only by one image location out of the eight locations for nearly all cases.

Visual explanation

We applied a probability heatmap to the classified images to understand the working of our model and to improve its performance. In the field of medical image analysis, it is crucial to establish that the prediction of the proposed model is based on proper detection of the relevant portion of the given image. Grad-CAM provides a visual mechanism to inspect which distinct region of the image influenced the decision of the model to assign the location information of the EGD image. Figure 6 shows several example images and their respective generated heatmaps using Grad-CAM.

Table 3 Model performance for location (multi-class) classification

Metrics	Mean value	Location							
		A	B	C	D	E	F	G	H
Accuracy	97.58	97.68	97.99	98.94	98.28	98.17	96.69	93.6	97.97
Sensitivity	97.42	97.68	97.99	98.94	98.28	98.17	96.69	93.6	97.97
Specificity	99.66	99.69	99.56	99.59	99.73	99.61	99.83	99.62	99.61
PPV	97.5	97.97	97.15	97.66	98.52	96.75	98.65	96.3	96.99
NPV	99.66	99.65	99.69	99.82	99.68	99.78	99.57	99.32	99.74

NPV negative predictive value, *PPV* positive predictive value

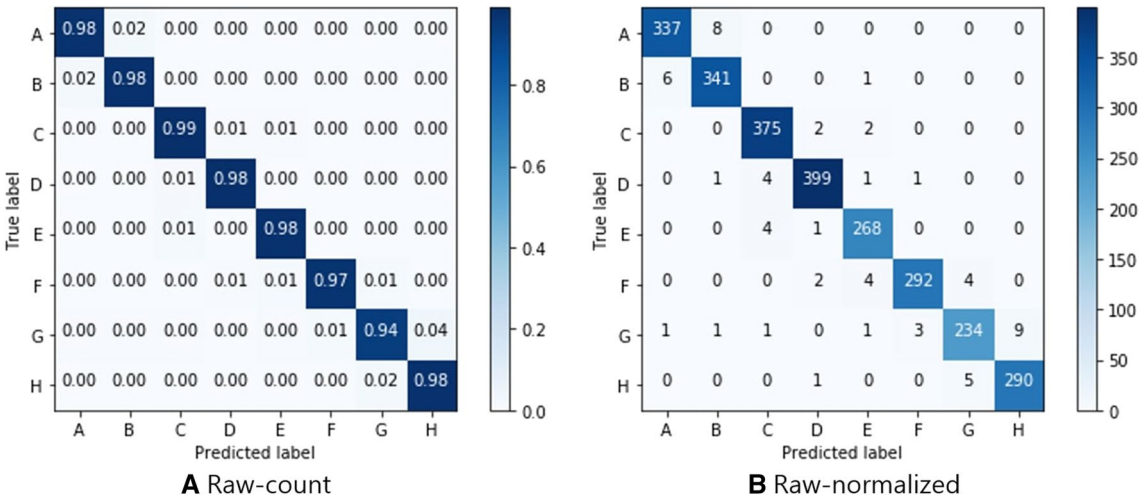

Fig. 4 Confusion matrices of classification results for the eight locations (multi-class classification): **A** raw count and **B** row-normalized values

