Anatomical classification of upper gastrointestinal organs under various image capture conditions using AlexNet

(AlexNet を用いた多様な撮影条件の上部消化管内視鏡画像群の臓器分類**)**

申請者　　弘前大学大学院医学研究科

病態制御科学領域　消化器内科学研究分野

氏名　　五十嵐　昌平

指導教授　　福田　眞作

Abstract

Background: Machine learning has led to several endoscopic studies about the automated localization of digestive lesions and prediction of cancer invasion depth. Training and validation dataset collection are required for a disease in each digestive organ under a similar image capture condition; this is the first step in system development. This data cleansing task in data collection causes a great burden among experienced endoscopists. Thus, this study classified upper gastrointestinal (GI) organ images obtained via routine esophagogastroduodenoscopy (EGD) into precise anatomical categories using AlexNet.

Method: In total, 85,246 raw upper GI endoscopic images from 441 patients with gastric cancer were collected retrospectively. The images were manually classified into 14 categories: 0) white-light (WL) stomach with indigo carmine (IC); 1) WL esophagus with iodine; 2) narrow-band (NB) esophagus; 3) NB stomach with IC; 4) NB stomach; 5) WL duodenum; 6) WL esophagus; 7) WL stomach; 8) NB oral–pharynx–larynx; 9) WL oral–pharynx–larynx; 10) WL scaling paper; 11) specimens; 12) WL muscle fibers during endoscopic submucosal dissection (ESD); and 13) others. AlexNet is a deep learning framework and was trained using 49,174 datasets and validated using 36,072 independent datasets.

Results: The accuracy rates of the training and validation dataset were 0.993 and 0.965, respectively.

Conclusions: A simple anatomical organ classifier using AlexNet was developed and found to be effective in data cleansing task for collection of EGD images. Moreover, it could be useful to both expert and non-expert endoscopists as well as engineers in retrospectively assessing upper GI images.

Key words: anatomical classification, upper gastrointestinal tract, endoscopy, convolutional neural network (CNN), artificial intelligence

1. Introduction

Physicians can process images and extract certain features from them via electronic endoscopy. Several studies have assessed the efficacy of feature extractions for computer-aided diagnosis (CAD). To determine the severity of ulcerative colitis, previous studies have used endoscopic images to characterize endoscopic features including mucosal patterns (spatial arrangements of mucosal color) and the degree of roughness on the mucosal surface [1][2]. A technique for characterizing the textural features that can be used to evaluate the risk of developing gastric cancers among *Helicobacter pylori*-positive patients was introduced [3]. Textural and color features were extracted from colonoscopy images to classify the status of the colon [4]. However, the diagnostic accuracy of feature engineering was limited due to challenges in extracting features for image analysis in gastrointestinal diseases.

The perceptron is a machine learning (ML) method and has long been used for the CAD of endoscopic images although this method used using trial and error to identify and extract feature quantities such as color, edge, and curve.

For several years, convolutional neural networks (CNNs), a deep learning method, have allowed engineers to solve limitations in feature engineering. The expansion of CNN-based supervised learning has significantly improved the diagnostic accuracy for different medical images. CNN could reduce the burden in feature extractions by transferring the engineering step

into a learning process using the backpropagation method [5]. This ML method has already been applied in the automated localization of gastric cancer in routine gastroscopies [6] and the automated detection of colon polyps [7].

The collection of required data is the first crucial step in supervised ML [8]. In clinical conventional upper gastrointestinal (GI) endoscopy, organ identification and knowledge about normal anatomical structures from the oral cavity to the duodenum are important in detecting abnormalities and diagnosing digestive diseases. When a lesion is suspected, a thorough assessment is performed under a variety of image capture conditions including white-light (WL) or narrow-band (NB) imaging with or without local coloring matter and magnification. Endoscopic treatments can also be provided under various image capture conditions as surveillance. These processes are not exceptional with the automatic detection of lesions using CNN-based CAD. Thus, to create a CNN based CAD system for endoscopic images, the dataset of one specific digestive disease in each organ must be collected.

Several studies about digestive lesion detectors have been conducted. However, only a few have reported organ classification and the anatomical location of upper GI organs. A previous study performed automatic anatomical classification of esophagogastroduodenoscopy (EGD) images into four main categories (larynx, esophagus, duodenum, and stomach) using regular WL images; the accuracy rate was 0.97 [9]. Wu et al. [10] used a high-quality deep CNN (DCNN) to

detect early gastric cancer and monitor blind spots in real-time endoscopy. Moreover, they classified WL EGD images into ten and 26 categories with an accuracy of 0.90 and 0.65, respectively.. Zhang et al. [11] developed a multi-task anatomy detection network (MT-AD), a real-time anatomical classification system, using the Single Shot MultiBox Detector (SSD). This system focused on three image types: informative, non-informative (including blur, defocus, specular reflections, and bubbles), and NB images. WL EGD images were classified into ten categories. The mean average precision rate was higher than 0.93 for each category except the squamocolumnar junction (SCJ) (0.84). However, the NB and non-informative images were not included in this study.

