國立臺灣大學電機資訊學院生醫電子與資訊學研究所

博士論文

Graduate Institute of Biomedical Electronics and Bioinformatics
College of Electrical Engineering and Computer Science
National Taiwan University
Doctoral Dissertation

應用深度學習於腸胃道內視鏡對解剖部位分類及自動品質評估

Deep Learning Based Gastrointestinal Endoscopic Anatomy Classification and Automatic Quality Assessment

張 元 嚴

Yuan-Yen Chang

指導教授：李百祺 博士

Advisor: Pai-Chi Li, Ph.D.

中華民國 111 年 9 月

September 2022

doi:10.6342/NTU202203401

口試委員審定書

國立臺灣大學博士學位論文
口試委員會審定書
PhD DISSERTATION ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

應用深度學習於腸胃道內視鏡對解剖部位分類及自動品質評估

Deep Learning Based Gastrointestinal Endoscopic Anatomy Classification and Automatic Quality Assessment

本論文係張元嚴（學號 D06945008）在國立臺灣大學生醫電子與資訊學研究所完成之博士學位論文，於民國 111 年 9 月 6 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department / Institute of Biomedical Electronics and Bioinformatics on 6 September, 2022 have examined a PhD dissertation entitled above presented by Yuan-Yen Chang (student ID D06945008) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

（指導教授 Advisor）		

系主任/所長 Director:

doi:10.6342/NTU202203401

致謝

五年的時間倏忽而逝，很感謝在台大遇到的所有人事物，都是我成長的養分，使我滿載而歸的離開。首先，感謝父母一路上的支持與鼓勵，讓我能無後顧之憂地專心致力於研究上，順利完成學業。

原本大學就讀的是電機系，但碩士班誤打誤撞進入一個不熟悉且陌生的領域-超音波，為了自我的實踐，我選擇繼續攻讀博士學位，而選擇進入李百祺教授的實驗室，剛開始當然也是碰到一堆挫折及瓶頸，實驗不順利是家常便飯，但老師給了我天空自由發揮，並適時提供指教及引導，才有今天即將畢業的我。也感謝所有在求學生涯指導過我的教授，有了各位老師的教授的指導，使我不僅在專業領域上精進知識，更能將書本上的知識與臨床使用融會貫通，是其他很多科系是無法去實踐的，希望我能在生醫電資相關的範疇開創新的道路。感謝彰化基督教醫院的顏旭亨醫師提供研究資料，提出了臨床上未被滿足的需求，才有了後續這篇論文的產出。感謝口試委員張立群醫師、曾宇鳳教授、楊東霖教授、羅崇銘教授及論文計畫審查委員邱瀚模醫師、傅楸善教授的建議，使我的論文更趨完備。感謝實驗室的夥伴們，大家辛苦的研究透過報告，可以迅速讓我學到東西，勝讀十年書。感謝資工系的學長們，在研究上碰到問題，學長們總是能第一時間幫我解惑，真的受益良多。

感謝在我生命中的每一個人，謹以此致上最深的謝意，並將這份成果呈獻給各位，謝謝。

doi:10.6342/NTU202203401

摘要

消化道內視鏡檢查過程中的照片記錄是檢查品質的指標之一，但是該指標在內視鏡檢查中心難以自動測量和審核，而新興的人工智慧技術可能可幫忙解決此問題。首先，我們將利用深度學習(Deep Learning, DL)依據歐洲胃腸內鏡學會指引對上消化道內視鏡影像進行分成八個特定解剖位置，然後再評估是否記錄了所有解剖位置的影像，以照片記錄率的完整性作為內視鏡檢查的自動化品質評估指標。同時，上消化道影像分類和品質指標系統也將擴展到下消化道內視鏡檢查。然而，一個好的 DL 模型需要大量的訓練數據來進行模型開發，為了減少醫師標記時間，我們開發了一種加速數據準備，在此提出的方法中，先使用較小的已標記數據集來訓練基礎模型，然後由基礎模型對另一個較大的尚未標記數據集進行分類，醫生將可以快速查看和修改由基礎模型分類的結果，隨後可以使用校正過的數據集重新訓練增強模型以提高性能，完成的基礎模型和增強模型的準確率分別達到 96.29% 和 96.64%。在開發了好的分類模型後，我們將利用 12 位內視鏡醫師進行的 472 次內視鏡檢查進行品質評估指標實驗，可發現腺瘤檢出率較高的內視鏡醫師從咽部到十二指腸(60.0% vs 38.7%，p<0.0001)和從食道到十二指腸(83.0% vs 65.7%，p<0.0001)有較高的完整檢查率。而在下消化道內視鏡檢查品質指標實驗中共分析了 761 個真實世界的報告和大腸鏡檢查影像，電子報告盲腸檢出率為 99.34%，而所提品質指標系統的盲腸檢出率為 98.95%；使用電子報告和品質指標系統評估息肉切除率的一致率為 0.87；使用品質指標系統計算的檢查時間與醫生輸入的檢查時間存在良好的相關性(r = 0.959，p < 0.0001)。由上述實驗結果可得知本研究建立的內視鏡影像品質自動評估系統應可提升內視鏡檢查品質並為病患提供更好的照顧。

關鍵字：消化道內視鏡；內視鏡品質；深度學習

doi:10.6342/NTU202203401

Abstract

Photodocumentation is one endoscopy quality performance indicator; however, manually auditing this indicator is challenging in clinical practice. Artificial intelligence technology may help to solve this problem. In this study, the upper gastrointestinal (GI) endoscopy images are classified into eight specific anatomical landmarks according to the society of Gastrointestinal Endoscopy (ESGE) guideline by the proposed deep learning (DL) system. Then, this classification model can be used to assess whether all images of anatomical locations are documented and the completeness of the photodocumentation rate could be used as the quality indicator. Also, the upper GI classification and quality indicator system could be extended to the lower GI endoscopy. However, a good DL model requires a large amount of training data for model development. In order to reduce the labeling time, we develop an accelerated data preparation approach. In this proposed approach, a smaller labeled data set is first used to train the base model, and then another larger unlabeled data set is classified by the base model. The base model and enhanced model achieve total accuracy of 96.29% and 96.64%, respectively. After developing a good classification model, we can use this DL system to assess whether all images of anatomical locations are documented. The photodocumentation completeness rate could be used as the quality indicator for the endoscopist performance. A total of 472 upper GI endoscopies performed by 12 endoscopists are enrolled. The higher adenoma detection endoscopists have a higher complete examination rate (83.0% vs. 65.7%). For the proposed lower GI quality indicator system, 761 real-world examinations are analyzed. The accuracy of the proposed algorithm for the cecal intubation rate is 98.95% and the polypectomy agreement rate of the electronic reports and the DL algorithm is 0.87. A good correlation of DL withdrawal time between and that entered by the physician is found (r = 0.959). From the above experiments, the proposed DL endoscopy quality indicator system could help to improve the endoscopy procedure's performance and provide better patient care.

Keywords: gastrointestinal endoscopy; quality in endoscopy; deep learning

doi:10.6342/NTU202203401

Table of Contents

doi:10.6342/NTU202203401

doi:10.6342/NTU202203401

doi:10.6342/NTU202203401

List of Figures

doi:10.6342/NTU202203401

doi:10.6342/NTU202203401

doi:10.6342/NTU202203401

List of Tables

doi:10.6342/NTU202203401

Chapter 1. Introduction

Emerging artificial intelligence (AI) is a technology impacting several aspects of health care. AI is now being used successfully to provide a diagnostic aid for endoscopists to detect lesions found during endoscopic procedures, such as esophageal cancer [1], small bowel ulcers [2], and colorectal lesions [3]. However, technical factors such as shorter observation time, ampulla photodocumentation, or inadequate biopsy fragments have been reported as predictive diagnostic failure factors [4-6]. Hence, the quality of gastrointestinal (GI) procedures should be audited to provide better patient care and improve patient outcomes [7, 8]. The systematically acquired photographs from specific landmarks are recommended as an endoscopy quality indicator in several societies' guidelines [9]. The first photodocumentation guideline was introduced by the European Society of Gastrointestinal Endoscopy (EGSE) [8]. Eight specific endoscopy landmarks are recommended for acquiring images in the EGSE guideline. Also, a high-quality colonoscopy should be conducted to reduce colorectal cancer incidence [7, 10-13]. Several quality indicators such as cecal intubation rate (CIR), adequate bowel preparation, and colonoscopy withdrawal time are used in centers with electronic report systems to improve colonoscopy quality and patient outcome. Thus, a dedicated AI automated colonoscopy quality assessment system can reduce efforts to evaluate the quality indicators from captured endoscopy images [14, 15].

However, auditing the GI procedure quality requires endoscopy experts and may need an additional workforce and considerable time to check images, making auditing work nearly impossible [16]. Also, most endoscopy units lack a quality evaluation computer system. AI is an ideal solution for auditing; therefore, an AI quality indicator system for GI endoscopy is proposed in this study. Before developing a quality indicator

doi:10.6342/NTU202203401

system, a GI classification system should be developed to classify the endoscopic images in each procedure based on their anatomical locations. However, a large quantity of data is required to train the AI model. Many open image datasets, such as ImageNet [17], and endoscopy datasets, such as the HyperKvasir [18], are available for training the DL model. However, there is no available open data for training the endoscopic anatomy classification. The trained endoscopists are needed to label the images for training a new model. Hence, to save endoscopists' labeling time, a deep learning (DL) based endoscopic anatomy classification system [14] with an accelerated data preparation approach is developed and validated for potential clinical use in the first part of this study.

Based on the developed anatomy classification system, the systematically acquired photographs from specific landmarks could be checked from the classified images. In the proposed upper GI endoscopy quality system, the completeness of the photodocumentation rate is used as the quality indicator in the second part of this study. The completeness of photodocumentation could be determined by whether all the images of anatomical locations for each procedure are saved. This completeness of photodocumentation automatically computed from endoscopic images is compared with the clinical performance indicators. Also, the upper GI classification and quality indicator system could be extended to the lower GI endoscopy. Thus, the third part of this study aims to develop a DL-based algorithm for screening colonoscopy quality assessment.

Finally, some future works using other AI algorithms are discussed. For natural language processing (NLP), there are several well-known methods, such as Transformer [19] and Bidirectional Encoder Representations from Transformers (BERT) [20]. Based on Transformer, Vision Transformer (ViT) [21], and Shifted Windows (Swin)

doi:10.6342/NTU202203401

Transformer [22] have recently been proposed for image classification. In the future, we could investigate these new techniques to improve our classification and quality indictor systems. Also, self-supervised learning [23-25] has been proposed for automatically training models without human labeling for large amounts of data. Hence, in the future, a self-supervised learning technique could be further studied to train a more robust model using large amounts of endoscopy images without the physicians' labeling.

The organization of the thesis is as follows. Chapter 2 briefly introduces related works on convolutional neural networks, explainable AI, performance metrics, and gastrointestinal endoscopy. Chapter 3 introduces the proposed upper gastrointestinal endoscopic anatomy classification with an accelerated data preparation approach. Chapter 4 illustrates upper endoscopy quality assessment based on the developed anatomy classification system. Colonoscopy quality assessment extended from the upper endoscopy quality assessment is proposed in Chapter 5. Some future works are discussed in Chapter 6. Finally, conclusions are drawn in Chapter 7.

doi:10.6342/NTU202203401

Chapter 2. Related Works

2.1. CNN Models

Recently, deep learning (DL) has been applied to various medical image analyses and research works. In this study, we use the DL technique for the classification of endoscopic images. There are a lot of DL techniques proposed by researchers. The LeNet [26], which was proposed by Lecun et al. in 1998, is made up of seven layers consisting of three convolutional layers, two subsampling layers, and two fully connected layers. The eight-layer AlexNet [27], the winner in the ImageNet LSVRC (Large Scale Visual Recognition Competition) 2012 contest, was designed to train a large convolutional neural network (CNN) to classify the 1.2 million images into 1,000 different classes. There are several famous CNN architectures such as VGGNet [28], ResNet [29], and GoogLeNet (Inception v1) [30]. Later, Inception net has other versions v2 [31], v3 [32], and v4 [33].

2.1.1 ResNet

The deeper CNN architecture contains several complex parameters and has the vanishing gradient problem [34]. To solve this problem, short paths (identity loop) from initial layers to subsequent layers [35, 36] are adopted in ResNet [29, 37]. In **Figure 1**, a ResNet building block is shown.

doi:10.6342/NTU202203401

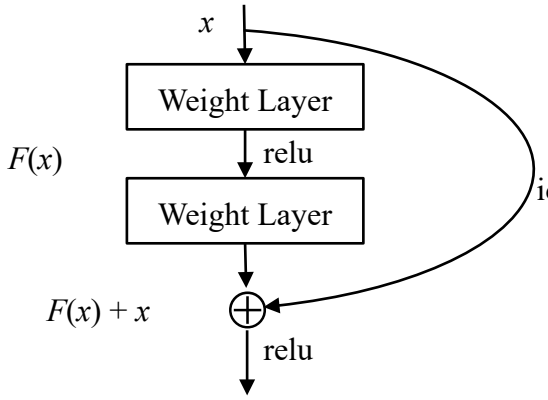


Figure 1. A ResNet building block.

2.1.2 ResNeXt

The Inception net [30] has an important common property, a split-transform-merge strategy. In an Inception block, as shown in **Figure 2**, the input is split into a few lower-dimensional 1×1 convolutions, transformed by a set of specialized filters such as 3×3, 5×5, etc., and then merged by concatenation. The Inception net is a successful multi-path architecture, and ResNet could be thought of as a two-path network in which one branch is the identity mapping. ResNeXt combines the multi-path and short-path concepts of Inception net and ResNet, as shown in **Figure 3**.

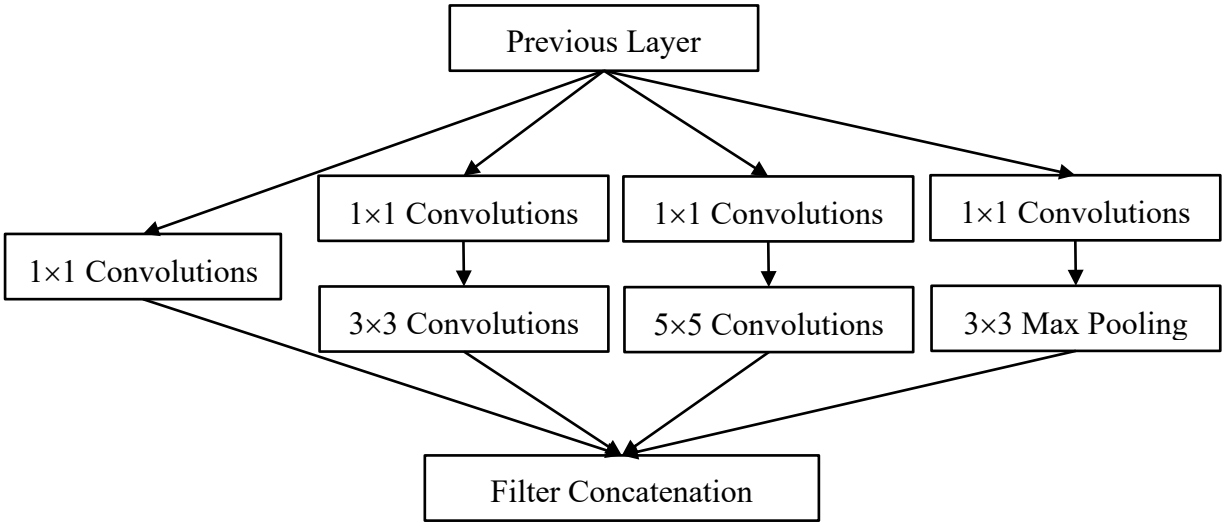


Figure 2. An Inception block.

doi:10.6342/NTU202203401

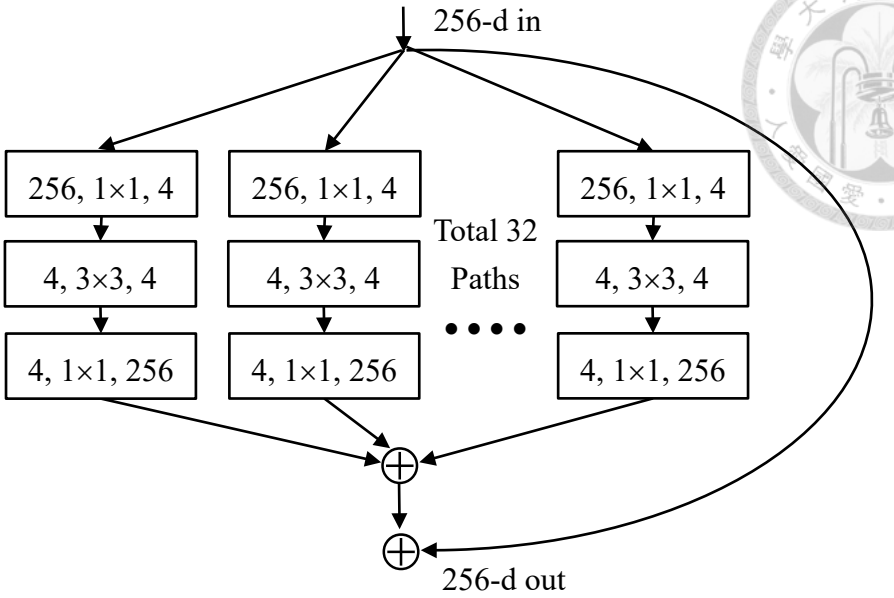


Figure 3. The ResNeXt net.