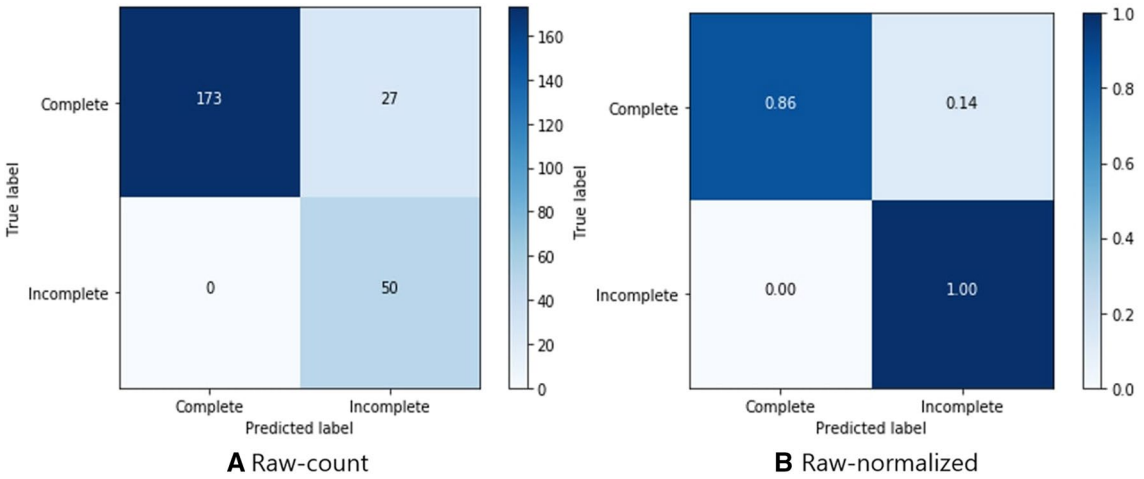

Fig. 5 Confusion matrices for validation of inspection completeness (binary classification): **A** raw count and **B** row-normalized values

Processing time analysis

The time requirements of a model vary based on the number of its trainable parameters. To test a single image, the mean prediction time, including the preprocessing time, was 0.26 s.

Discussion

Our model successfully classified the EGD images into eight regions of the UGI tract with 97.58% accuracy and evaluated the completeness of EGD studies with 89.20% accuracy. Using these two phases, we were able to construct a model that could be used for quality control of EGD photo documentation.

There are a few studies that used AI to recognize the anatomical location of EGD images. Takiyama et al. used 27,335 EGD images and classified them into four classes (larynx, esophagus, stomach, and duodenum) and subclassified the images of the stomach into three classes (upper, middle, and lower) [25]. This study highlighted the potential of using AI in EGD images but had limited clinical application because they categorized EGD images to four organs: larynx, esophagus, stomach and duodenum. WISENSE, an AI system introduced by Wu et al. made a significant progress in the application of AI into EGD, not only in the area of images but also in real-time [26, 27]. The study utilized the VGG-16 network to classify 26 locations of EGD and applied WISENSE to real-time EGD for monitoring blindspots through reinforcement learning with an accuracy of 90.02%. Compared to previous studies, we tried to build