The NB imaging mode and local coloring matters are widely used in routine clinical endoscopy. However, whether the application of training and validation datasets in this approach is practical has not yet been validated. Thus, we previously focused on anatomical classification in routine GI endoscopic images under various image capture conditions. That is, GI endoscopic images were retrospectively classified into eight categories with the exclusion of no or less informative images such as blur, halation, bleeding, unidentifiable images, and therapeutic findings. An artificial neural network with a multi-perceptron was used. We identified color and edge in endoscopic images and developed hand-engineered features of color and spatial gradient histograms, which were considered to be feature quantities of endoscopic images. The accuracy

rate of the system for the validation dataset was 0.961 [12]. However, this study had a limitation that the use of exclusion criteria was not practical for the retrospective analysis of clinical endoscopic images because non-cleansed clinical endoscopic examinations include non-informative images.

To the best of our knowledge, there is no report about anatomical classification in routine inspective upper GI endoscopies including all endoscopic images. This study aimed to develop a CNN based CAD system for classifying upper GI organ images from a large set of routine endoscopic images taken under various clinical image capture conditions. Hence, the burden of data cleansing is reduced.

2. Methods

2.1 Preparation of Endoscopic Images

Upper GI endoscopy images were obtained from our hospital (Hirosaki University Hospital, Aomori, Japan) and were used for this single-center retrospective study. There were 85,246 anonymized images collected from 441 patients with gastric cancer who underwent upper GI endoscopies from 2017/01/01 to 2018/09/30. The endoscopes were GIF-H260, GIF-H260Z, GIF-Q260J, GIF-XP260N, GIF-2TQ260M, GIF-H290, GIF-H290Z, and GIF-HQ290. A EVIS LUCERA ELITE CV-290/CLV-290SL (Olympus Medical Systems, Co., Ltd., Tokyo, Japan) video system was used.

We manually classified these images into 14 categories according to the major pattern classification by anatomical organs with or without NB imaging, local coloring matters, and treatable devices. The subclassification of stomach location was not adopted in this study. Figure 1 shows 14 categories as follows: 0) WL stomach with indigo carmine (IC); 1) WL esophagus with iodine; 2) NB esophagus; 3) NB stomach with IC; 4) NB stomach; 5) WL duodenum; 6) WL esophagus; 7) WL stomach; 8) NB oral–pharynx–larynx; 9) WL oral–pharynx–larynx; 10) WL scaling paper for checking magnification; 11) specimens under WL or NB or WL with IC; 12) WL muscle fibers during the procedure of endoscopic submucosal dissection (ESD), and 13) others (unidentifiable images and images under procedures such as biopsy, marking, injection, mucosal excisions, and bleeding with or without treatable devices under all image capture conditions).

We established the category of "NB stomach with IC" because the presence of IC appeared as a green matter on the mucosal surface in those images and was clearly different from those without IC that appeared cyanic. The images were first classified into four organs (oral–pharynx–larynx, esophagus, stomach, and duodenum). Second, each organ image was subclassified according to imaging modes under WL or NB images. The images were then sorted according to whether there was existing coloring matter, iodine staining, and other treatable devices and classified into the 14 categories.

The examination of gastric cancer includes more information obtained from one GI endoscopic examination because of natural bleeding in the lesion and the characteristics of the mucosa such as presence of *H. pylori*-related inflammation, coloring matters, and procedures including ESD. We believed that the diversity of images could lead good performance system. Unidentifiable images in clinical endoscopy were not classified even by expert endoscopists. Thus, the category "others" was developed. The "WL muscle fibers during the procedures" was established because it was different from "others" as the images of the muscle fibers were taken at closer range rather than that of "others", which included only white muscle fibers in one frame.

Two endoscopists classified the endoscopic images in this study. An expert of endoscopy (YS) with more than 30 years of experience manually classified all images into 14 categories. Another endoscopist (SI) on training with 4 years of experience reclassified some images under the correct category. This task is not extremely difficult for endoscopists regardless of years of experience. This study was approved by the ethics committee at Hirosaki University Graduate School of Medicine on November 10, 2018 (approval number: 2018-1171).

2.2 Inclusion and Exclusion Criteria for the data

The study did not have any inclusion or exclusion criteria. The study aimed to establish an effective organ classifier that can be used in any common clinical settings and all types of images such as lesions, treatment devices, bleeding, magnification, and unidentifiable images (halation,

blackout, and defocused).

2.3 Training and Validation Dataset

The training data set comprised 49,174 images from 242 patients with gastric cancer who underwent upper GI endoscopies from January 01, 2017 to December 31, 2017. All images were manually classified into 14 categories as mentioned above.