2.1.3 ResNeSt

The Selective Kernel Network (SK-Net) [38] consists of three operators: Split, Fuse, and Select. Multiple paths with various kernel sizes are generated by the Split operator, the Fuse operator combines and aggregates the information from multiple paths, and the Select operator aggregates the differently sized kernel feature maps according to the selection weights. The SK-Net is shown in **Figure 4**.

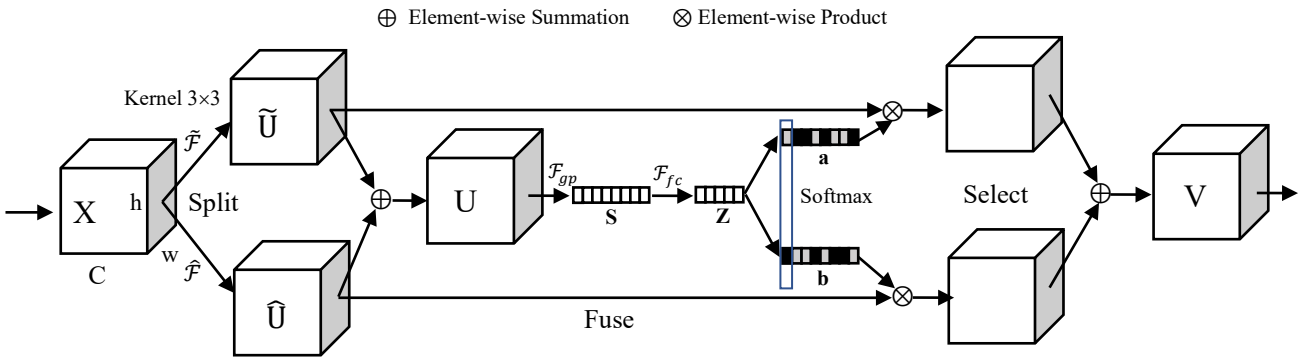


Figure 4. Selective Kernel network.

doi:10.6342/NTU202203401

The split-attention network (ResNeSt) [13] combines the multi-path and split-attention concepts of ResNeXt and SK-Net, as shown in **Figure 5**.

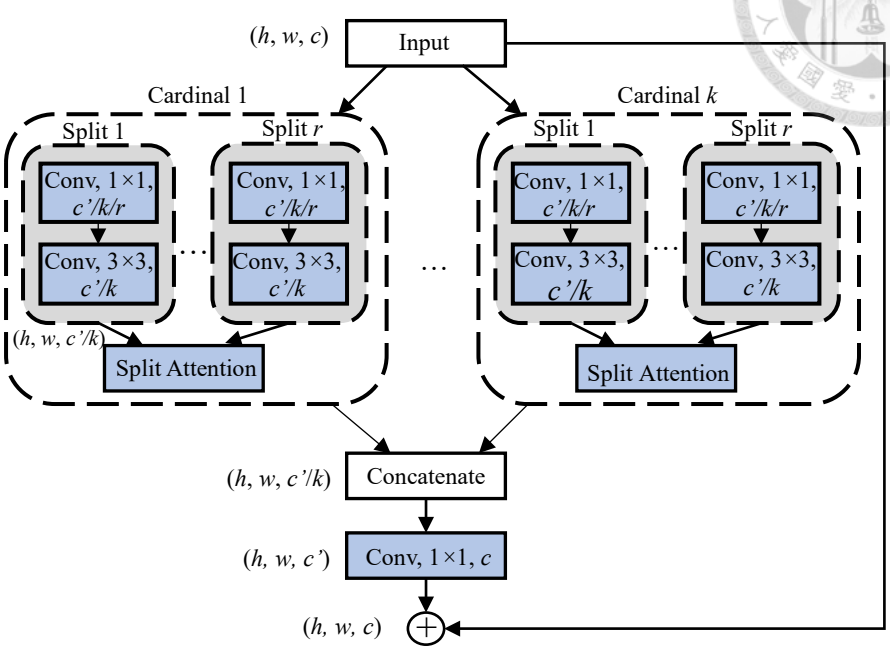


Figure 5. Split-attention network.

2.2. Explainable AI and Performance Metrics

2.2.1 Explainable AI

The gradient-weighted class activation mapping (Grad-CAM) [39] and anchor image [40] are applied to evaluate the model prediction interpretably. Also, the test result is visualized by using the t-distributed stochastic neighbor embedding (t-SNE) [41].

2.2.2 Performance Measures

Accuracy, precision, recall, and *F1-score*, which are defined as follows, are used to evaluate the study's performance. The harmonic average of *precision* and *recall*, *F1-*

doi:10.6342/NTU202203401

score, might be a better measure for seeking a *precision-recall* balance.

$$Accuracy = \frac{Corr}{N} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\text{-}score = \frac{2\times Precision\times Recall}{(Precision+Recall)} \quad (4)$$

where *Corr* is the corrected case number, *TP* is the true-positive case number, *FP* is the false-positive case number, *FN* is the false-negative case number, and *N* is the total case number. There are two averaging methods, macroaverage and microaverage, for these measures for multi-class classification defined as follows.

$$\text{Macro\_Precision} = \frac{1}{C}\sum_1^C Precision_i \quad (5)$$

$$Macro\_Recall = \frac{1}{C}\sum_1^C Recall_i \quad (6)$$

$$Macro\_F1\text{-}score = \frac{2\times Macro\_Precision\times Macro\_Recall}{(Macro\_Precision+Macro\_Recall)} \quad (7)$$

$$Weighted\_Precision = \frac{\sum_1^C N_i\times Precision_i}{\sum_1^C N_i} \quad (8)$$

$$Weighted\_Recall = \frac{\sum_1^C N_i\times Recall_i}{\sum_1^C N_i} \quad (9)$$

$$Weighted\_F1\text{-}score = \frac{2\times Weighted\_Precision\times Weighted\_Recall}{(Weighted\_Precision+Weighted\_Recall)} \quad (10)$$

where *C* is the class number, $Precision_i$ is the $precision_i$ for class *i*, $Recall_i$ is the $recall_i$ for class *i*.

2.2.3 McNemar's test

In this study, McNemar's test [42] is used to compare the predictive accuracy of two models. At first, the contingency table of the two tests is defined as **Table 1** reconstructed. The McNemar test statistic is

doi:10.6342/NTU202203401

$$\chi^2 = (|b - c|-1)^2/(b + c) \quad (11)$$

where χ^2 is chi-squared distribution.

Table 1. The contingency table for McNemar's test.

	Test 2 positive	Test 2 negative
Test 1 positive	*a*	*b*
Test 1 negative	*c*	*d*

2.3. Gastrointestinal Endoscopy

GI endoscopy is the most used procedure for diagnosing and treating patients of GI diseases. Auditing the procedure's quality has been advocated for better patient care.

2.3.1 Upper GI Endoscopy

The European Society of Gastrointestinal Endoscopy (ESGE) recommended acquiring eight specific endoscopy landmark images [8]. In this study, nine anatomical locations are used to classify the upper GI endoscopy images (0: pharynx, 1: esophagus, 2: esophageal-gastric junction, 3: gastric cardia, fundus or retroflex view, 4: gastric body, 5: gastric angle, 6: gastric antrum, 7: duodenal 1st portion, and 8: duodenal 2nd portion). The nine anatomical locations in upper GI are shown in **Figure 6**.

doi:10.6342/NTU202203401

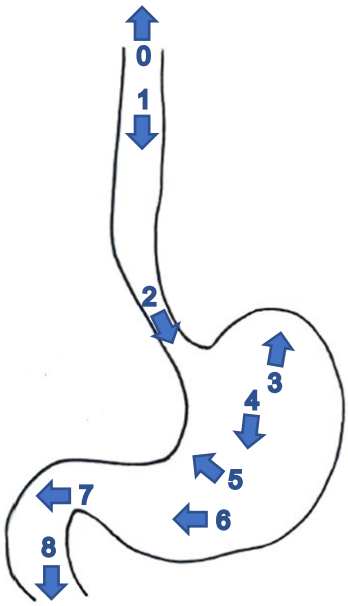


Figure 6. Nine anatomical locations in upper GI.

2.3.2 Duodenal Papilla

The L8 location can be further evaluated whether the whole or partial area of the duodenal papilla is present or not [6]. The ResNeSt model could be used to train the images of L8 location to classify into L8A ones without ampulla and L8B ones with ampulla. An image of the duodenal papilla is shown in **Figure 7**.

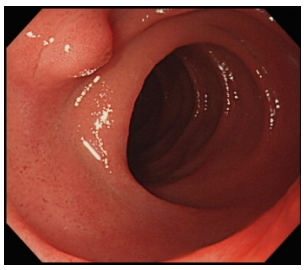


Figure 7. Duodenal Papilla.

2.3.3 White Light Endoscopy and Narrowband Endoscopy

White light endoscopy (WLE) is the standard endoscopy image, and narrowband endoscopy images (NBI) is an advanced optical method to improve the visualization of

doi:10.6342/NTU202203401

the mucosal surface architecture and microvascular pattern. The WLE and NBI images are shown in **Figure 8**.

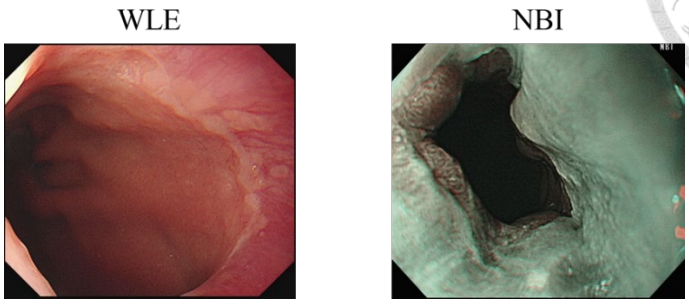


Figure 8. White light endoscopy (WLE) and narrowband endoscopy image (NBI).

2.3.4 Colonoscopy Quality Indicators

A high-quality colonoscopy image should be adopted to reduce the interval colorectal cancer incidence. Some quality indicators (**Figure 9**) have been used to improve the colonoscopy quality and patient outcomes. For example, a longer colonoscopy withdrawal time has been related to a better adenoma detection rate. Cecal intubation rate (CIR) is a complete colonoscopy indicator. To prevent missed polyps, adequate bowel preparation can help an endoscopist assess colonic mucosa, and a short follow-up interval is needed for inadequate bowel preparation patients.

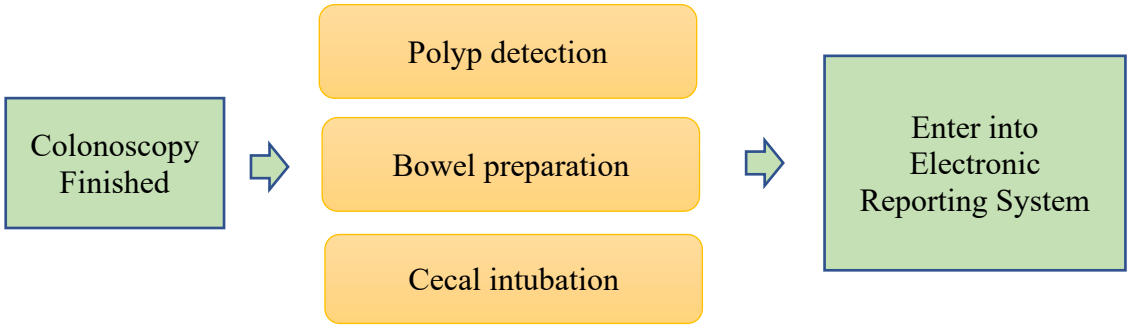


Figure 9. The current workflow for quality indicator acquisition.

doi:10.6342/NTU202203401

Chapter 3. Upper Gastrointestinal Endoscopic Anatomy Classification

3.1. Materials & Methods

3.1.1 Patients and Data Preparation

The patients who underwent diagnostic upper endoscopic examination at Changhua Christian Hospital from January 2020 to December 2020 are retrospectively reviewed. Two expert endoscopists with 15 years of therapeutic endoscopy experience reviewed the first white light endoscopy (WLE) dataset with the Olympus 260 or 290 series. Nine anatomical locations are used to classify the endoscopy images (0: pharynx, 1: esophagus, 2: esophageal-gastric junction, 3: gastric cardia, fundus or retroflex view, 4: gastric body, 5: gastric angle, 6: gastric antrum, 7: duodenal 1st portion, and 8: duodenal 2nd portion). The second image dataset from endoscopy records of patients from September 2020 to October 2020 at Changhua Christian Hospital is obtained for subsequent analysis. These datasets contain WLE and NBI. The WLE images constitute the majority of images captured in a typical upper endoscopy procedure, and NBI is minor. An independent dataset from Changhua Christian Hospital is used to validate the DL model. In **Figure 10**, the images included in this study are illustrated.

doi:10.6342/NTU202203401

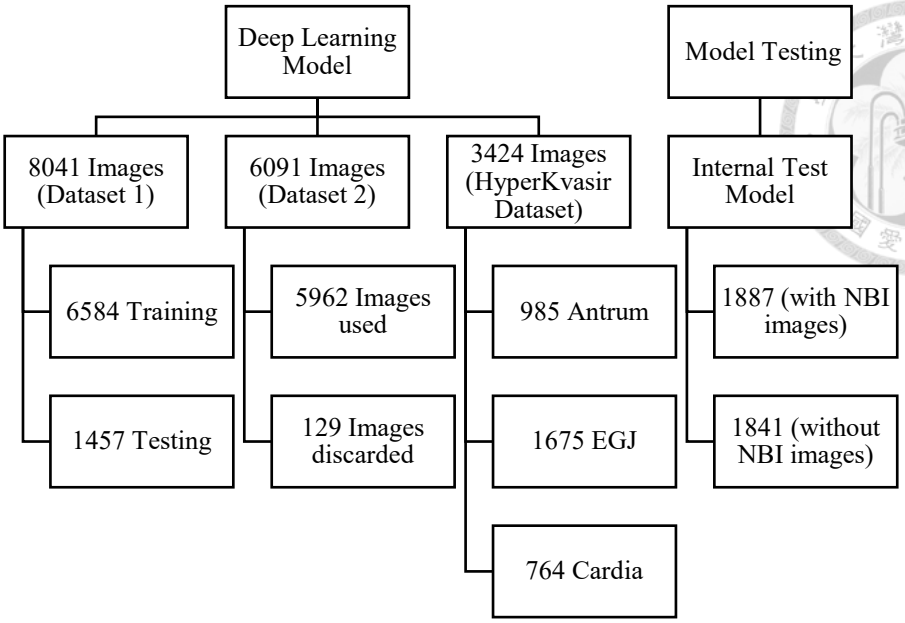
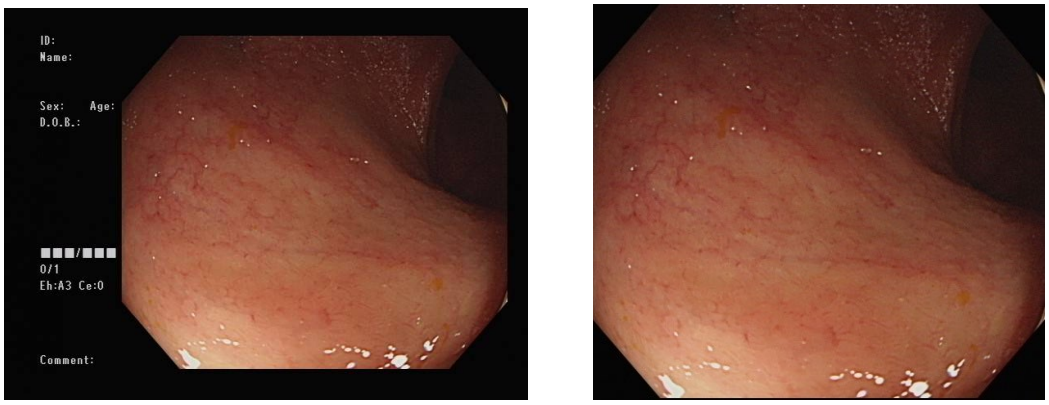


Figure 10. Flowchart of the images included in this study.