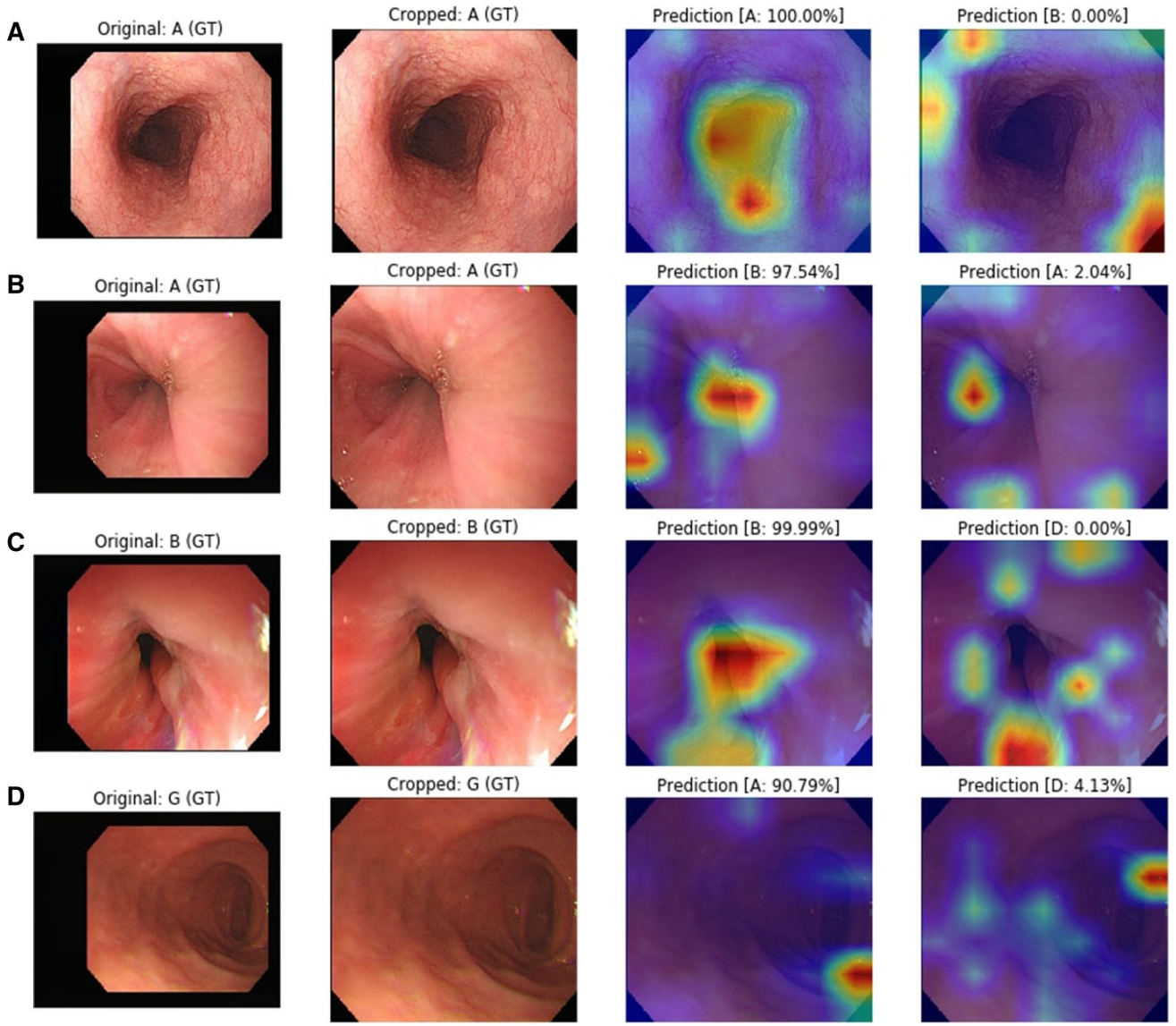

Fig. 6 Heatmaps of endoscopic images. Images **A**, **B** are of mid-esophagus at the same location with different air-insufflation levels, **C** is of the gastroesophageal junction, and **D** is of the duodenal bulb. **A**, **C** are examples of correctly classified images while **B**, **D** are incorrectly classified images. An area displayed in red is more activated than an area in blue in a particular region

our model especially useful for evaluating the completeness of visualization of EGD procedures in large number, considering its application in the EGD quality control program. Currently, EGD quality control, especially related to photo documentation, can only be performed by experts in the endoscopy, but reviewing all EGD procedures takes a lot of time, energy, and resources, making it nearly impossible. Instead, we constructed our model to aid the experts in hopes of taking over the task in the near future. To build a model for this purpose, our study applied widely accepted guidelines for EGD photo documentation with eight locations, which is suitable for the purpose of classification and determination of EGD completeness in already stored images. Our study was also designed in two steps, multi-class classification for evaluation in image level and binary classification for evaluation in patient/procedure level not only to match the images with anatomical locations but also to determine the completeness of the procedure. Our study showed 97.58% accuracy in image classification, which was similar or higher than 97% and 90.02% accuracies of previous studies, and 89.20% of accuracy in evaluation of the completeness of EGD procedures. Our results support the possibility of substituting artificial intelligence for repetitive tasks in the field of experts through our success in this field.

In the domain of medical image analysis, it is crucial to find the reason a model fails for particular images. The

probability heatmap helps us not only understand how the model works but also learn why the model failed so that we can improve it. Figure 6A and B were taken at the same location, mid-esophagus, with a different air-insufflation level. Figure 6A was correctly classified as location A, while Fig. 6B was incorrectly classified as location B. The original images of Fig. 6B and C look similar, and our model predicted both images to be in location B, but they were captured at A and B, respectively. These examples indicate that in order for the model to classify the image correctly, the image quality and confounding factors need to be controlled.

Besides algorithmic error, there are several features of collected images that render them incorrectly classified, including the aforementioned image quality and confounding factors. However, some of these features could not be avoided because of the retrospective design of the study. The same location or lesion can be seen or documented differently depending on many factors such as the level of air inflation, distance from the wall, luminosity, and rotation angle. In addition, images may contain several characteristics that hinder proper classification, such as mucous, inflammation, gastric juice, bubbles, blood, abnormal lesion, external compression, and out-of-focus. Moreover, there is a clear distinction between organs, but there is no clear cut-off margin within the organ. Thus, a disagreement among the clinicians regarding the location is possible. Additionally, images occasionally cover more than one location, which makes it difficult to classify the locations. We did not control these factors because they are common in endoscopic images; thus, the model should be able to compensate for them in order to be effective and widely used in the clinical setting.