To assess the performance of the proposed CNN as an organ classifier, another set of 36,072 images from 199 patients with gastric cancer who underwent upper GI endoscopies from January 01, 2018 to September 30, 2018 was used. These images were manually classified into 14 categories to validate the accuracy of the CNN using the same method. The training and validation dataset is shown in Table 1.

2.4 Architecture of the CNN

AlexNet (a CNN) and Pytorch (a moving framework) were utilized [13]. The architecture of AlexNet comprises five convolutional layers followed by maximum pooling layers, three fully connected layers, and finally a 1000-way Softmax classifier [14]. We obtained the source codes for AlexNet from the internet [15]. The original acquired images with 1000 × 870 pixels were converted into images with 227 × 227 pixels. We tuned hyper parameters, which were set by a human, as follows: number of training epochs, 50; batch size, 128; learning rate, 0.00025 via trial and error; and number of the outer layers, 14 classes.

3. Results

The accuracy rates for the training and validation datasets were 0.993 and 0.965, respectively. The accuracy rate for each identical organ is shown in Table 2. The accuracy rate of the validation data was greater than 0.96. Thus, this system could be used in classifying daily routine endoscopic images into the correct category with an accuracy of 0.965.

In contrast, the accuracy rate for some categories such as “NB stomach with IC”, “WL duodenum”, and “others” was lower than 0.900. Table 3 shows the confusion matrix diagram indicating the results of the classification using the CNN. The true classes were on the vertical axis, and the predicted classes were on the horizontal axis. The NB images were more likely to be classified into different organs with NB, and the same is true for the WL images. The representative features of incorrect images in each category are shown in Figure 2. A clear hood was attached to the top of the scope for procedures to maintain a good distance to the lesion. This was misidentified as the esophagus because its linear lumen was similar to that of the esophagus. Similarly, the linear lumen of the larynx was misclassified as the esophagus. The “WL duodenum” was misclassified as the “WL stomach” because the fold of the duodenum was similar to that of the fornix and antrum of the stomach. The “specimens” and “others” were classified into various categories because they were sorted without identifying the image capture conditions for the WL or NB images.

4. Discussion

In this study, we successfully developed an organ classifier system with a high accuracy for upper GI endoscopic images obtained under various clinical image capture conditions. There were no exclusion criteria for 85,246 training and validation datasets. This system has a simple architecture with generalization capabilities and less overfitting; it does not require any data-cleansing task and can be used to retrospectively collect data when developing the artificial intelligence (AI)-based supervised learning system.

The advancement of ML using CNN has enabled physicians to apply CAD of medical images in their specialized field. Supervised machine learning has been applied to various fields of medical images using their own techniques and datasets although there is no consensus on the exact metrics and datasets that should be used in each field [16]. The collection of image datasets and the diversity of medical images such as rotation, blur, halation, light contrast, materials during procedure, and matters on organs make CNN classification and detection difficult. Therefore, various techniques and devices such as preprocessing, transfer learning (TL) to avoid a data shortage in rare diseases, and various schemes that were implemented in the CNN have been applied. In the field of gastroenterology, various modalities such as the EGD, colonoscopy, and wireless capsule endoscopy (WCE) have been used.

In the esophagus, Mendel et al. [17] developed a GoogleNet-based CNN with an F1 score

of 0.91 to distinguish Barrett's epithelium from cancerous lesions using TL; this was developed with 100 qualified Barrett's esophagus images from publicly available datasets. In the colon, Riberio et al. [18] reported the feasibility of TL in detecting colon polyps by comparing two full-trained CNNs with pre-trained CNN. Misawa [7] developed CNN for polyp detection with an area under the curve (AUC) of 0.87, and Urban [19] achieved an accuracy of 0.964 and AUC of 0.991 using the original colon polyps. Stidham et al. [20] estimated the severity of ulcerative colitis (UC) using a GoogleNet-based CNN. They classified colonoscopy images of patients with UC into two groups: the normal to mild group (Mayo score 0 or 1, which was the endoscopic evaluation of severity of UC) and the moderate to severe group (Mayo 2 or 3). These metrics had an AUC of 0.966, sensitivity (SN) of 0.83, specificity (SP) of 0.96, positive predictive value (PPV) of 0.87, and negative predictive value (NPV) of 0.94. The CNN performance was at the level of the human reviewers. In the small intestine, WCE is used worldwide because it is less invasive and convenient. However, data cleansing and lesion detection are required as one WCE examination yields tens of thousands of images. Aoki et al. [21] developed a SSD that trained 5,360 images to detect erosions and ulcers in WCE images. The accuracy, SN, SP, and AUC were 0.908, 0.882, 0.909, and 0.958, respectively. Segui et al. [22] focused on small intestinal motility characterization, and their system automatically classified WCE images into six categories (turbid, bubbles, clear blob, wrinkles, wall, and undefined) with a mean accuracy of 0.96. Zhou et al. [23]

successfully developed a GoogleNet-based CNN architecture using a pre-rotating scheme to quantitatively evaluate WCE images of celiac disease with a SN and SP both of 1.0. Leenhardt et al. [24] introduced the CNN for detection of GI angioectasia with SN of 1.0, SP of 0.96, PPV of 0.96, and NPV of 1.0. Thus, CNNs with interesting devices have been used to classify and detect lesions in medical images of GI tract.