3.1.2 Preparation of Endoscopy Images for Deep Learning

In **Figure 11**, the original images are 640 × 480 pixels and are cropped to size 469 × 410 pixels. Because transfer learning with the pre-trained ImageNet model is applied to reduce training time, the 469 × 410 pixels images are downsampled into 224 × 224 pixels.



(a) (b)

Figure 11. (a) original image and (b) cropped image.

doi:10.6342/NTU202203401

3.2. Results

ResNeSt [43] ,which a deep residual network (ResNet) [29] variant, is used to train the DL model. This study uses ResNeSt with fifty layers deep pre-trained with ImageNet [17]. In order to solve the data imbalance problem, the focus loss function [44] is used as the loss function in the training process. The PyTorch framework using Python programming language on an Intel i7-10700 personal computer with Windows 10 and Nvidia GeForce RTX^{TM} 3090 GPU is used to develop the proposed system. Normalize and RandomResizedCrop in PyTorch package Transforms are used for training images.

3.2.1 Deep Learning Base Model Training on 1st Dataset

In **Table 2**, the training, validation, and test subsets are extracted from the first 8,041 image dataset to build the base model. The training and validation accuracies and losses are illustrated in **Figure 12**(a). **Figure 13**(a) shows the testing confusion matrix. **Table 3** lists the base model performance. The total accuracy is 96.29%, the macro-average precision is 96.17%, and the weighted average precision is 96.41%. The macro-average recall is 95.42% and the weighted average recall is 96.29%. The macro-average F1-score is 95.73% and the weighted average F1-score is 96.22%.

doi:10.6342/NTU202203401

Table 2. The image numbers of the training set, validation set, and test set.

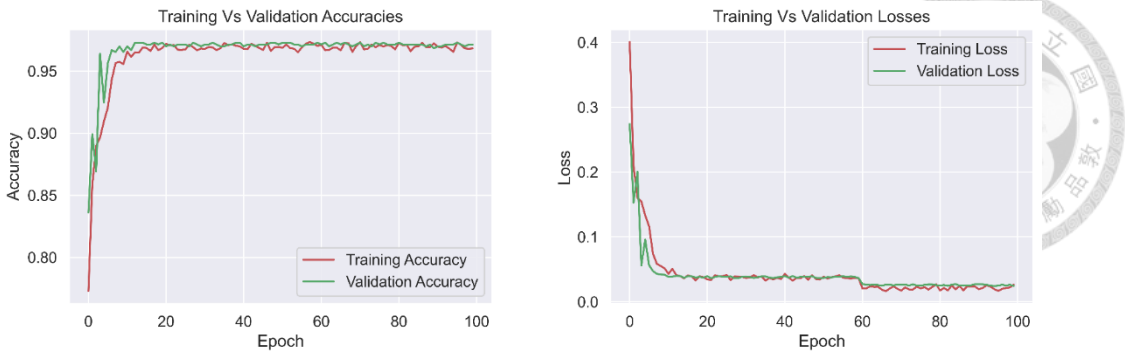
		Training	Validation	Test
0	Pharynx	168	19	32
1	Esophagus	800	93	114
2	Esophageal-gastric junction	638	69	123
3	Gastric cardia, fundus or retroflex view	929	103	231
4	Gastric body	909	101	417
5	Gastric angle	322	35	97
6	Gastric antrum	1050	120	167
7	Duodenal 1st portion	422	49	86
8	Duodenal 2nd portion	681	76	190
Total number		5919	665	1457

Table 3. Performance comparison for the base and enhanced models.

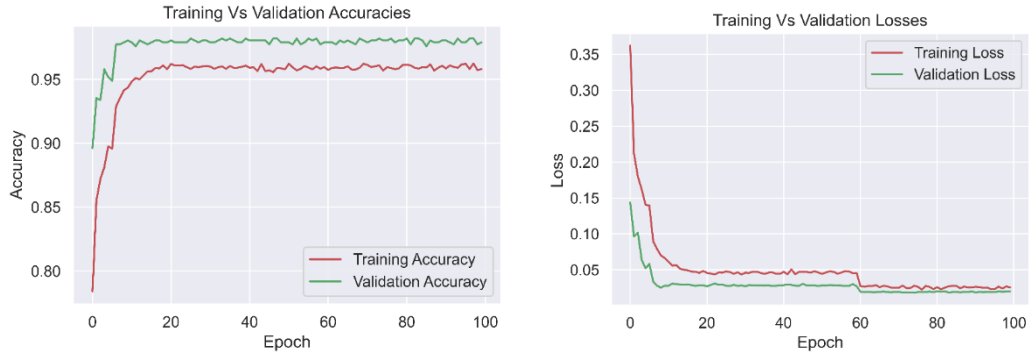
	Metrics	Class											
		All	Macro-average	Weighted average	0	1	2	3	4	5	6	7	8
Base	Accuracy	96.29											
	Precision		96.17	96.41	96.77	98.21	92.25	97.84	97.81	100.00	89.56	95.18	97.91
	Recall		95.42	96.29	93.75	96.49	96.75	97.84	96.40	89.69	97.60	91.86	98.42
	F1-score		95.73	96.30	95.24	97.35	94.44	97.84	97.10	94.57	93.41	93.49	98.16
Enhanced	Accuracy	96.64											
	Precision		96.51	96.68	96.97	97.32	93.65	96.62	97.84	98.86	93.10	96.34	97.88
	Recall		96.02	96.64	100.00	95.61	95.93	99.13	97.60	89.69	97.01	91.86	97.37
	F1-score		96.22	96.63	98.46	96.46	94.78	97.86	97.72	94.05	95.01	94.05	97.63
p-value		0.53			0.48	1.00	1.00	0.25	0.18	0.68	1.00	0.72	0.68

*A *p*-value less than 0.05 is considered statistically significant.

doi:10.6342/NTU202203401



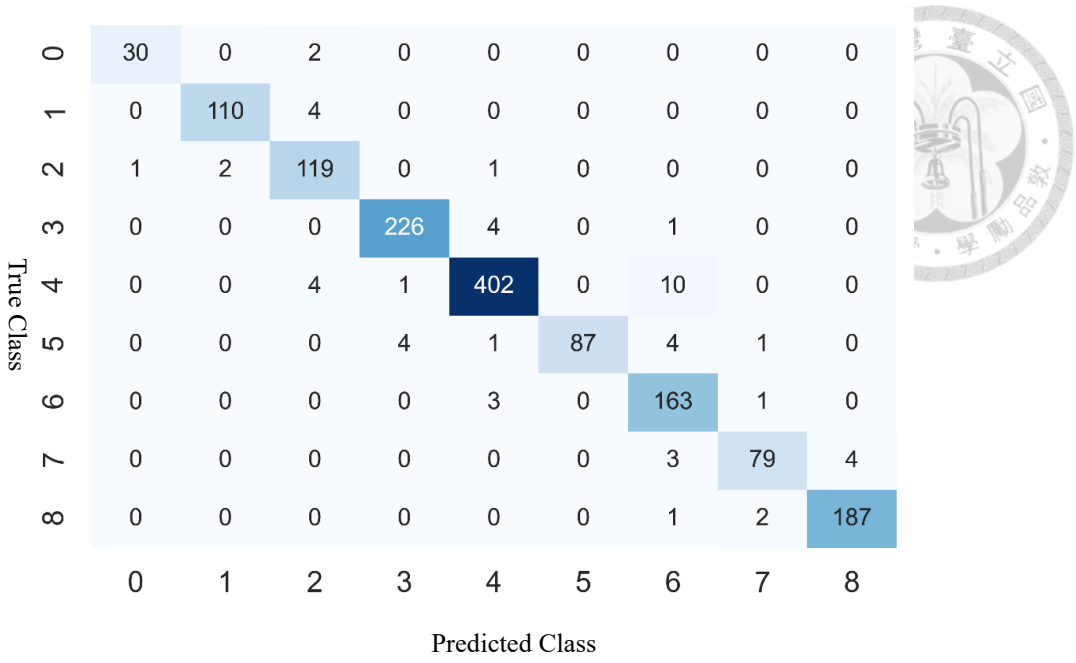
(a)



(b)

Figure 12. The training and validation accuracies and losses of training for the base and (b) enhanced models.

doi:10.6342/NTU202203401



(a)

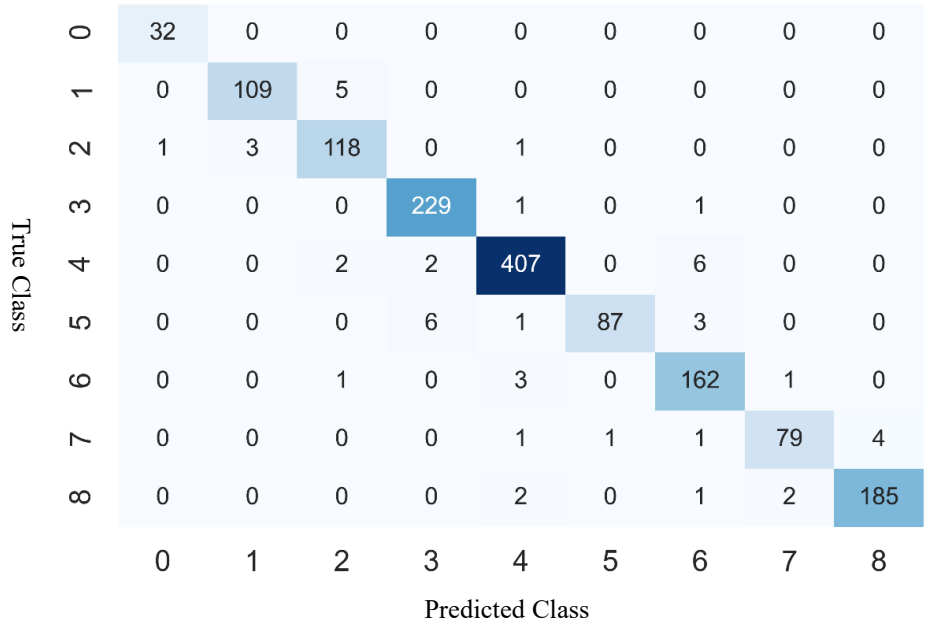


Figure 13. The testing confusion matrixes for (a) base and (b) enhanced models.

3.2.2 Training the Deep Learning Enhanced Model on the 2nd Dataset

Additional unclassified 6,091 images are added to the training set to improve the robustness of the trained model. The 6,091 unclassified images are first classified by

doi:10.6342/NTU202203401

the base model and then rechecked by a physician. Only 409 images (6.67%) in the 6,091 images are reclassified, and 129 images are discarded. Because the base model first classifies the unclassified images, the physician's review time is significantly reduced. There are 277 NBI images and 132 WLE images in 409 misclassified images. Because the base model is trained only with WLE images, most misclassified images are NBI images. For WLE images, the base model has already achieved excellent results. Moreover, we added 3,424 images that contain locations 2, 3, and 6 from the HyperKvasir dataset to enrich the training material [18] to train the enhanced model. **Table 4** lists the number of images of each class. **Figure 3**(b) shows the training and validation accuracies and losses. **Figure 4**(b) shows the testing confusion matrix. **Table 3** lists the performance comparison for the base and enhanced models. The total accuracy of enhanced model is 96.64%, the macro-average precision is 96.51% and the weighted average precision is 96.68%. The macro-average recall is 96.02% and the weighted average recall is 96.64%. The macro-average F1-score is 96.22% and the weighted average F1-score is 96.63%. The enhanced model is slightly better after using more training images than the previous base model. All the *p*-values are larger than 0.05 using the McNemar test.

doi:10.6342/NTU202203401

Table 4. The image number of training set for the enhanced model.

		Training
0	Pharynx	364
1	Esophagus	2166
2	Esophageal-gastric junction	3228
3	Gastric cardia, fundus or retroflex view	2247
4	Gastric body	2070
5	Gastric angle	551
6	Gastric antrum	2782
7	Duodenal 1st portion	777
8	Duodenal 2nd portion	1120
Total number		15305

3.2.3 Model Performance Evaluation for the Internal Test Dataset

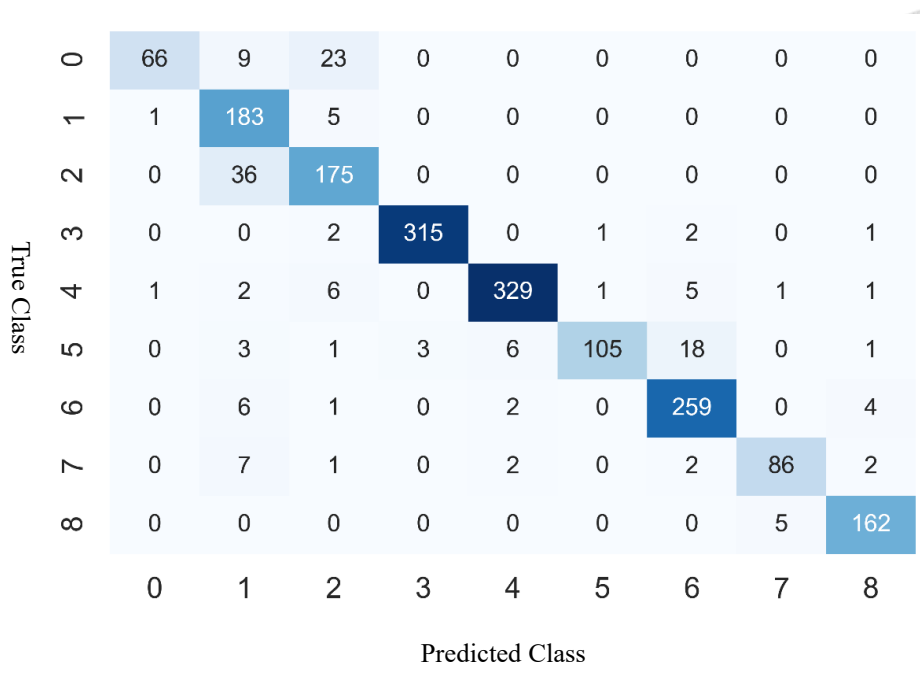
Because the enhanced model includes both WLE and NBI images and the internal test dataset includes 46 NBI images, the results without and with NBI images are compared in **Figure 14**, **Figure 15**, **Table 5**, and **Table 6**. For the base model of the internal test dataset without NBI images, the total accuracy is 91.25%, the macro-average precision is 91.80% and the weighted average precision is 92.04%. The macro-average recall is 88.36% and the weighted average recall is 91.25%. The macro-average F1-score is 89.44% and the weighted average F1-score is 91.23%. For the enhanced model of the internal test dataset without NBI images, the total accuracy is 93.05%, the macro-average precision is 93.91% and the weighted average precision is 93.35%. The macro-average recall is 90.89% and the weighted average recall is 93.05%. The macro-average F1-score is 92.01% and the weighted average F1-score is 92.92%. Compared with the performances of two models without NBI images, the enhanced model has significantly better overall accuracy and accuracy in classes 0, 2, and 8.

For the base model of the internal test dataset with NBI images, the accuracy is 90.46%, the macro-average precision is 90.82% and the weighted average precision is 90.99%. The macro-average recall is 87.97% and the weighted average recall is 90.46%.