Based on the confusion matrix (Fig. 4), the most confusing area in our study was between locations A and B, corresponding to the esophagus and GEJ, respectively. The esophagus is a long narrow cylindrical organ that is collapsed at rest and has a peristaltic contraction. These two conditions can make the esophagus similar in appearance to GEJ, as shown in Fig. 6B and C. Furthermore, if GEJ is loose or open during the contraction, these conditions could lead GEJ to be incorrectly classified or labeled as a more proximal esophagus. GEJ can be recognized visually from the rest of the esophagus by squamocolumnar junction called the Z-line. However, occasionally the Z-line cannot be seen clearly even if the images were taken at GEJ, and it requires a lot of effort by endoscopists to identify the GEJ correctly. To improve the accuracy in differentiating the esophagus from GEJ, we propose adding a location or sequence information to the photo, but this proposal requires further research.

There are several approaches to improve the overall accuracy of our study. For our model to be able to cope with different clinical settings and overcome the error factors mentioned above, additional image data are necessary. In addition, minimization of a data loss during preprocessing, modification of our network algorithm, and optimization of the hyperparameter may be able to improve the outcome of the model. In addition, using our model in a real-time setting, although not currently applicable, could provide the location information and solve the problems resulting from the limited number of images.

The average number of images endoscopists take per EGD procedure is unknown, but classification using our model takes about 2.1 s for eight images, 5.2 s for 20 images, and 7.8 s for 30 images, including preprocessing time. If we estimate the number of images for a single EGD procedure to be 20, it would take less than 9 min for the model to classify images from one hundred patients and to differentiate complete EGD tests from incomplete ones. About 6 million EGD procedures are performed annually in the United States, so, using a single computer, it would take our model about 361 days and 2 h to classify and determine the completeness of all EGD procedures performed in the United States in a year [28].

Our study has a few limitations. The study was performed in a single tertiary center using endoscopes from a single supplier, so our model needs further validation in different settings with larger image datasets. We did not use endoscopic images of patients with the mucosal lesion, such as gastric ulcer or cancer, or with surgically altered anatomy. Our model needs to be validated with a larger dataset before it can be used in clinical settings. Moreover, the model flexibility is quite low, so if quality control was to be applied to different photo documentation methods, the model would have to be re-trained with re-annotated data. The number of experts reviewing the images was relatively small.

Conclusion

In summary, we developed a ready-to-use program for EGD quality control in terms of completeness of EGD. Our model successfully classified the EGD images into its anatomical locations and evaluated the completeness of EGD procedures. With its validation at multi-institutional level, we expect our program to be able to relieve the burden of the endoscopists. Our classification results could also be the basis for further AI research in detecting abnormal lesions in the UGI tract and ultimately developing robotic EGD and capsule EGD.

Acknowledgements This work was supported by National IT Industry Promotion Agency (A0105-20-1007) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A1B07048202).

Compliance with ethical standards

Disclosures Seong Ji Choi, Mohammad Azam Khan, Hyuk Soon Choi, Jaegul Choo, Jae Min Lee, Soonwook Kwon, Bora Keum, and Hoon Jai Chun have no conflicts of interest or financial ties to disclose.

Ethical approval This study was conducted in accordance with the Helsinki Declaration, and the Ethics Committee of Korea University Anam Hospital approved this study (2019AN0253).