EGD is the standard examination for the diagnosis of gastric cancer, atrophic gastritis, esophageal cancer, esophagitis, gastric or duodenal ulcer, and other digestive diseases. In a previous study with EGD, 11.3% of upper GI cancers including esophageal and gastric cancer were still missing three years before diagnosis [25]; this was attributed to limited experience in endoscopy and subtle changes in atrophic, inflammatory gastric mucosa and the lesion after *H. pylori* eradication. Recently, the development of image recognition using CNN or other deep learning methods has solved these problems, improved the accuracy of diagnosis, and prevented upper GI cancers from being overlooked [26]. Diagnostic systems with a high SN of 0.98 for esophageal cancer were developed using CNN. A total of 8,428 training datasets from 384 patients and 1,118 test images from 47 patients under WL and NB imaging were used in a previous study [27]. Another study reported the accuracy of a deep neural network system for the diagnosis of esophageal cancer. This was trained using 2,428 images and validated with 187 images; the accuracy reached 0.914 [28]. Hirasawa et al. [6] reported that the overall SN of a CNN-based

diagnostic system for detecting gastric cancers (trained using 13,584 images and validated with 2,296 images) was 0.922. These systems may have high performance at a level similar to that of an experienced endoscopist. Thus, several studies focused on the development of a CNN-based CAD for the detection of digestive lesions using EGD data. Furthermore, CNN recently contributed to the prediction of the invasion depth of esophageal carcinoma [29] and gastric carcinoma [30].

As stated above, since the most common form of ML is supervised learning, a large amount of adequate training and validation data were required to develop and validate such a high-performance system [31]. The number of studies on organ classifiers or the anatomical location of each organ is lower than studies about the automatic disease detection. However, some studies focused on and described the anatomical classification of upper GI organs (Table 4). Takiyama et al. [9] developed an automatic organ classifier that could group images into four main categories (larynx, esophagus, stomach, and duodenum) using regular WL images with 27,335 training images and 17,081 validation images. This system had a high accuracy of 0.97 although the training and test datasets were limited to WL and identifiable images alone. Wu et al. [10] developed a deep CNN (DCNN) to detect early gastric cancers and monitor blind spots in a real-time EGD examination. They showed that the DCNN could classify 24,549 EGD images into ten categories (esophagus, squamocolumnar junction [SCJ], duodenal bulb, descending duodenum,

antrum, lower body in forward view, middle-upper body, angle, fundus, and retroflex view of the middle-upper body) with 26 categories. The accuracy rates were 0.90 and 0.65, respectively. Only WL images were used for datasets in this study, and detailed classification into 26 categories in each organ was slightly complicated. Zhang et al. [11] developed an MT-AD that can be used for the classification of EGD images according to quality (informative, noninformative, and NB images) using a combination of anatomical detection and classification processes. During anatomical detection, the constructed SSD model was trained using 59,513 images that were divided into ten categories (esophagus, SCJ, cardia, fundus, body, antrum, angle, pylorus, duodenal bulb, and descending duodenum) and validated using 15,762 images. The average precision was high in all categories (0.937) except SCJ (0.84). In clinical practice, no or less informative images were included and should not be excluded in other analyses. In this respect, we could successfully classify all EGD images in accordance with image capture conditions including NB and noninformative images.

Cogan et al. [32] developed high accuracy techniques of the modular adaptive preprocessing for GI tract images (MAPGI) for anatomical landmarks and disease states without overfitting. This technique (data preprocessing and augmentation methods along with edge removal, contrast enhancement, filtering, color mapping, and scaling) maximized the performance of constructed CNN. In this study, the Kvasir dataset was adopted and is the dataset

of GI tract images confirmed by endoscopists. The images consisted of eight categories: three "anatomical landmarks" as esophagogastric junction (EGJ), pylorus, and cecum; three "abnormal findings" as esophagitis, colon polyps, and UC; and two "polyp removal images" as polyps with indigo carmine or injected in submucosal tissues; and post resection images. However, it is difficult to apply the Kvasir dataset in the classification of clinical upper GI endoscopic images because less information was obtained from it, and these were limited to only three organs. Moccia et al. presented publicly obtainable datasets recorded during interventional medicine such as the Kvasir dataset [16]. The use of these datasets is practical although they are biased and limited to few specific regions (e.g., gastric cancer, Barrett's esophagus, colon polyps, and celiac disease in the GI field). Because of this, researchers created an original dataset in each organ and disease from several images recorded in their institution. Similarly, we collected EGD images from the database of our hospital. In general, the pictorial data of EGD were obtained from patient records in examination databases. The endoscopic data of each patient include mixed upper GI endoscopic images in routine examinations conducted under various imaging conditions. These might be in or out of focus with or without various magnifications and exhibit excessive halation (or blackout) under the condition of WL or NB imaging, local coloring matter to enhance the contrast, iodine staining, and therapeutic artifacts. Thus, endoscopists would have difficulty in selecting gastric cancer images alone from one patient record. Data cleansing must be performed

and selects each organ including the stomach for gastric diseases. However, the task is not always feasible, and data cleansing and classification are time-consuming processes for clinicians. The classification of all images (85,246) into 14 categories in this study was of note.