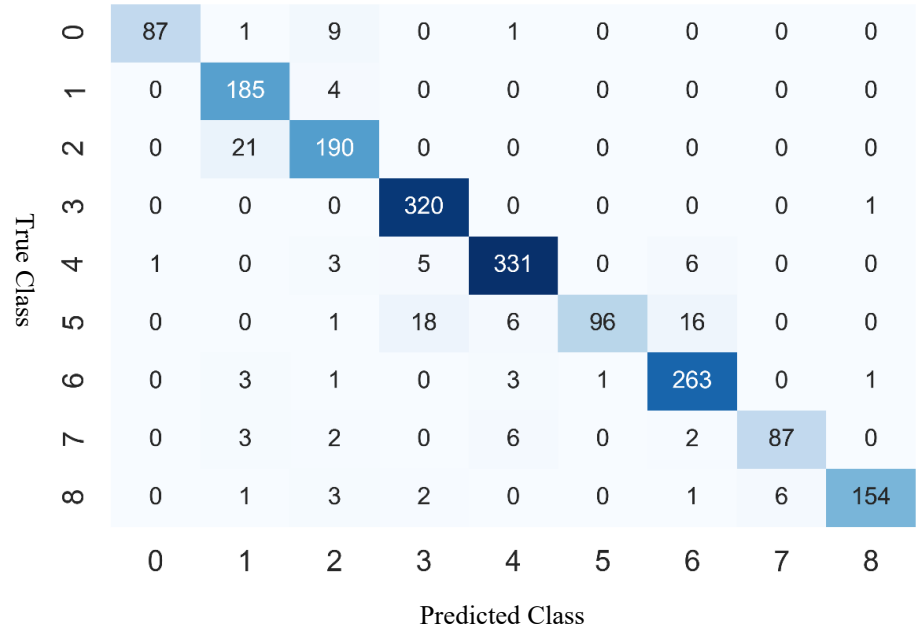
doi:10.6342/NTU202203401

The macro-average F1-score is 88.93% and the weighted average F1-score is 90.40%. For the enhanced model of the internal test dataset with NBI images, the accuracy is 92.74%, the macro-average precision is 93.70% and the weighted average precision is 93.12%, and the macro-average recall is 90.45% and the weighted average recall is 92.74% The macro-average F1-score is 91.64% and the weighted average F1-score is 92.62%. Compared with the performances of two models containing NBI images, the enhanced model has significantly better overall accuracy and accuracy in classes 0, 1, 2, and 8.

doi:10.6342/NTU202203401



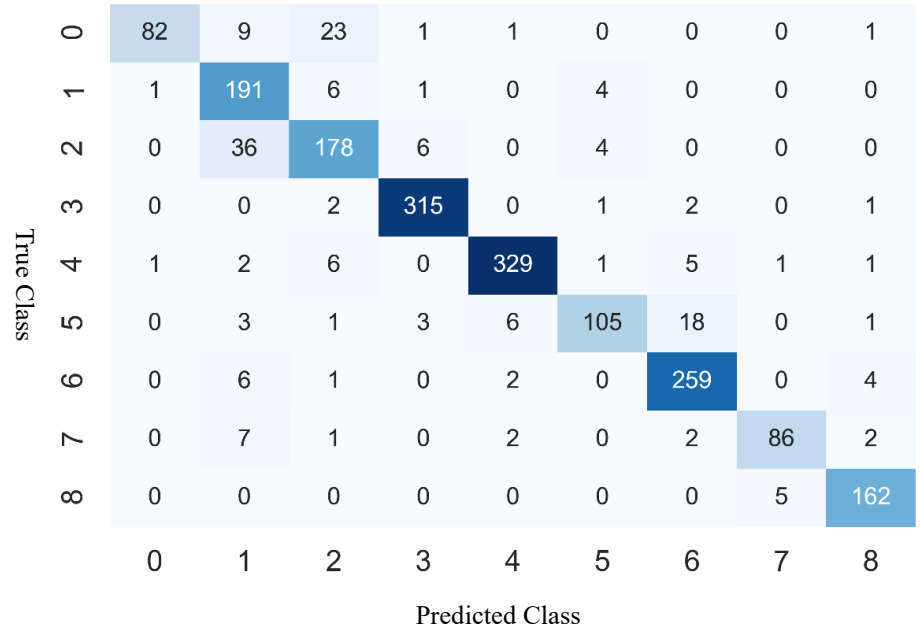
(a)



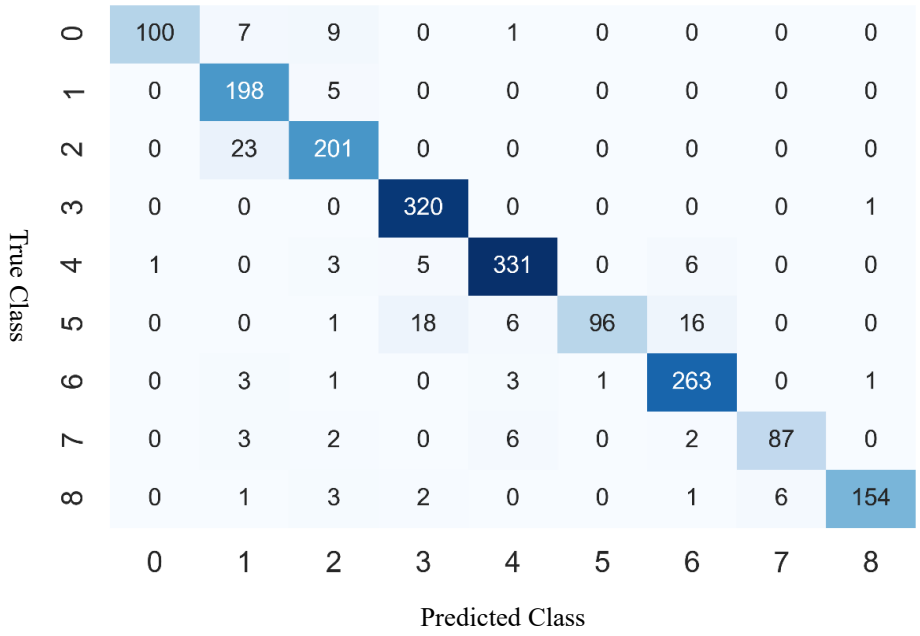
(b)

Figure 14. The confusion matrixes for test dataset #1 without NBI images for (a) base and (b) enhanced models.

doi:10.6342/NTU202203401



(a)



(b)

Figure 15. The confusion matrixes of the internal test dataset containing NBI images for (a) base and (b) enhanced models.

doi:10.6342/NTU202203401

Table 5. Performance comparison of the internal test dataset without NBI images.

	Metrics				Class								
		All	Macro-average	Weighted average	0	1	2	3	4	5	6	7	8
Base	Accuracy	91.25											
	Precision		91.80	92.04	97.06	74.39	81.78	99.06	97.05	98.13	90.56	93.48	94.74
	Recall		88.36	91.25	67.35	96.83	82.94	98.13	95.09	76.64	95.22	86.00	97.01
	F1-score		89.44	91.23	79.52	84.14	82.35	98.59	96.06	86.07	92.83	89.58	95.86
Enhanced	Accuracy	93.05											
	Precision		93.91	93.35	98.86	86.45	89.20	92.75	95.39	98.97	91.32	93.55	98.72
	Recall		90.89	93.05	88.78	97.88	90.05	99.69	95.66	70.07	96.69	87.00	92.22
	F1-score		92.01	92.92	93.55	91.81	89.62	96.10	95.53	82.05	93.93	90.16	95.36
p-value		<0.01*			<0.01*	0.48	<0.01*	0.07	0.82	0.11	0.22	1.00	0.01*

*A *p*-value less than 0.05 is considered statistically significant.

Table 6. Performance comparison of the internal test dataset containing NBI images.

	Metrics				Class								
		All	Macro-average	Weighted average	0	1	2	3	4	5	6	7	8
Base	Accuracy	90.46											
	Precision		90.82	90.99	97.62	75.20	81.65	96.63	96.76	91.30	90.56	93.48	94.19
	Recall		87.97	90.46	70.09	94.09	79.46	98.13	95.09	76.64	95.22	86.00	97.01
	F1-score		88.93	90.40	81.59	83.59	80.54	97.37	95.92	83.33	92.83	89.58	95.58
Enhanced	Accuracy	92.74			85.47	97.54	89.73	99.69	95.66	70.07	96.69	87.00	92.22
	Precision		93.70	93.12	99.01	84.26	89.33	92.75	95.39	98.97	91.32	93.55	98.72
	Recall		90.45	92.74	85.47	97.54	89.73	99.69	95.66	70.07	96.69	87.00	92.22
	F1-score		91.64	92.62	91.74	90.41	89.53	96.10	95.53	82.05	93.93	90.16	95.36
p-value		<0.01*			<0.01*	0.046*	<0.01*	0.07	0.82	0.11	0.22	1.00	0.01*

*A *p*-value less than 0.05 is considered statistically significant.

doi:10.6342/NTU202203401

The test image heatmaps for class 2 esophageal-gastric junction images generated by the Grad-CAM algorithm [18] for both models are shown in **Figure 16**. Both models are correctly classified as class 2, but the enhanced model has higher probabilities of 96.98% and 91.39% for the WLE and NBI images, respectively. The esophageal-gastric junction features are also illustrated as the highlighted spots in this figure. In **Figure 17**, the test result corresponding to the features extracted from the model is visualized by the t-distributed stochastic neighbor embedding (t-SNE) [41]. Because there is almost no intersection between the clusters, the model's classification accuracy is up to 96.64%.

3.3. Discussion

In this study, a high accuracy DL approach is developed for endoscopic anatomic classification. At first, manually labeled images are used to train the base model and then unlabeled images are added based on the trained base model with better performance than the original base model. The approach is feasible for endoscopy units with an inadequate workforce of data preparation [19, 20]. On the other hand, as datasets are the primary drivers for developing an excellent deep learning algorithm, the greater the number of images entered into the model for training, the better the model performed. Hence, it is a good research topic to study how many cases are needed to achieve the best performance.

doi:10.6342/NTU202203401

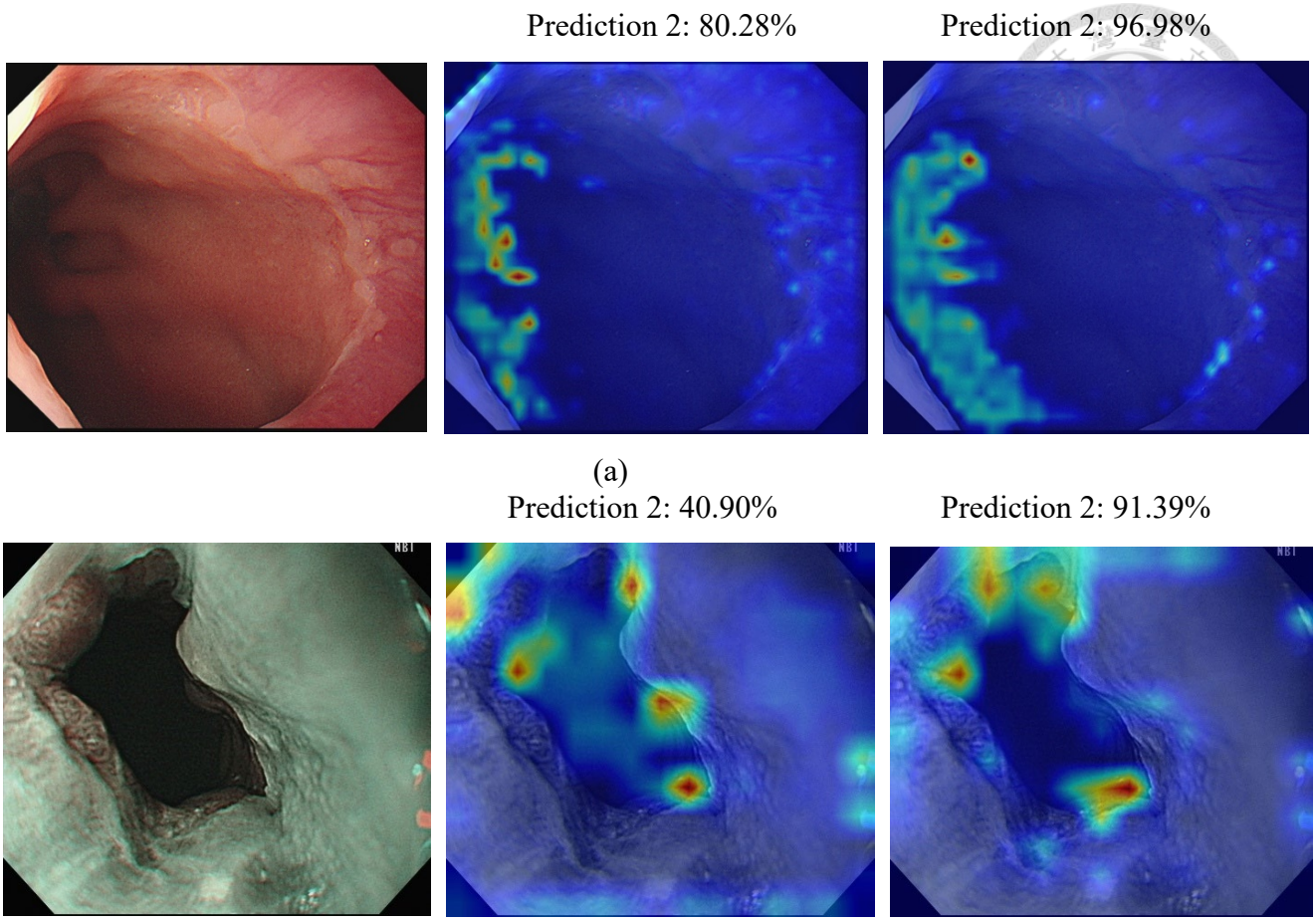


Figure 16. Heatmaps of the test images using base (middle column) and enhanced (right column) models for class 2 images. (a) WLE (b) NBI.

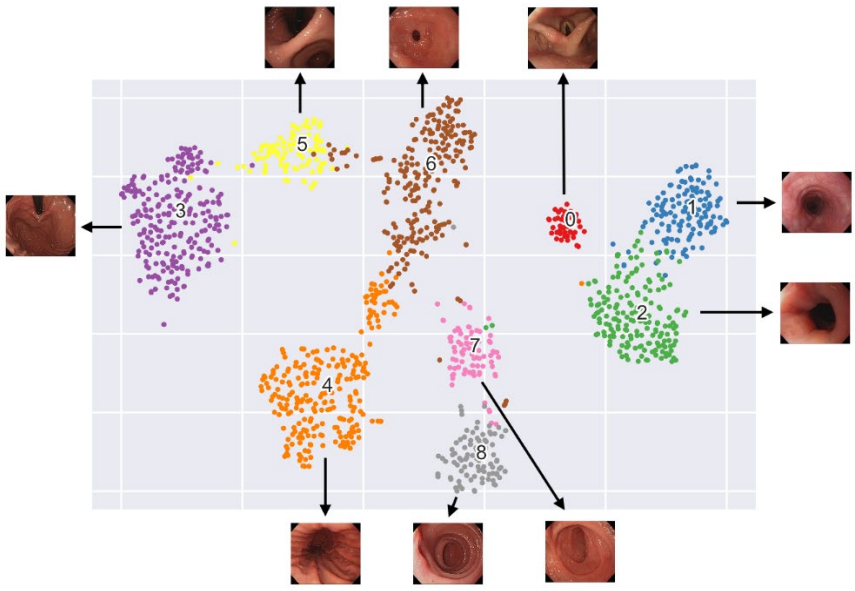


Figure 17. The t-distributed stochastic neighbor embedding (t-SNE) map for endoscopy anatomy DL classification.

doi:10.6342/NTU202203401

Chapter 4. Upper Endoscopy Quality Assessment

4.1. Materials & Methods

4.1.1 Patients and Data Preparation

The endoscopy records of upper endoscopy screening patients under general anesthesia from Jan to Jun 2020 at Changhua Christian Hospital are retrospectively reviewed. There are 14 board-certified endoscopists in the endoscopy center and 12 of them performed both upper endoscopy and screening colonoscopy. The Olympus EVIS LUCERA ELITE 260 or 290 series system is used to perform the endoscopy. Since 2015, the ADR result for quality assurance has been fed back to the endoscopist to improve their performance [45, 46].

4.1.2 Deep Learning Model and Endoscopy Image Processing

15,305 images are used to train Split-Attention Networks (ResNeSt) model [43] on an Intel i7-10700 personal computer with Windows 10 and Nvidia GeForce RTX^{TM} 3090 GPU. In this study, the retrieved original endoscopy images are 640 × 480 pixels and the images are cropped to size 469 × 410 pixels. The 15,723 images are tested by the trained ResNeSt model. Eight anatomical locations are used to classify all the endoscopy images of each procedure (L1: esophagus, L2: esophageal-gastric junction, L3: gastric cardia, fundus or retroflex view, L4: gastric body, L5: gastric angle, L6: gastric antrum, L7: duodenal 1st portion, L8: duodenal 2nd portion) by the DL model according to the ESGE classification. In this study, an additional location of the pharynx is defined as L0 and poor-quality images are defined as L9. The L8 location is further evaluated whether the whole or partial area of the duodenal ampulla is present or not [6]. Hence, in this study, 896 images of L8 location are used to train the second ResNeSt

doi:10.6342/NTU202203401

model to classify into L8A ones without ampulla (accuracy rate 92.41%, 134/145) and L8B ones with ampulla (accuracy rate 86.08%, 68/79) at the accuracy rate of 90.18%. The proposed automatic quality indicator system is shown in **Figure 18**.