References

1. Hamashima C, Systematic Review Group and Guideline Development Group for Gastric Cancer Screening Guidelines (2018) Update version of the Japanese guidelines for gastric cancer screening. Jpn J Clin Oncol 48:673–683
2. Jun JK, Choi KS, Lee HY, Suh M, Park B, Song SH, Jung KW, Lee CW, Choi IJ, Park EC, Lee D (2017) Effectiveness of the Korean National cancer screening program in reducing gastric cancer mortality. Gastroenterology 152:1319–1328
3. Choi KS, Jun JK, Park EC, Park S, Jung KW, Han MA, Choi IJ, Lee HY (2012) Performance of different gastric cancer screening methods in Korea: a population-based study. PLoS ONE 7:e50041
4. Choi KS, Suh M (2014) Screening for gastric cancer: the usefulness of endoscopy. Clin Endosc 47:490–496
5. ASGE Standards of Practice Committee, Ben-Menachem T, Decker GA, Early DS, Evans J, Fanelli RD, Fisher DA, Fisher L, Fukami N, Hwang JH, Ikenberry SO, Jain R, Jue TL, Khan KM, Krinsky ML, Malpas PM, Maple JT, Sharaf RN, Dominitz JA, Cash BD (2012) Adverse events of upper GI endoscopy. Gastrointest Endosc 76:707–718
6. Mema SC, Yang H, Vaska M, Elnitsky S, Jiang Z (2016) Integrated cancer screening performance indicators: a systematic review. PLoS ONE 11:e0161187
7. Rey JF, Lambert R, ESGE Quality Assurance Committee (2001) ESGE recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower GI endoscopy. Endoscopy 33:901–903
8. Marques S, Bispo M, Pimentel-Nunes P, Chagas C, Dinis-Ribeiro M (2017) Image documentation in gastrointestinal endoscopy: review of recommendations. GE Port J Gastroenterol 24:269–274
9. Bisschops R, Areia M, Coron E, Dobru D, Kaskas B, Kuvaev R, Pech O, Ragunath K, Weusten B, Familiari P, Domagk D, Valori R, Kaminski MF, Spada C, Bretthauer M, Bennett C, Senore C, Dinis-Ribeiro M, Rutter MD (2016) Performance measures for upper gastrointestinal endoscopy: a European Society of gastrointestinal endoscopy (ESGE) quality improvement initiative. Endoscopy 48:843–864
10. Buch VH, Ahmed I, Maruthappu M (2018) Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract 68:143–144
11. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, Baldi P (2018) Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155:1069–1078
12. Kominami Y, Yoshida S, Tanaka S, Sanomura Y, Hirakawa T, Raytchev B, Tamaki T, Koide T, Kaneda K, Chayama K (2016) Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. Gastrointest Endosc 83:643–649
13. Chen PJ, Lin MC, Lai MJ, Lin JC, Lu HH, Tseng VS (2018) Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology 154:568–575
14. Nakashima H, Kawahira H, Kawachi H, Sakaki N (2018) Artificial intelligence diagnosis of Helicobacter pylori infection using blue laser imaging-bright and linked color imaging: a single-center prospective study. Ann Gastroenterol 31:462–468
15. Kanesaka T, Lee TC, Uedo N, Lin KP, Chen HZ, Lee JY, Wang HP, Chang HT (2018) Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging. Gastrointest Endosc 87:1339–1344
16. de Souza LA, Palm C, Mendel R, Hook C, Ebigbo A, Probst A, Messmann H, Weber S, Papa JP (2018) A survey on Barrett's esophagus analysis using machine learning. Comput Biol Med 96:203–213
17. Trindade AJ, McKinley MJ, Fan C, Leggett CL, Kahn A, Pleskow DK (2019) Endoscopic surveillance of Barrett's esophagus using volumetric laser endomicroscopy with artificial intelligence image enhancement. Gastroenterology 157:303–305
18. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L (2019) PyTorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems, pp 8024–8035
19. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. In: Neural information processing systems, p 25
20. Hu J, Shen L, Albanie S, Sun G, Wu E (2019) Squeeze-and-excitation networks. In: IEEE transition pattern analytical machine intelligence
21. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316:2402–2410
22. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? Adv Neural Inf Process Syst 27:3320–3328
23. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. In: International conference on learning representations
24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV), pp 618–626
25. Takiyama H, Ozawa T, Ishihara S, Fujishiro M, Shichijo S, Nomura S, Miura M, Tada T (2018) Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. Sci Rep 8:7497
26. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, Jiang X, Huang X, Mu G, Wan X, Lv X, Gao J, Cui N, Hu S, Chen Y, Hu X, Li J, Chen D, Gong D, He X, Ding Q, Zhu X, Li S, Wei X, Li X, Wang X, Zhou J, Zhang M, Yu HG (2019) Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut 68:2161–2169
27. Chen D, Wu L, Li Y, Zhang J, Liu J, Huang L, Jiang X, Huang X, Mu G, Hu S, Hu X, Gong D, He X, Yu H (2020) Comparing blind spots of unsedated ultrafine, sedated, and unsedated conventional gastroscopy with and without artificial intelligence: a prospective, single-blind, 3-parallel-group, randomized, single-center trial. Gastrointest Endosc 91:332–339
28. Peery AF, Crockett SD, Murphy CC, Lund JL, Dellon ES, Williams JL, Jensen ET, Shaheen NJ, Barritt AS, Lieber SR, Kochar B, Barnes EL, Fan YC, Pate V, Galanko J, Baron TH, Sandler RS (2019) Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2018. Gastroenterology 156:254–272