From the perspective of CNN models, the literatures employed AlexNet-based CNN and indicated good performance. Zou et al. [33] classified WCE images into three anatomical categories (stomach, small intestine, and colon). They achieved an accuracy of 0.95 versus other classification models such as support vector machine. Su et al. [34] developed the CNN using TL for detection and classification of ascites cytopathology and then compared the CNN models (AlexNet, VGG16, GoogleNet, ResNet18, and ResNet50). All CNN models achieved high performance with an AUC >0.85 and precision of 0.958. The ResNet50 model achieved the best performance (AUC of 0.885, precision of 0.968, false negative rate (FNR) of 0.0473) in contrast to the AlexNet model (AUC of 0.868, precision of 0.964, FNR of 0.05) although the AlexNet model achieved a good performance.

Various CNN models have been adopted to examine the dataset in each organ. However, there are no established metrics in CNN strategies and thus collecting enough unbiased and qualified images for training and validation is important for CNN-based CAD. In this respect, our training dataset was adequate to train AlexNet for the classification of upper GI organs without any complex preprocessing or exclusions. The AlexNet that was retrained on our datasets showed

good performance versus other anatomical classification studies of EGD images (Table 4). Moreover, its performance was comparable to other medical classification systems [7]–[11] and [17]–[28] although it was somewhat difficult to compare.

The accuracy rate of the “WL duodenum” (0.806), “NB with IC stomach” (0.870), and “others” (0.884) was lower than that of the other categories (Table 2). The stomach with a clear hood attached to the top of the scope was misclassified as the esophagus because they have a similar linear lumen. Using a similar method, the linear lumen of the trachea was misclassified as the esophagus. The “WL duodenum” was categorized into the “WL stomach” because the fold of the duodenum is similar to that of the fornix and antrum of the stomach. The “WL muscle fibers during the procedures” were different from “others” because the muscle fiber images were captured at close range. The “specimens” and “others” were classified into various categories because they were sorted without proper identification of filming conditions. The differences in air volume, site of organs, digestive juice (bile, which appears yellow in WL mode), and peristaltic movement may be associated with the misclassification of images.

The errors committed by the system are usually understandable (Figure 2). A chronological order of images exists when classifying continuous image data from one patient record. This sequence of images helps endoscopists to identify each organ. However, even an expert endoscopist can misclassify the images without the use of a sequencer, which may be associated

with irresistible force. These unidentifiable images might be excluded from the training and validation dataset. However, the exclusion of unidentifiable images causes selection bias. In this context, future studies must be conducted to assess this paradox and establish a category of images considered unidentifiable by an endoscopist. Notably, this study showed that the classifier had a high accuracy even if routine endoscopic data were not excluded.

Therefore, the constructed CNN offered good performance for data collection and cleansing of upper GI endoscopic images due to its generalization capability and limited overfitting. This system may help clinicians including expert and non-expert endoscopists. This approach can facilitate the cleansing of endoscopic images.

5. Limitations

This study had several limitations. First, this is a single-center retrospective study. The system cannot be used in real-time examinations, and conducting a prospective study is challenging. Thus, the use of data from other institutions may improve degradation performance and prevent overfitting because each endoscopist capture images differently.

Second, the training or test data comprised selected images of gastric cancer lesions, which might have caused selection bias. Supervised learning requires the use of several datasets selected by endoscopists. Thus, selection bias might not be prevented. However, this study did not have any exclusion criteria, and we included all types of images unlike previous studies.

Third, we did not consider the structure-weighted level and color enhancement of the endoscopic system and the type of scopes. This might have affected the accuracy due to the difference in mucosal color. Only a few clinicians pay attention to the default settings of the endoscopic system. These settings were modified in minute examinations including magnifying endoscopy before endoscopic treatment; these changes may be better for examinations including magnifying endoscopy with similar settings.

6. Conclusion

The anatomical organ classifier used here was found to be effective in the retrospective analyses and collection of endoscopic data for supervised learning. It can also be used by both expert endoscopists and non-expert endoscopists in assessing upper GI images. In the future, we hope that endoscopists and engineers will apply CNN-based systems to collect training and validation dataset to develop new AI systems for various endoscopic fields.

7. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

8. References

1)Y. Sasaki, R. Hada, A. Munakata, Computer-aided grading system for endoscopic severity in patients with ulcerative colitis, Digestive Endoscopy (2003) 15: 206-209, https://doi.org/10.1046/j.1443-1661.2003.00246.x .