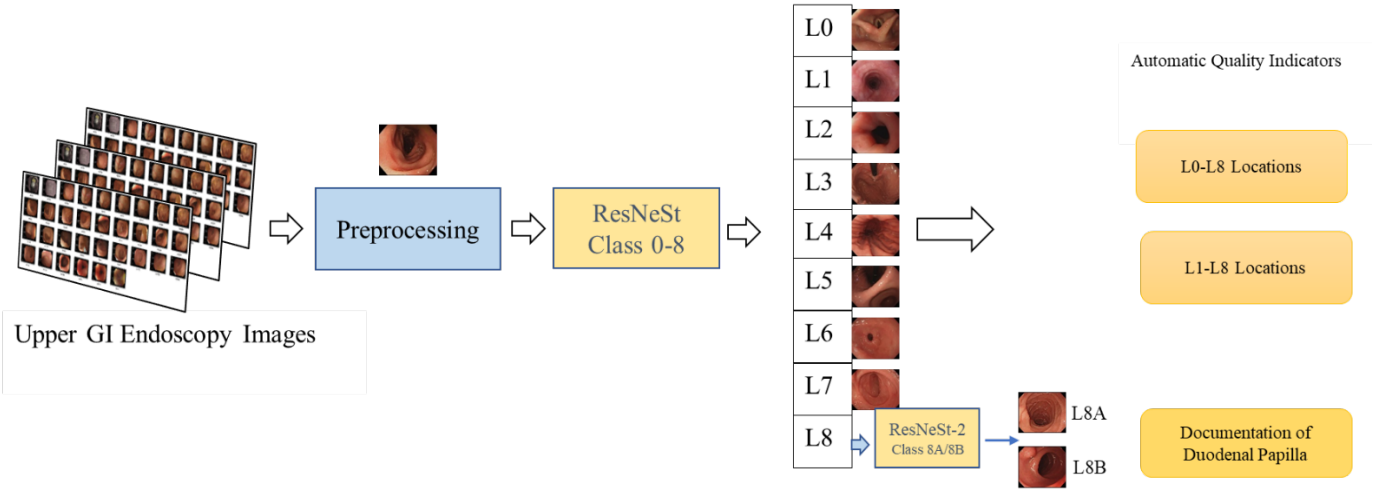


Figure 18. The proposed deep learning based automatic upper GI quality indicator system. L0: pharynx, L1: esophagus, L2: esophageal-gastric junction, L3: gastric retroflex view, L4: gastric body, L5: gastric angle, L6: gastric antrum, L7: duodenal 1st portion, L8: duodenal 2nd portion, L8A: duodenal 2nd portion without ampulla, L8B: duodenal 2nd portion with ampulla.

4.1.3 Complete Photodocumentation Rate Assessment

The DL model can compute the endoscopy image number at each anatomical location. Complete photodocumentation per location is defined as the presence of at least one image in a specific location. Complete photodocumentation of the entire procedure is defined as the presence of complete examination at all endoscopic anatomic locations. Complete ESGE 0-8 photodocumentation requires the presence of complete photodocumentation at location L0 to L8 and ESGE 1-8 requires the presence of complete photodocumentation at location L1 to L8.

4.1.4 Statistical Analyses

Continuous variables are presented as means (standard deviation) or median

doi:10.6342/NTU202203401

(interquartile range) according to the normality of the data distribution. Categorical variables are presented as proportions and compared by the chi-square test or the Fisher exact tests as appropriate. The Medcalc version 19.3 is used for all statistical analyses, with p values of <0.05 indicating statistical significance.

4.2. Results

4.2.1 Endoscopy Image Data Characteristics

Four hundred seventy-two upper endoscopy procedures with 15,723 images were performed during the study period. The median age of the population is 52 year-old, with 55% male patients. Totally, 98 (0.6%) poor-quality images (L9) are recognized. The median endoscopy image per examination is highest in the esophagus (L1), followed by the gastric body (L4). The pharynx (L0) and gastric angle (L5) had the lowest endoscopy image number at each procedure (**Table 7**). In **Table 7**, complete examination rate is 100% in the esophagus (L1) followed by L2 (98.9%), L6 (98.9%), L8 (98.9%), L3 (98.5%), L4 (97.9%), L7 (94.7%), L5 (85.8%) and L0 (66.7%). The overall complete photodocumentation rate is 78.0% from esophagus to the duodenum (location L1-L8) and 53.8% from pharynx to the duodenum (location L0-L8). The duodenal ampulla (L8B) photodocumentation is found in 287 procedures (60.8%).

doi:10.6342/NTU202203401

Table 7. Metrics of 9 anatomy site photodocumentation during EGD.

Endoscopy Location	Median Photo Number per Exam, *n*, IQR	Complete Examination Rate of the Location
0 Pharynx	1 (0-2)	315/472(66.7%)
1 Esophagus	7 (5-10)	472/472(100.0%)
2 EGJ	5 (3-6)	467/472(98.9%)
3 Retroflex view	3 (2-4)	465/472(98.5%)
4 Gastric body	5 (2-8)	462/472(97.9%)
5 Gastric angle	1 (1-2)	405/472(85.8%)
6 Gastric antrum	4 (2-6)	467/472(98.9%)
7 Duodenum 1st	2 (1-3)	447/472(94.7%)
8 Duodenum 2nd	2 (1-3)	467/472(98.9%)

4.2.2 Characteristics of Endoscopist Performance

In this study, 12 endoscopists perform 18 to 63 endoscopy procedures for analysis. **Table 8** summarizes the screening colonoscopy performance metrics after positive fecal immunochemical tests from 2018 to 2020**.** The ADR rate ranges from 43.57% to 67.74%, and the polyp detection rate (PDR) ranges from 55.37% to 74.19%. The captured endoscopy image number per procedure by the individual endoscopist ranges from 19.71 to 51.55 images that are all above the recommended at least 10 photos by the ESGE [47]. The complete photodocumentation rate ranges from 0% to 95.0% for location L0-L8 and 55.6% to 95.0% for location L1-L8.

doi:10.6342/NTU202203401

Table 8. Performance of endoscopists on screening colonoscopy in 2018-2020.

Endoscopist	Screening Colonoscopy Procedures in 2018-2020	ADR	PDR	EGD Procedures in this Analysis	Mean EGD Images per Procedure (Mean, SD)
A	22	67.74%	74.19%	18	22.56, 0.28
B	151	67.46%	84.13%	20	51.55, 0.12
C	202	62.23%	70.59%	63	34.65, 0.17
D	271	60.96%	72.80%	51	30.71, 0.18
E	131	60.33%	73.91%	50	30.32, 0.21
F	318	58.81%	73.46%	37	46.89, 0.21
G	134	57.64%	63.76%	42	24.14, 0.25
H	132	55.38%	71.79%	54	48.53, 0.16
I	91	53.90%	69.50%	28	33.57, 0.23
J	57	52.38%	60.00%	47	21.98, 0.18
K	36	46.15%	75.00%	28	32.50, 0.26
L	346	43.57%	55.37%	34	19.71, 0.21

ADR: Adenoma detection rate; PDR: Polyp detection rate.
Note: The ADR and PDR rates for screening colonoscopy after a positive fecal occult blood test are audited annually in the unit.

4.2.3 Comparison of Endoscopist Colonoscopy Performance and Completeness of Upper Endoscopy Examination Photodocumentation

All the endoscopists in the unit have an ADR rate above the national average ADR rate of 39.5% [48]. The overall ADR is 53.6% in the recent multicenter Asia-Pacific study [49]; therefore, a goal is set for achieving ADR above 55% for further quality improvement. The endoscopists are divided into high ADR (Endoscopist A-H) and low ADR (Endoscopist I-L) groups according to their ADR detection rate at a 55% cut-off for subsequent analysis in **Figure 19**. High ADR endoscopists have higher complete photodocumentation rates at locations L0-L8 (60%) and L1-L8 (83%) compared with low ADR endoscopists. The duodenal ampulla photodocumentation is similar among

doi:10.6342/NTU202203401

these two endoscopist groups.

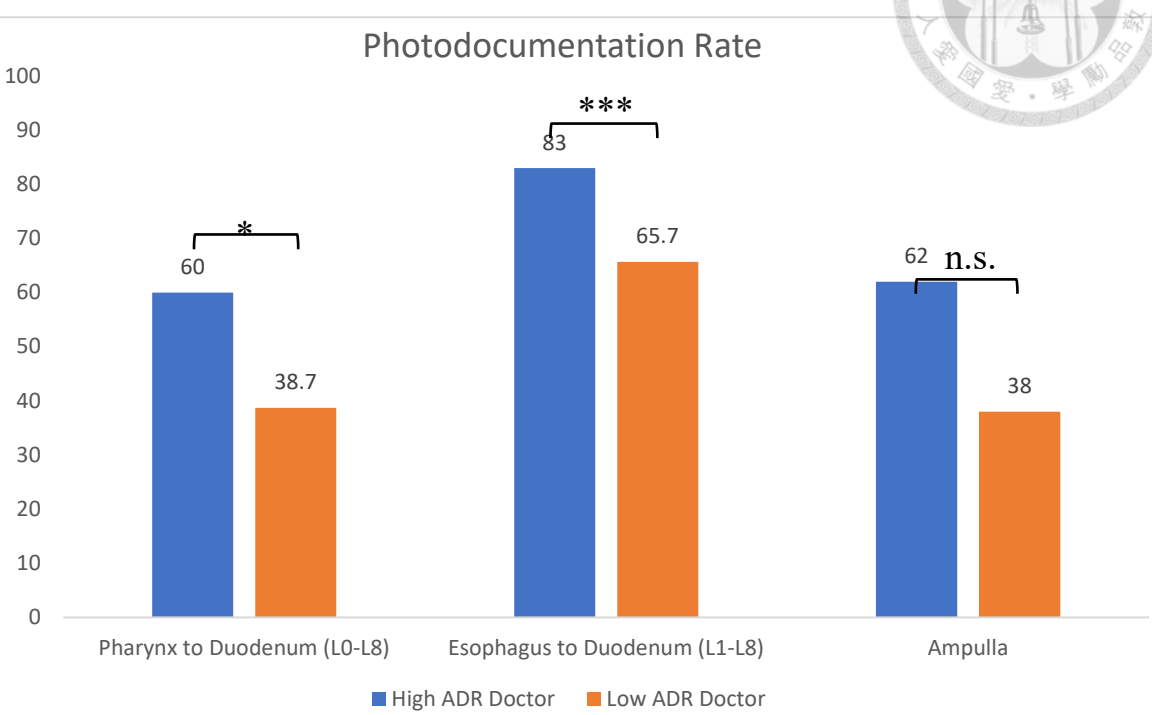


Figure 19. Comparison of endoscopist performance on colonoscopy and the completeness of photodocumentation during upper endoscopy. The rate of complete photodocumentation is significantly higher for doctors with high adenoma detection rates (ADR) from pharynx to duodenum and from esophagus to duodenum (*** $p < 0.0001$). The ampulla photodocumentation rate is similar.

In **Table 9**, the differences in the photodocumentation completeness at each endoscopy location are compared. High ADR endoscopists have higher complete photodocumentation rates at L0, L5, L6, and L7 than low ADR endoscopists. **Table 8** lists the median number of images captured by individual endoscopists.

doi:10.6342/NTU202203401

Table 9. Comparison of endoscopist performance on colonoscopy and the completeness of photodocumentation at each endoscopy location.

Location, number of complete examination/total examination	High ADR Doctors	Low ADR Doctors	*p*-value
L0	233/335(69.6%)	82/137(59.9%)	*P* = 0.0426*
L1	335/335(100%)	137/137(100%)	*P* =1
L2	333/335(99.4%)	134/137(97.8%)	*P* = 0.1254
L3	332/335(99.1%)	133/137(97.1%)	*P* = 0.0990
L4	328/335(97.9%)	134/137(97.8%)	*P* = 0.9453
L5	295/335(88.1%)	110/137(80.3%)	*P* = 0.0283*
L6	334/335(99.7%)	133/137(97.1%)	*P* = 0.0117*
L7	325/335(97.0%)	122/137(89.1%)	*P* = 0.0005*
L8	330/335(98.5%)	137/137(100%)	*P* = 0.1510

* A *p*-value less than 0.05 is considered statistically significant.

4.3. Discussion

This study is the first report for a thorough check-up quality of endoscopy photodocumentation using a robust endoscopic anatomic classification DL model in this study. The endoscopists with better colonoscopy performance have a higher complete photodocumentation rate. This finding can support the photodocumentation rate as an endoscopy quality indicator [50, 51], as it became feasible in the era of artificial intelligence [52, 53].

Accurate upper GI endoscopy photodocumentation was recommended as one of the top 6 quality indicators in the 2019 survey [16]. A performance target of 98% [7] is set by the American Society of Gastrointestinal Endoscopy (ASGE), and a minimal photo-documentation rate of 90% is recommended by ESGE [47] via yearly sampling of 100 consecutive UGI endoscopies reports [46, 47]. However, automated information capture is lacking in most units, and auditing this quality indicator is impossible [16, 54].

Based on captured endoscopy images, we utilize the DL model for this endoscopy

doi:10.6342/NTU202203401

automatic quality auditing. The proposed system could count the classified endoscopy photos per endoscopy procedure and the quality auditing results are generated immediately. AI is a promising tool for improving medical care [53, 55]. The current AI development in endoscopy mainly focuses on lesion detection or classification, such as colon polyps, esophageal cancer, or gastric cancers [56]. For endoscopy quality control, there are few AI studies as photodocumentation as in our study. CNN for the anatomical classification of still images into four anatomy sites was firstly proposed by Takiyama et al. [57]. Other groups [58, 59] and our reported DL model found such an AI approach could achieve a high accuracy rate for photodocumentation quality control in esophagogastroduodenoscopy. This study is the first to implement its use for reports generated in daily practice, and the ampulla photodocumentation rate was previously not reported.

Although adequate photodocumentation is generally accepted, no studies link the photodocumentation rate to improving patient outcomes or diagnostic yield [47, 54, 60]. Longer withdrawal time is associated with more high-risk gastric lesions during endoscopy, as shown in previous studies [5, 6, 61, 62]. The more images are taken, the higher the complete photodocumentation rate should theoretically correlate with a longer procedure time because it takes time for the endoscopist to capture the clear endoscopic image. In this study, depending on each endoscopist, there is a variation in the average image number of an EGD procedure. Despite theoretically the more endoscopy images were taken, the high photodocumentation rate can be achieved. We do not find a correlation between the endoscopist performance with the captured image number. One reason is that the endoscopist might take several images at one anatomical site, such as, in this study, the esophagus with the highest image number; in contrast, other anatomical sites are neglected. Hence, the photo number may not be a good

doi:10.6342/NTU202203401

endoscopy quality indicator; instead, the well-known ADR rate [48, 63-65] is used as an endoscopist performance indicator to explore its relationship to the photodocumentation quality in this study. We can find that high ADR endoscopists have a significantly higher complete photodocumentation rate, especially in specific locations.

There are some advantages for endoscopy units in the development of AI systems for auditing quality indicators. First, using the AI system as in this study may allow real-time quality metric feedback to endoscopists to improve examination quality, as we have learned from colonoscopy quality improvement [45]. Secondly, the system may explore the potential endoscopist blind spots and be used as an evaluation tool for the trainee endoscopist competence. For instance, in the study, all endoscopists have a lower number of photos taken from the gastric angle (L5). Understanding this might help the endoscopist improve their performing endoscopy to increase the upper GI diagnostic yield in the future. Third, to upgrade the existing endoscopy operation system, the AI system may not require additional costs. The DL model with an accelerated approach [14] will be feasible and applicable to any endoscopy unit.