2) Y. Sasaki, S. Fukuda, T. Mikami, Endoscopic Quantification of Mucosal Surface roughness for grading severity of ulcerative colitis, Digestive Endoscopy (2008) 20: 2891-2898, https://doi.org/10.1111/j.1443-1661.2008.00778.x

3) Y. Sasaki, R. Hada, T. Yoshimura, et al, Computer-aided estimation for the risk of development of gastric cancer by image processing, Artificial Intelligence in Theory and Practice Ⅲ (2010) 197-204, https://doi.org/10.1007/978-3-642-15286-3_19.

4) M. P Tjoa and S. M Krishnan, Feature extraction for the analysis of colon status from the endoscopic images, BioMedical Engineering Online (2003) 2:9, https://doi.org/10.1186/1475-925X-2-9.

5) D. Shen, G. Wu, and H. Suk, Deep learning in medical image analysis, Annu Rev Biomed Eng (2017) 19: 221-248, https://doi.org/10.1146/annurev-bioeng-071516-044442.

6) T. Hirasawa, K. Aoyama, T. Tanimoto, et al, Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images, Gastric Cancer (2018) 21: 653-660, https://doi.org/10.1007/s10120-018-0793-2.

7) M. Misawa, S. Kudo, Y. Mori, et al, Artificial Intelligence-Assisted Polyp Detection for Colonoscopy:Initial Experience, Gastroenterology (2018) 154: 2027-2029, https://doi.org/10.1053/j.gastro.2018.04.003.

8)S.B.Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica (2007) 31: 249-268.

9) H. Takiyama, T. Ozawa, S. Ishihara, et al, Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks, SCIENTIFIC REPORTS (2018) 8:1-8, https://doi.org/10.1038/s41598-018-25842-6.

10) L. Wu, W. Zhou, X. Wan, et al, A deep neural network improves endoscopic detection of early gastric cancer without blind spots, Endoscopy (2019) 51(6): 522-531, https://doi.org/10.1055/a-0855-3532.

11) X. Zhang, T. Yu, W. Zheng, et al. Upper gastrointestinal anatomy detection with multi-task convolutional neural networks, Healthcare Technology Letters (2019) 6:176-180, https://doi.org/10.1049/htl.2019.0066.

12) S. Igarashi, Y. Sasaki, et al. Neural network system for identifying upper-gastrointestinal organs in endoscopic images. UEG journal abstract book (2019) 7(8S): 49

13) Available at https://github.com/pytorch/pytorch

14) A. Krizhevsky, I. Sutskever, G E. Hinton, ImageNet Classification with Deep Convolutional

Neural Networks, Communications of the ACM (2017) 60 (6), https://doi.org/10.1145/3065386.

15) Available at https://github.com.

16) S. Moccia, L. Romeo, L. Migliorelli, et al. Supervised CNN Strategies for Optical Image Segmentation and Classification in Interventional Medicine. Deep Learner Descriptors for Medical Applications (2020) 186: 213-236. https://doi.org/10.1007/978-3-030-42750-4_8

17) R. Mendel, A. Ebigbo, A. Probst, et al. Barrett's Esophagus Analysis Using Convolutional Neural Network. Bildverarbeitung für die Medizin (2017): 80-85. https://doi.org/10.1007/978-3-662-54345-0_23

18) E. Ribeiro, A. Uhll, G. Wimmer, et al. Exploring Deep Learning and Transfer Learning for Colonic Polyp Detection. Computational and Mathematical Methods in Medicine (2016) Article ID 6584725 https://dx.doi.org/10.1155/2016/6584725

19) G. Urban, P. Tripathi, T. Alkayali, et al, Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy, Gastroenterology (2018) 155(4): 1069-1078, https://doi.org/10.1053/j.gastro.2018.06.037.

20) R. Stidham, W. Liu, S. Bishu, et al. Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis.

Gastroenterology and Hepatology (2019); 2(5): e193963.

https://doi.org/10.1001/jamanetworkopen.2019.3963

21) T. Aoki, A. Yamada, K. Aoyama, et al, Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network, Gastrointestinal Endoscopy, (2019) 89 (2): 357-363e2, https://doi.org/10.1016/j.gie.2018.10.027.

22) S. Segui, M. Drozdzal, G. Pascual, et al. Generic feature learning for wireless capsule endoscopy analysis. Computers in Biology and Medicine (2016) 79:163-172.

http://dx.doi.org/10.1016/j..compbiomed.2016.10.011

23) T. Zhou, G. Han, Z. Lin, et al. Quantitative analysis of patients with celiac disease by video capusule endoscopy: A deep learning method. Computers in Biology and Medicine (2017) 85: 1-6. https://dx.doi.org/10.1016/j.compbiomed.2017.03.031

24) R. Leenhardt, P. Vasseur, C. Li, et al, A neural network algorithm for detection of GI angioectasia during small-bowel capsule endoscopy, Gastrointestinal Endoscopy, (2019) 89 (1) 189-194, https://doi.org/10.1016/j.gie.2018.06.036.

25) S. Menon, N. Trudgill. How commonly is upper gastrointestinal cancer missed at endoscopy? A meta-analysis, Endoscopy International Open (2014) 2:E46-50, http://dx.doi.org/s-0034-1365524.

26) Young Joo Yang, Chang Seok Bang, Application of artificial intelligence in gastroenterology. World Journal of Gastroenterology (2019) 25(14):1666-1683, https://doi.org/10.3748/wjg.v25.i14.1666.

27) Y. Horie, T. Yoshio, K. Aoyama, et al, Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks, Gastrointestinal Endoscopy, (2019) 89(1):25-32, https://doi.org/10.1016/j.gie.2018.07.037.

28) S. Cai, B. Li, W. Tan, et al, Using a deep learning system in endoscopy for screening of early esophageal squamous cell carcinoma (with video), Gastrointestinal Endoscopy (2019) 90(5): 25-32, https://doi.org/10.1016/j.gie.2019.06.044.

29) K. Nakagawa, R. Ishihara, K. Aoyama, et al, Classification for invasion depth of esophageal squamous cell carcinoma using a deep neural network compared with experienced endoscopists, Gastrointestinal endoscopy (2019) 90 (3): 407-414, https://doi.org/10.1016/j.gie.2019.04.245.

30) Y. Zhu, Q. Wang, M. Xu, et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy, Gastrointestinal endoscopy (2019) 89 (4):806-815, https://doi.org/10.1016/j.gie.2018.11.011.

31) Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. NATURE (2015) 521: 436-444, https://doi.org/10.1038/nature14539.

32)T. Cogan, M. Cogan, and L. Tamil, MAPGI: Accurate identification of anatomical landmarks

and diseased tissue in gastrointestinal tract using deep learning, Computers in Bio and Med (2019) 111:103351.https://doi.org/10.1016/j.compbiomed.2019.103351.

33) Y. Zou, L. Li, Y. Wang, et al. Classifying Digestive Organs in Wireless Capsule Endoscopy Images Based on Deep Convolutional Neural Network. IEEE International Conference on Digital Signal Processing. (2015): 1274-1278. https://doi.org/10.1109/ICDSP.2015.7252086

34) F. Su, Y. Sun, Y. Hu, et al. Development and validation of a deep learning system for ascites cytopathology interpretation. Gastric Cancer (2020). https://doi.org/10.1007/s10120-020-01093-1

Figure 1. Representative images of the 14 categories: 0) white-light (WL) stomach with indigo carmine (IC); 1) WL esophagus with iodine; 2) narrow-band (NB) esophagus; 3) NB stomach with IC; 4) NB stomach; 5) WL duodenum; 6) WL esophagus; 7) WL stomach; 8) NB oral–pharynx–larynx; 9) WL oral–pharynx–larynx; 10) WL scaling paper for checking magnification; 11) specimens under WL or NB or WL with IC; 12) WL muscle fibers during ESD; and 13) others (stomach under procedures, including biopsy, marking, injection, mucosal excisions, bleeding with or without treatable devices under all image capture conditions).

(please print in color and 2 column fitting images)

Figure 2. Examples of misclassified pictures. A: Duodenum resembles the fornix of the stomach due to the folds. B: The picture of the antrum and pylorus is similar to that of the linear lumen of the esophagus. C: The clear hood at the top of the endoscopic image forms a linear lumen similar to that of the esophagus.

(please print in color and 2 column fitting images)

Table 1. Number of images in each training and validation dataset.

		Training data set	Validation data set
No	Categories	Number of pictures	Number of pictures
0	WL stomach with indigo carmine	7868	5816
1	WL esophagus with iodine	395	88
2	NB esophagus	2596	1616
3	NB stomach with IC	1878	1882
4	NB stomach	8076	5250
5	WL duodenum	1377	770
6	WL esophagus	2417	1225
7	WL stomach	17145	10996
8	NB oral-pharynx-larynx	1103	803
9	WL oral-pharynx-larynx	440	229
10	WL scaling paper	215	112
11	specimens	2794	3599
12	WL muscle fibers during procedures	941	1272
13	others	1929	2414
	Total	49174	36072

Table 2. Accuracy of the training and test dataset. The accuracy for each category is presented next to the number of images.