In this study, there are some limitations. Firstly, this study retrospectively analyzes the real-world endoscopy documentation rate using a DL model. This model is trained based on the Olympus endoscopy system and diagnostic endoscopy with general anesthesia. Further studies are needed before its broad generalization to other different endoscopy systems and other units. Second, no generally accepted outcome indicator exists for subsequent risk of esophageal or gastric cancers during screening upper GI endoscopy, and a limited case number is enrolled in this study. Hence, we cannot use the gastric neoplasia detection rate as the outcome target due to its low incidence in Taiwan. Third, we only assess the AI model performance in the eight anatomy location

doi:10.6342/NTU202203401

proposed by the ESGE and BSG [8, 66]. In the future, there is a further development need to extend the DL model's ability to recognize the more complex anatomical locations, such as 28 anatomy locations proposed by the world endoscopy association [54, 67].

doi:10.6342/NTU202203401

Chapter 5. Colonoscopy Quality Assessment

5.1. Materials & Methods

5.1.1 Patients and Data Preparation

Three hundred patients who underwent diagnostic colonoscopy examination were retrospectively reviewed at Changhua Christian Hospital from January 2020 to December 2020. The Olympus 260 or 290 series system is used for all the endoscopic procedures. Two expert endoscopists with 15 years of experience in therapeutic endoscopy assess the retrieved images for subsequent analysis. A discussion with another independent endoscopist resolves a disagreement between reviewers. Five key intraprocedural quality indicators, including withdraw time, bowel preparation, cecal intubation, polypectomy or biopsy, and rectal retroflexion proposed by the UK quality assurance standards [11] are selected as a possible DL measurable target.

Initially, the retrieved endoscopy images of each colonoscopy procedure are classified into eight classes: 0, patient information images; 1, endoscopy information images; 2, the anal rectal region images; 3, poor-quality images (such as out-of-focus images and those with excessive light reflection); 4, instrument images, such as polypectomy snare and biopsy forceps; 5, colon images; 6, cecum images; and 7, rectal retroflexion images. Then, class 5 images are further classified according to the bowel preparation level: 5A, Boston Bowel Preparation Scale (BBPS) scores of 2-3 and 5B, BBPS scores of 0-1 [68].

Because, in real-world endoscopy reports, the numbers of those with a BBPS score of 0-1 cecal, rectal, and retroflexion images are relatively low, these three-class images from the Hyper-Kvasir open dataset [18] are included for model training.

doi:10.6342/NTU202203401

5.1.2 Preparation of Endoscopy Images and Model Training

The 469 × 410 regions of interest are cropped from the original 640 × 480 pixels images to contain colonoscopy images alone. The 469 × 410 images are rescaled into 224 × 224 images due to the pre-trained ImageNet model is used for transfer learning to reduce training time. **Figure 20** shows the number of endoscopy images in the trained model.

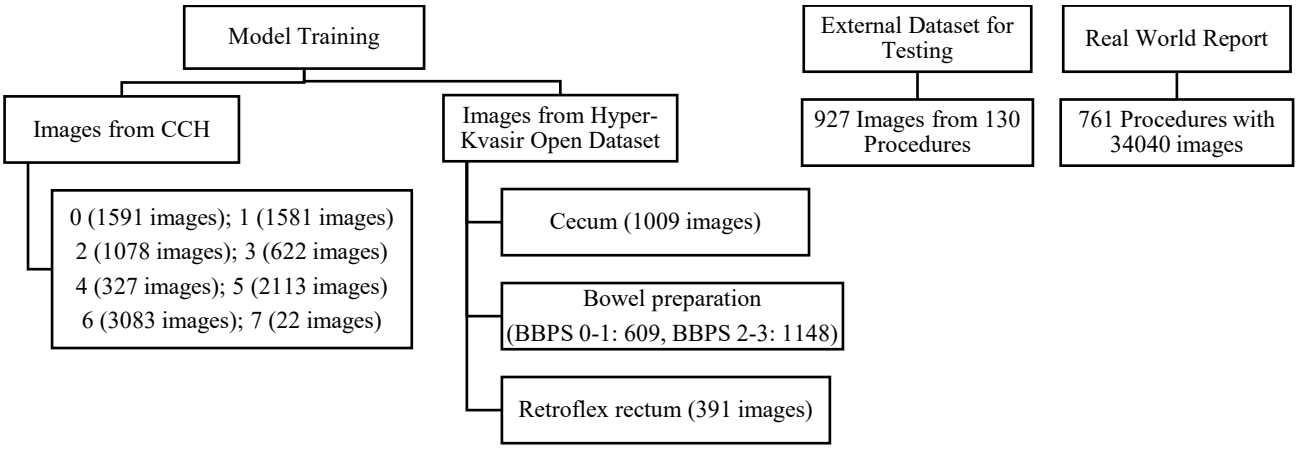


Figure 20. Endoscopy image number in the trained model, external test dataset, and real-world report.

The DL model is trained by the Split-Attention Networks (ResNeSt) [43]. In this study, ResNeSt with 50 layers pretrained with ImageNet [17] is used. **Figure 21** depicts the manual quality indicators of the electronic endoscopy reporting system and the automatic proposed DL-based quality indicators. *Accuracy*, *Precision*, *Recall*, and *F1-score* of the trained model are calculated. For multi-class classification, two average methods are used to obtain the different measures, macro-average and micro-average [69]. Additionally, both Grad-CAM [39] and anchor image [40] are applied to assess the model prediction visually and interpretably. The test result is also visualized by the t-SNE [41].

doi:10.6342/NTU202203401

5.1.3 Deep Learning Model Performance with External Testing Image Data

As shown in **Figure 20**, a 130-patient independent dataset is utilized as external testing for the developed DL model.

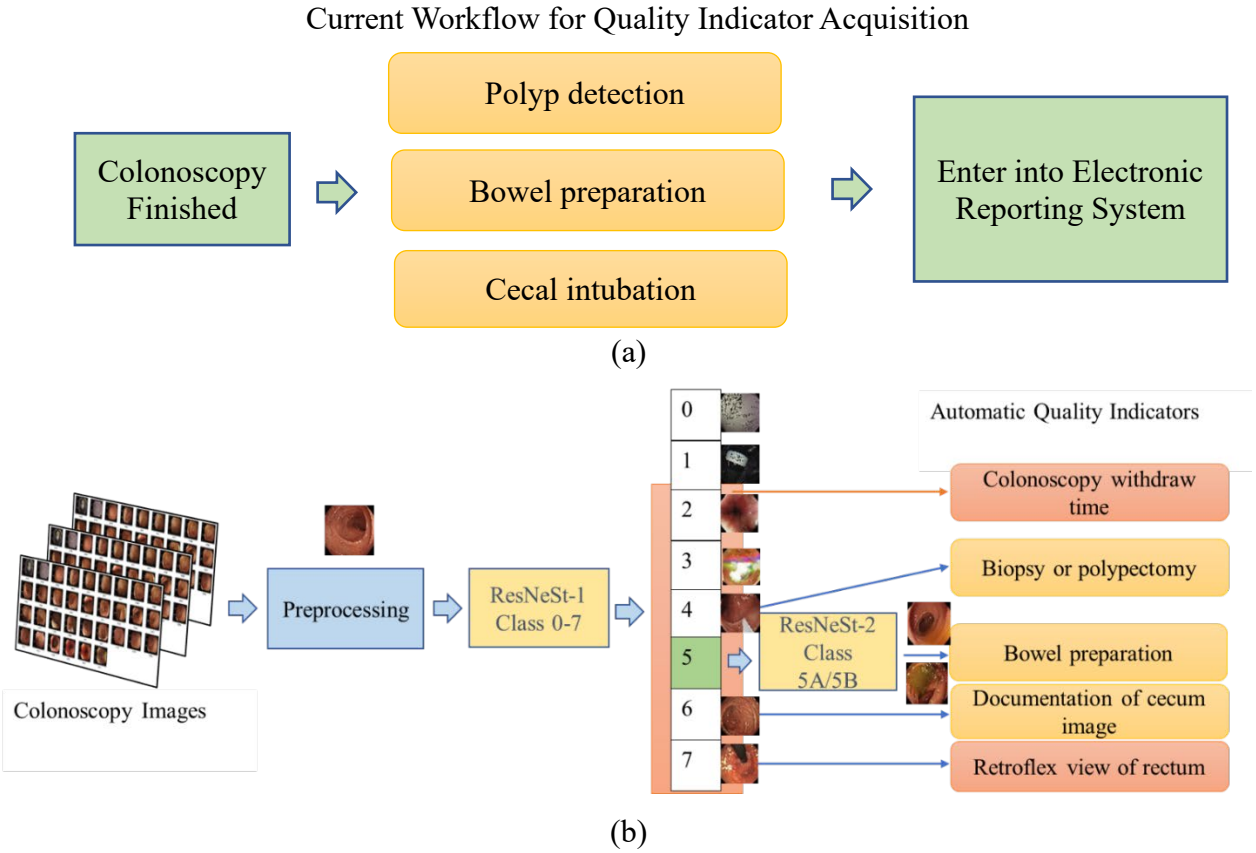
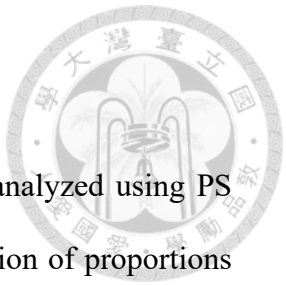


Figure 21. (a) The current workflow in the endoscopy quality indicator report. (b) Automatic quality indicators in the proposed deep learning-based system.

5.1.4 DL model Performance with Real-world Colonoscopy Reports

Real-world colonoscopy images and reports from May 2020 to December 2020 are manually reviewed, as shown in **Figure 20**. Two endoscopists checked the captured colonoscopy images for the performance comparison of the electronic report, our developed DL model, and captured images.

doi:10.6342/NTU202203401



5.1.5 Statistical Analysis

The extracted data are organized using Microsoft Excel and analyzed using PS IMAGO Pro 7. Kruskal-Wallis test is used to compare the distribution of proportions of good preparation images per procedure at different bowel preparation grades with Dunn's post hoc tests. The strength of association between withdrawal times from the record and from our DL model is measured by Spearman's correlation analysis.

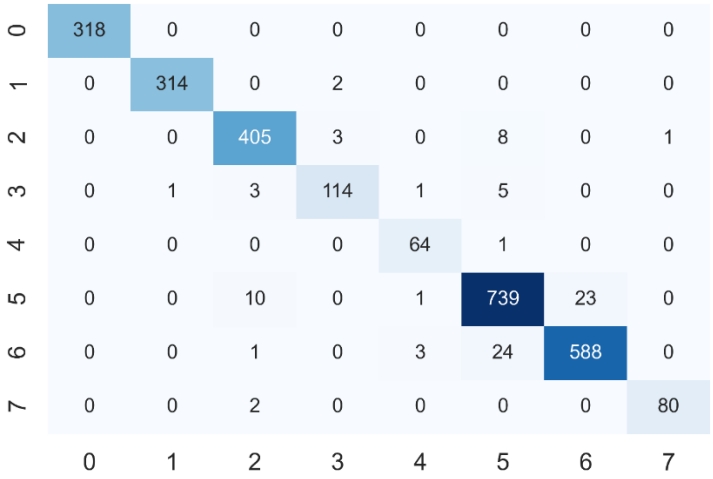
5.2. Results

5.2.1 Performance of Trained Model

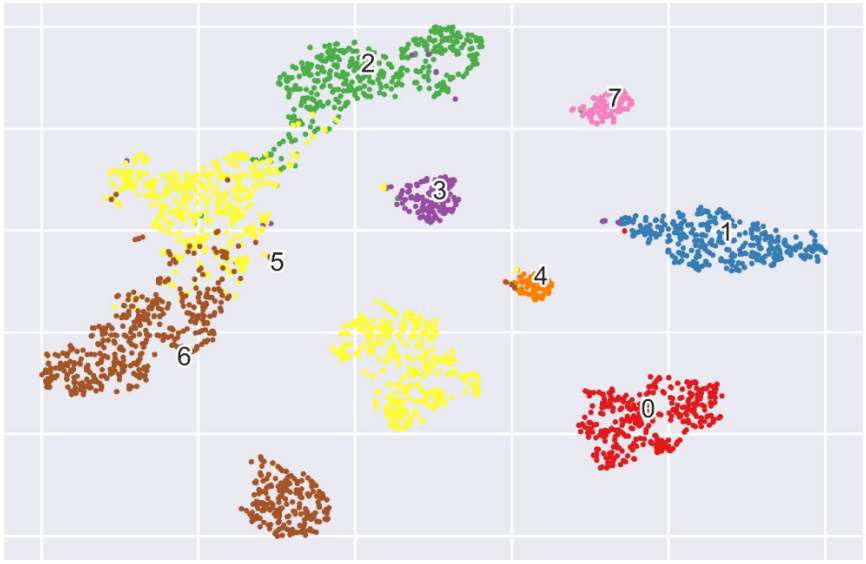
13,574 images are divided into the training, validation, and testing subsets to establish the model, as shown in **Figure 20**. The accuracies and losses of training and validation, confusion matrix, and t-SNE map are shown in **Figure 22**. The model performance is listed in **Table 10**. The overall accuracy is 96.72%, and the macro-average precision is 96.82% and the weighted average precision is 96.73%. Further, the macro-average recall is 96.94% and the weighted average recall is 96.72%. The macro-average F1-score is 96.86% and the weighted average F1-score is 96.72%. **Figure 23** depicts the Grad-CAM, anchor, and t-SNE images.

doi:10.6342/NTU202203401

(a)

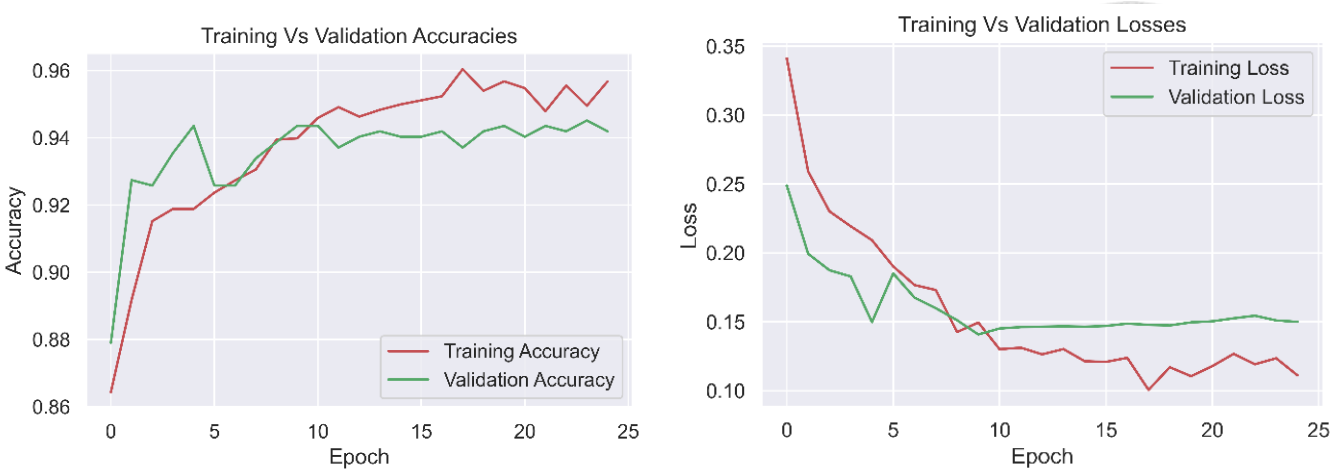


(b)

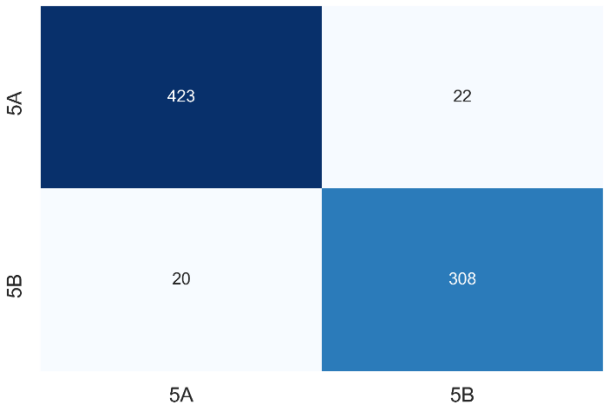


(c)

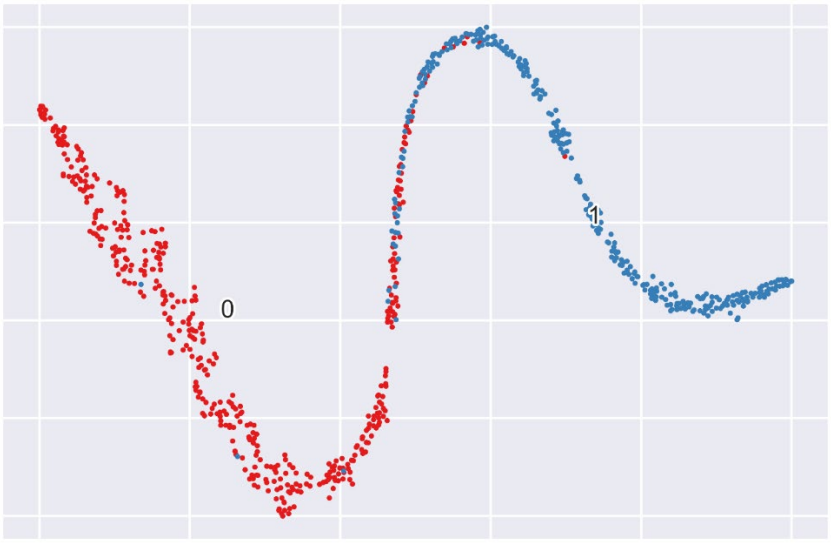
doi:10.6342/NTU202203401



(d)



(e)



5A: BBPS score 2–3, 5B: BBPS score 0–1

(f)

Figure 22. The training and validation accuracies and losses for classes 0–7 (a) and classes 5A and 5B (d). The confusion matrix of testing for classes 0–7 (b) and 5A and 5B (e). The t-SNE map of our deep learning model for classes 0–7 (c) and 5A and 5B (f).

doi:10.6342/NTU202203401

Table 10. Performance for the trained model.

Metrics				Class								
Testing	All	Macro average	Weighted average	0	1	2	3	4	5	6	7	5A/5B
ACCURACY	96.72											94.57
PRECISION		96.82	96.73	100.00	99.68	96.20	95.80	92.75	95.11	96.24	98.77	93.33
RECALL		96.94	96.72	100.00	99.37	97.12	91.94	98.46	95.60	95.45	97.56	93.90
F1-SCORE		96.86	96.72	100.00	99.52	96.66	93.83	95.52	95.35	95.84	98.16	93.62

5.2.2 Model Performances of External Test Dataset

An independent dataset is used to evaluate the trained model. The overall accuracy is 94.71%. The macro-average precision is 96.14% and the weighted average precision is 94.79% and the macro-average recall is 94.06% and the weighted average recall is 94.71%. The macro-average F1-score is 94.98% and the weighted average F1-score is 94.65%, as shown in **Table 11.** Performance for the external test dataset model.

Table 11. Performance for the external test dataset model.

Metrics				Class								
Testing	All	Macro average	Weighted average	0	1	2	3	4	5	6	7	5A/5B
ACCURACY	94.71											94.62
PRECISION		96.14	94.79	100.00	100.00	95.10	90.91	100.00	88.34	94.79	100.00	99.07
RECALL		94.06	94.71	100.00	100.00	100.00	77.78	92.93	93.36	97.09	91.30	90.60
F1-SCORE		94.98	94.65	100.00	100.00	97.49	83.83	96.34	90.78	95.92	95.45	94.64

doi:10.6342/NTU202203401

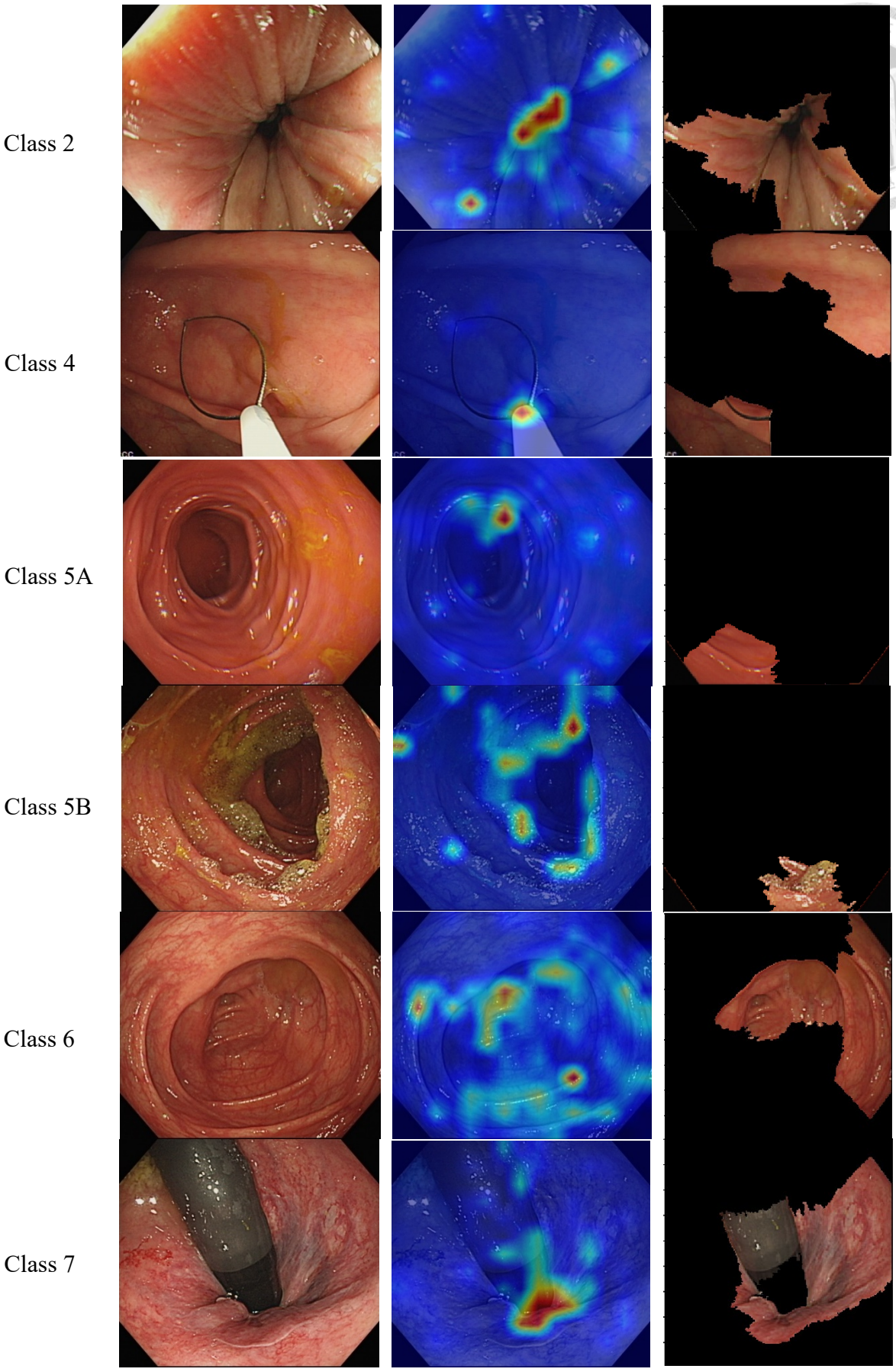


Figure 23. (a) Original images. (b) Grad-CAM images. (c) Anchor images.

doi:10.6342/NTU202203401

5.2.3 DL Model in Assessing Real-world Colonoscopy Images and Reports

Seven hundred sixty-one electronic endoscopy reports and their captured images (mean: 44.7 per procedure) are included in the analysis. The median patient age is 57 years, and 54.7% are males. The CIR is 99.1%, the polypectomy rate is 36.8, and bowel preparation is graded as excellent (29.3%), good (63.9%), fair (5.9%), and poor (0.9%).

Based on the reviewed image, the physician's electronic report CIR accuracy is 99.34% and the CIR accuracy of the DL system is 98.95%. **Table 12** depicts the CIR based on the electronic report and the proposed DL system with a human assessment. The DL system accuracy for accessing the polypectomy rate is 93.56%, based on an electronic report. The polypectomy agreement rate based on the electronic reports and DL system is 0.87 (95% confidence interval: 0.83-0.90). According to the electronic reports and the proposed DL system, the biopsy and polypectomy results are listed in **Table 13**.

Table 12. Cecum intubation comparison of the electronic report and proposed DL systems with the photo documentation of cecal image

		Photo Documentation of Cecal Image	
		No	Yes
Cecum intubation by electronic report	No	6	1*
	Yes	4**	750
Cecum intubation by DL system	No	3	1
	Yes	7	750

* Typing error of exam result

** No photodocumenetation of cecal image

doi:10.6342/NTU202203401

Table 13. Comparison of biopsy or polypectomy at the electronic report with proposed DL system.

Biopsy or Polypectomy at Electronic Report	Instrument detected by DL system	
	No	Yes
No	448	32
Yes	16*	265

* 9 of the 16 procedures have no image evidence of biopsy or polypectomy after review of the procedure image.

According to different bowel preparation levels (excellent, good, fair, poor) from the electronic report, the images obtained per procedure in terms of the proportion of well-prepared images [5A / (5A + 5B)] are compared in **Figure 24**.

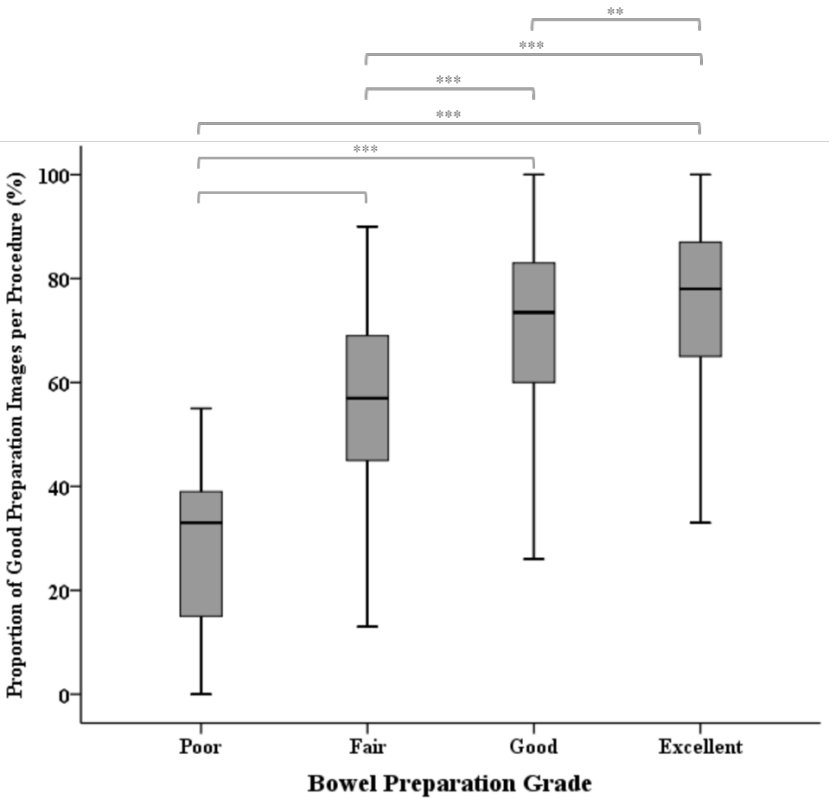


Figure 24. Median proportion of good preparation images per procedure with the levels of bowel preparation based on the electronic report. $^{**}p<0.01$, $^{****}p<0.001$

The proposed DL model could classify each captured image with its timestamp. The DL-based withdrawal time, which is the time interval between the first cecal image and the last colon image during each procedure, and that entered by the physician in the

doi:10.6342/NTU202203401

electronic report are compared with a good correlation (correlation coefficient: 0.959, $p < 0.0001$) in **Figure 25**.

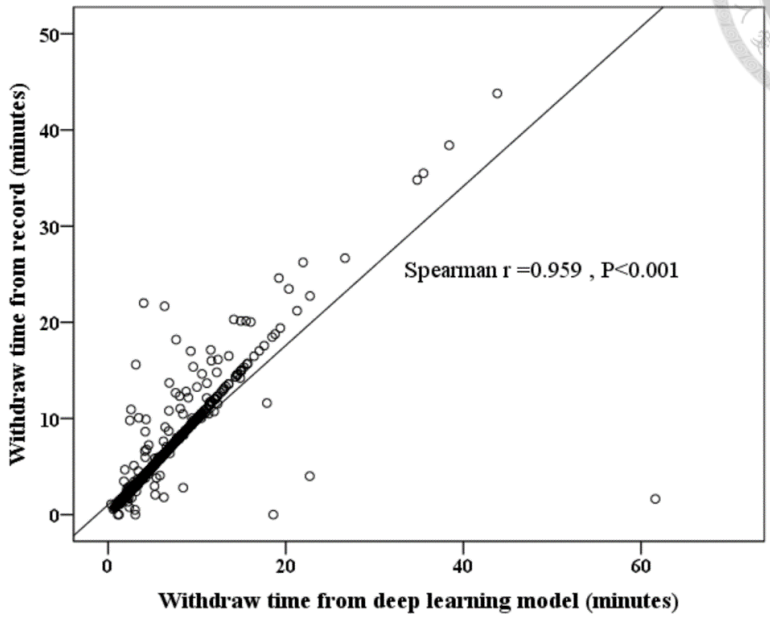


Figure 25. Withdrawal times are calculated using the DL and those entered by the physician. Correlation coefficient: $r = 0.959$, $p < 0.0001$.

5.3. Discussion

The current study develops a high-accuracy DL-based model for auditing colonoscopy quality. Colorectal cancer is the fourth leading cause of death worldwide, and colonoscopy can effectively reduce the incidence and mortality [70]. However, colorectal cancer screening is based on high-quality colonoscopy to identify lesions [11, 71]. A novel image-based system is proposed to assess colonoscopy quality. The accuracy of identifying cecal, instrument, and retroflexion images is 95.45%, 98.46%, and 97.56%, respectively, and that of differentiating different bowel preparation statuses is 94.57%. The feasibility of the DL model is shown in this study. The proposal DL model could automatically extract the colonoscopy quality indicators based only on captured colonoscopy images.

Yao et al. [72] recently reported the endoscopy quality control systems, which

doi:10.6342/NTU202203401

require the integration of endoscopic images, endoscopic descriptions, and pathology reports, to improve endoscopist performance. CNN was used to identify cecal images in their studies. The entire captured endoscopy images rather than the single image are analyzed for quality evaluation in our novel model. Also, the proposed model can identify invalid images with a high accuracy rate, such as patient information, scope type, and poor quality images. The accuracy rate of rectal retroflexion identification could achieve 97.6%. The identifying cecal images rate is 95.59%, which is better than a previous report [73]. For cecal photodocumentation, the DL accuracy rate is 98.95% compared with the electronic report. One entering error was found with confirmed cecal photodocumentation and no cecal intubation was found in the electronic report. Also, there are four procedures in electronic report cecal intubation documentation that lack cecal photodocumentation during the reviewing process.

ADR is a crucial colonoscopy quality indicator related to interval colorectal cancer [13, 71, 74]. However, a combination of information associating pathology and endoscopy reports is required to calculate this ADR indicator [75]. Polyp detection rate had been considered a surrogate marker of ADR [10, 76, 77] and could be obtained from the endoscopy electronic reporting system after each procedure. Although DL had been applied to detecting polyps during colonoscopy procedures to improve ADR [78, 79], the false-positive polyp results from the captured images [77] may overestimate ADR. Because endoscopists will obtain high-quality images during polyp therapy, identifying instruments such as biopsy forceps and polypectomy snare could be a reliable marker for biopsy or polypectomy. The instrument detection accuracy rate is 92.93% and 93.56% for external test image data and real-world data, respectively. Also, a good correlation between colon withdrawal time can be obtained by using the DL model. Hence, the proposed AI system may improve the endoscopy practice workflow.

doi:10.6342/NTU202203401

As the primary procedure indicator is automatically calculated, endoscopists can focus on carefully examining the colonic mucosa and capturing adequate images.