		Training data set		Validation data set	
No	Categories	correct images	accuracy	correct images	accuracy
0	WL stomach with indigo carmine	7812	0.993	5773	0.993
1	WL esophagus with iodine	395	1.000	88	1.000
2	NB esophagus	2588	0.997	1604	0.993
3	NB stomach with IC	1859	0.990	1638	0.870
4	NB stomach	8028	0.994	4988	0.950
5	WL duodenum	1300	0.944	621	0.806
6	WL esophagus	2412	0.998	1212	0.989
7	WL stomach	17094	0.997	10946	0.995
8	NB oral-pharynx-larynx	1072	0.972	794	0.989
9	WL oral-pharynx-larynx	423	0.961	223	0.974
10	WL scaling paper	215	1.000	112	1.000
11	specimens	2769	0.991	3534	0.982
12	WL muscle fibers during procedures	938	0.997	1153	0.906
13	others	1919	0.995	2134	0.884
	Total	48824	0.993	34820	0.965

Table 3. Confused matrix showing the classification results using the established convolutional learning network.

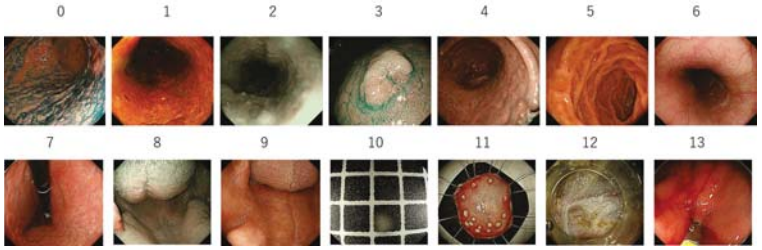
true category/predicted category		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
WL stomach with indigo carmine	**0**	5773	0	0	2	3	0	3	25	0	0	0	5	0	5
WL esophagus with iodine	**1**	0	88	0	0	0	0	0	0	0	0	0	0	0	0
NB esophagus	**2**	0	0	1604	0	4	0	0	0	8	0	0	0	0	0
NB stomach with IC	**3**	2	0	0	1638	235	0	0	0	1	0	0	6	0	0
NB stomach	**4**	1	1	99	144	4988	0	0	3	0	0	0	14	0	0
WL duodenum	**5**	0	1	0	0	0	621	0	148	0	0	0	0	0	0
WL esophagus	**6**	2	0	0	0	0	0	1212	11	0	0	0	0	0	0
WL stomach	**7**	10	1	0	0	2	14	16	10946	0	0	0	1	0	6
NB oral-pharynx-larynx	**8**	0	0	5	0	4	0	0	0	794	0	0	0	0	0
WL oral-pharynx-larynx	**9**	0	0	0	0	0	0	1	5	0	223	0	0	0	0
WL scaling paper	**10**	0	0	0	0	0	0	0	0	0	0	112	0	0	0
specimens	**11**	11	0	0	1	47	0	1	5	0	0	0	3534	0	0
WL muscle fibers during procedures	**12**	0	0	0	0	0	0	0	1	0	0	0	0	1153	118
others	**13**	60	0	0	2	17	2	0	45	0	1	0	4	149	2134

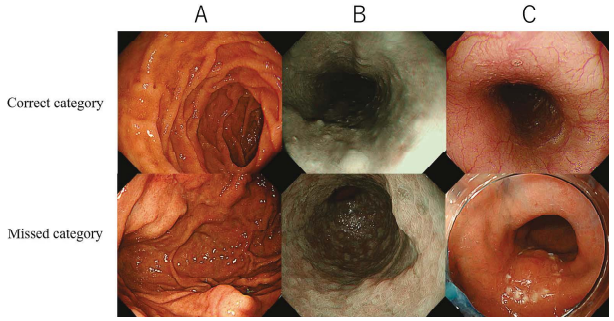
Table 4. List of convolutional learning network applications to upper gastrointestinal anatomical classification.

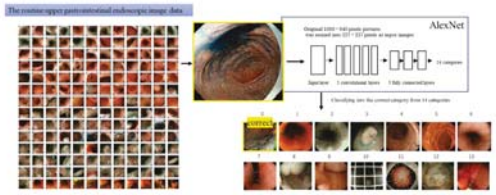
year	author	modality	features of dataset	training dataset	test dataset	categories	accuracy
2018	Takiyama	CNN(GoogLeNet)	only WL images, which excluded magnifying or unidentifiable images	27335	17081	4	0.974
2019	Wu	CNN(VGG-16, ResNet)	only WL images	19503	4876	10, 26	0.900, 0.659
2019	Zhang	SSD (MT-AD)	combination of process regognition of three types (WL informative, noninformative, NB images)	59513	15762	10	0.937*
			only WL informative images were classified into 10 categories	—	—	—	—
2020	Our study	CNN(AlexNet)	all images including WL, NB images with or without unidentifiable images	49174	36072	14	0.965

SSD:Single Shot MultiBox Detector, MT-AD:multi-task anatomy detection

*mean average precision


0
1
2
3
4
5
6
7
8
9
10
11
12
13


A
B
C
Correct category
Missed category


The routine upper gastrointestinal endoscopic image data
AlexNet
Original 1350 × 940 pixels pictures
Was resized into 227 × 227 pixels as input images
Input layer
5 convolutional layers
3 fully connected layers
14 categories
Classifying into the correct category from 14 categories
NORMAL