The current study has several limitations. First, it uses images from a single endoscopy system; therefore, whether the system works on other endoscopy systems remains unknown. Second, only screening endoscopy images are included and patients with inflammatory bowel disease or surgically resected colon are not included in the model training. Third, the poor-quality image classification (class 3) is subjective. Due to no standard for evaluating the captured image quality, an ambiguous classification between classes 3 and 5 might occur. Fourth, in the current study, bowel preparation is only divided into two classes rather than four classes based on the BPS score. Also, the proposed DL model could not assess bowel preparation status at different bowel segments, which requires extra information on colonoscope movement [80].

doi:10.6342/NTU202203401

Chapter 6. Future Works

In the future, other advanced machine learning methods will be further investigated to improve the performance of the proposed DL model. In natural language processing (NLP), the Transformer by Vaswani et al. in 2017 for machine translation [19] has become the state-of-the-art method in many NLP tasks. The Bidirectional Encoder Representations from Transformers (BERT) [20] is designed to pre-train deep bidirectional representations from the unlabeled text. As a result, the pre-trained BERT model can be fine-tuned to create state-of-the-art models for a wide range of tasks. The Vision Transformer (ViT) [21] has recently demonstrated promising results in image classification. On image classification, ViT achieves an impressive speed-accuracy tradeoff compared to convolutional networks. Recently, based on the ViT architecture, the Shifted Windows (Swin) Transformer [22] was proposed using the shifted window partitioning technique. Recently, Facebook AI Research enhanced the standard ConvNet (ResNet) as ConvNeXts [81] according to the design of the Swin Transformer, and the ConvNeXt can outperform the Swin Transformer. Hence, we can try to implement the Swin Transformer and ConvNeXts for the GI colonoscopy image classification.

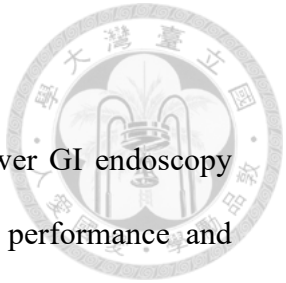
The proposed accelerated data preparation approach is similar to semi-supervised learning [82]. In semi-supervised learning, only a small amount of labeled data is used to train the model, and then the trained model is used to predict the unlabeled data. Then, the model trained with labeled data can be retrained by using labeled data and predicted unlabeled data. In this study, the physicians recheck the predicted unlabeled data. Recently, self-supervised learning [23-25] has been proposed for automatically generating labels without human intervention for large amounts of data. Contrastive

doi:10.6342/NTU202203401

learning is a self-supervised learning technique to find similar and dissimilar information from a dataset. For example, Momentum Contrast (MoCo) [83], Pretext-Invariant Representation Learning (PIRL), Google's SimCLR [84, 85], and WAV2VEC [86] are contrastive self-supervised learning techniques. For GI endoscopy with large amounts of images, self-supervised learning might help reduce the physicians' labeling works.

doi:10.6342/NTU202203401

Chapter 7. Conclusion



This study aims to resolve the unmet needs of upper and lower GI endoscopy quality indicator systems to improve the endoscopy procedure's performance and provide better patient care. A good endoscopy quality indicator system is based on a successful GI image classification system. Hence, we have proposed a DL model for classifying upper GI endoscopy images according to the anatomical locations with an accelerated data preparation approach in the first part of this study. At first, a base model is trained from a smaller data set labeled by the endoscopists in order to save their labeling time. Then, the base model is used to classify another unlabeled data set, and the classification results are reviewed and revised by the endoscopists. Combining the first smaller data set and the revised data set, an enhanced model is retrained to improve the classification performance. Because the enhanced model is also validated with another dataset with high accuracy, the model could be used in practical clinical uses. Also, such an accelerated data preparation approach can help endoscopy units quickly build automatic upper GI endoscopy image classification systems. The proposed accelerated data preparation approach is also applied in the third part of this study, the DL-based colonoscopy quality model, to reduce the radiologist's labeling time. The picture archiving and communication system (PACS) could combine the proposed AI-based classification system for endoscopists to extract images at a specific location for reviewing.

In the second part of this study, we report the first report for a thorough check-up of endoscopy images for the photodocumentation quality with a robust endoscopic anatomic classification DL model in the first part of this study. From the experiments, better performance endoscopists have a higher complete photodocumentation rate. This

doi:10.6342/NTU202203401

supports the photodocumentation rate as a quality indicator, and it becomes feasible in the AI era. There are some advantages for endoscopy units using AI-based auditing quality indicators. The AI system may allow real-time quality metric feedback to endoscopists to improve examination quality and explore the potential endoscopist blind spots.

In the third part of this study, we propose a novel DL-based quality assurance model based only on captured colonoscopy images. The entire captured endoscopy image sequence is analyzed by the proposed novel model rather than the single image for quality assurance evaluation. Without additional integration of different reporting systems, it can be used in real-world settings. Several key colonoscopy quality indicators are accurately evaluated by the model and it is feasible in clinical practice. Because the colonoscopy quality indicators are automatically calculated after each endoscopy procedure, endoscopists can focus on careful inspection of the colonic mucosa and capturing adequate images rather than inputting data into the reporting system after the procedure. Because the endoscopy units in the national colonoscopy screening program need to be routinely accredited, the endoscopists have to input some quality indicators in daily examinations manually. The proposal DL model could automatically extract several colonoscopy quality indicators based only on captured colonoscopy images. The automatic quality indicators will be more objective and reliable than manual ones.

In the future, using other AI algorithms will be further studied. For example, Swin Transformer and ConvNeXts could be used to improve the performance of GI colonoscopy image classification. Also, the new self-supervised learning technique could be explored to train a more robust model using large amounts of endoscopy images and reduce the physicians' labeling works.

doi:10.6342/NTU202203401

References

[1] H. Luo *et al.*, "Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study," *Lancet Oncol,* vol. 20, no. 12, pp. 1645-1654, Dec 2019, doi: 10.1016/S1470-2045(19)30637-0.

[2] T. Aoki *et al.*, "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointest Endosc,* vol. 89, no. 2, pp. 357-363 e2, Feb 2019, doi: 10.1016/j.gie.2018.10.027.

[3] S. Attardo *et al.*, "Artificial intelligence technologies for the detection of colorectal lesions: The future is now," *World J Gastroenterol,* vol. 26, no. 37, pp. 5606-5616, Oct 7 2020, doi: 10.3748/wjg.v26.i37.5606.

[4] A. R. Pimenta-Melo, M. Monteiro-Soares, D. Libanio, and M. Dinis-Ribeiro, "Missing rate for gastric cancer during upper gastrointestinal endoscopy: a systematic review and meta-analysis," *Eur J Gastroenterol Hepatol,* vol. 28, no. 9, pp. 1041-9, Sep 2016, doi: 10.1097/MEG.0000000000000657.

[5] J. M. Park, S. M. Huo, H. H. Lee, B. I. Lee, H. J. Song, and M. G. Choi, "Longer Observation Time Increases Proportion of Neoplasms Detected by Esophagogastroduodenoscopy," *Gastroenterology,* vol. 153, no. 2, pp. 460-469 e1, Aug 2017, doi: 10.1053/j.gastro.2017.05.009.

[6] J. M. Park *et al.*, "The effect of photo-documentation of the ampulla on neoplasm detection rate during esophagogastroduodenoscopy," *Endoscopy,* vol. 51, no. 2, pp. 115-124, Feb 2019, doi: 10.1055/a-0662-5523.

[7] J. Cohen and I. M. Pike, "Defining and measuring quality in endoscopy,"

doi:10.6342/NTU202203401

Gastrointest Endosc, vol. 81, no. 1, pp. 1-2, Jan 2015, doi: 10.1016/j.gie.2014.07.052.

[8] J. F. Rey, R. Lambert, and E. Q. A. Committee, "ESGE recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower GI endoscopy," *Endoscopy,* vol. 33, no. 10, pp. 901-3, Oct 2001, doi: 10.1055/s-2001-42537.

[9] S. Marques, M. Bispo, P. Pimentel-Nunes, C. Chagas, and M. Dinis-Ribeiro, "Image Documentation in Gastrointestinal Endoscopy: Review of Recommendations," *GE Port J Gastroenterol,* vol. 24, no. 6, pp. 269-274, Nov 2017, doi: 10.1159/000477739.

[10] M. F. Kaminski *et al.*, "Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative," *Endoscopy,* vol. 49, no. 4, pp. 378-397, Apr 2017, doi: 10.1055/s-0043-103411.

[11] C. J. Rees *et al.*, "UK key performance indicators and quality assurance standards for colonoscopy," *Gut,* vol. 65, no. 12, pp. 1923-1929, Dec 2016, doi: 10.1136/gutjnl-2016-312044.

[12] M. K. Rizk *et al.*, "Quality indicators common to all GI endoscopic procedures," *Gastrointest Endosc,* vol. 81, no. 1, pp. 3-16, Jan 2015, doi: 10.1016/j.gie.2014.07.055.

[13] S. Muthukuru *et al.*, "Quality of Colonoscopy: A Comparison Between Gastroenterologists and Nongastroenterologists," *Dis Colon Rectum,* vol. 63, no. 7, pp. 980-987, Jul 2020, doi: 10.1097/DCR.0000000000001659.

[14] Y. Y. Chang *et al.*, "Deep learning-based endoscopic anatomy classification: an accelerated approach for data preparation and model validation," *Surg Endosc,*

doi:10.6342/NTU202203401

Sep 29 2021, doi: 10.1007/s00464-021-08698-2.

[15] Y. Y. Chang *et al.*, "Upper Endoscopy Photodocumentation Quality Evaluation with Novel Deep Learning System," *Dig Endosc,* Oct 30 2021, doi: 10.1111/den.14179.

[16] R. Bisschops *et al.*, "Overcoming the barriers to dissemination and implementation of quality measures for gastrointestinal endoscopy: European Society of Gastrointestinal Endoscopy (ESGE) and United European Gastroenterology (UEG) position statement," *Endoscopy,* Jan 7 2021, doi: 10.1055/a-1312-6389.

[17] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 20-25 June 2009 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[18] H. Borgli *et al.*, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Sci Data,* vol. 7, no. 1, p. 283, Aug 28 2020, doi: 10.1038/s41597-020-00622-y.

[19] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 2017: Curran Associates Inc., pp. 6000–6010, doi: 10.5555/3295222.3295349.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://arxiv.org/abs/1810.04805

[21] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," presented at the International Conference on

doi:10.6342/NTU202203401

Learning Representations, Addis Ababa, ETHIOPIA, October, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929.

[22] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030,* 2021.

[23] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis,* vol. 58, p. 101539, 2019/12/01/ 2019, doi: https://doi.org/10.1016/j.media.2019.101539.

[24] K. Ohri and M. Kumar, "Review on self-supervised image recognition using deep neural networks," *KNOWLEDGE-BASED SYSTEMS,* vol. 224, JUL 19 2021, doi: 10.1016/j.knosys.2021.107090.

[25] F. Haghighi, M. R. H. Taher, Z. W. Zhou, M. B. Gotway, and J. M. Liang, "Transferable Visual Words: Exploiting the Semantics of Anatomical Patterns for Self-Supervised Learning," *IEEE TRANSACTIONS ON MEDICAL IMAGING,* vol. 40, no. 10, pp. 2857-2868, OCT 2021, doi: 10.1109/TMI.2021.3060634.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the Ieee,* vol. 86, no. 11, pp. 2278-2324, Nov 1998, doi: 10.1109/5.726791.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the Acm,* vol. 60, no. 6, pp. 84-90, Jun 2017, doi: 10.1145/3065386.

[28] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv 1409.1556,* 09/04 2014.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image

doi:10.6342/NTU202203401

Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[30] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[31] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv e-prints,* p. arXiv:1502.03167. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2015arXiv150203167I

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *arXiv e-prints,* p. arXiv:1512.00567. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2015arXiv151200567S

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *arXiv e-prints,* p. arXiv:1602.07261. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2016arXiv160207261S

[34] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks,* vol. 5, no. 2, pp. 157-166, 1994.

[35] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377-2385.

[36] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *arXiv preprint arXiv:1605.07648,* 2016.

doi:10.6342/NTU202203401

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR,* vol. abs/1512.03385, / 2015. [Online]. Available: http://arxiv.org/abs/1512.03385.

[38] X. Li, W. Wang, X. Hu, and J. Yang, "Selective Kernel Networks," p. arXiv:1903.06586. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190306586L

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 22-29 Oct. 2017 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[40] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982.

[41] B. Xie, Y. Mu, D. Tao, and K. Huang, "m-SNE: Multiview Stochastic Neighbor Embedding," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 41, no. 4, pp. 1088-1096, 2011, doi: 10.1109/TSMCB.2011.2106208.

[42] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," (in English), *Neural Comput,* vol. 10, no. 7, pp. 1895-1923, Oct 1 1998, doi: Doi 10.1162/089976698300017197.

[43] H. Zhang *et al.*, "ResNeSt: Split-Attention Networks," p. arXiv:2004.08955. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2020arXiv200408955Z

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," p. arXiv:1708.02002. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2017arXiv170802002L

doi:10.6342/NTU202203401

[45] K. Bishay *et al.*, "Associations between endoscopist feedback and improvements in colonoscopy quality indicators: a systematic review and meta-analysis," *Gastrointest Endosc,* vol. 92, no. 5, pp. 1030-1040 e9, Nov 2020, doi: 10.1016/j.gie.2020.03.3865.

[46] M. D. Rutter *et al.*, "The European Society of Gastrointestinal Endoscopy Quality Improvement Initiative: developing performance measures," *Endoscopy,* vol. 48, no. 1, pp. 81-9, Jan 2016, doi: 10.1055/s-0035-1569580.

[47] R. Bisschops *et al.*, "Performance measures for upper gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative," *Endoscopy,* vol. 48, no. 9, pp. 843-64, Sep 2016, doi: 10.1055/s-0042-113128.

[48] S. Y. Chiu *et al.*, "Faecal haemoglobin concentration influences risk prediction of interval cancers resulting from inadequate colonoscopy quality: analysis of the Taiwanese Nationwide Colorectal Cancer Screening Program," *Gut,* vol. 66, no. 2, pp. 293-300, Feb 2017, doi: 10.1136/gutjnl-2015-310256.

[49] J. C. T. Wong *et al.*, "Adenoma detection rates in colonoscopies for positive fecal immunochemical tests versus direct screening colonoscopies," *Gastrointest Endosc,* vol. 89, no. 3, pp. 607-613 e1, Mar 2019, doi: 10.1016/j.gie.2018.11.014.

[50] A. N. Barkun *et al.*, "Management of Nonvariceal Upper Gastrointestinal Bleeding: Guideline Recommendations From the International Consensus Group," *Ann Intern Med,* vol. 171, no. 11, pp. 805-822, Dec 3 2019, doi: 10.7326/M19-1795.

[51] H. A. Penny *et al.*, "Changing trends in the UK management of upper GI bleeding: is there evidence of reduced UK training experience?," *Frontline*

doi:10.6342/NTU202203401

Gastroenterol, vol. 7, no. 1, pp. 67-72, Jan 2016, doi: 10.1136/flgastro-2014-100537.

[52] H. H. Yen, P. Y. Wu, M. F. Chen, W. C. Lin, C. L. Tsai, and K. P. Lin, "Current Status and Future Perspective of Artificial Intelligence in the Management of Peptic Ulcer Bleeding: A Review of Recent Literature," *J Clin Med,* vol. 10, no. 16, p. 3527, Aug 11 2021, doi: 10.3390/jcm10163527.

[53] H.-H. Yen *et al.*, "Performance Comparison of the Deep Learning and the Human Endoscopist for Bleeding Peptic Ulcer Disease," *J Med Biol Eng,* vol. 41, no. 4, pp. 504-513, 2021/08/01 2021, doi: 10.1007/s40846-021-00608-0.

[54] F. Emura *et al.*, "Principles and practice to facilitate complete photodocumentation of the upper gastrointestinal tract: World Endoscopy Organization position statement," *Dig Endosc,* vol. 32, no. 2, pp. 168-179, Jan 2020, doi: 10.1111/den.13530.

[55] R. Pannala *et al.*, "Artificial intelligence in gastrointestinal endoscopy," *VideoGIE,* vol. 5, no. 12, pp. 598-613, Dec 2020, doi: 10.1016/j.vgie.2020.08.013.

[56] T. L. Ang and G. Carneiro, "Artificial intelligence in gastrointestinal endoscopy," *J Gastroenterol Hepatol,* vol. 36, no. 1, pp. 5-6, Jan 2021, doi: 10.1111/jgh.15344.

[57] H. Takiyama *et al.*, "Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks," *Sci Rep,* vol. 8, no. 1, p. 7497, May 14 2018, doi: 10.1038/s41598-018-25842-6.

[58] S. J. Choi *et al.*, "Development of artificial intelligence system for quality control of photo documentation in esophagogastroduodenoscopy," *Surg Endosc,*

doi:10.6342/NTU202203401

Jan 7 2021, doi: 10.1007/s00464-020-08236-6.

[59] C. Y. Liao *et al.*, "Improving medication safety by cloud technology: Progression and value-added applications in Taiwan," *Int J Med Inform,* vol. 126, pp. 65-71, Jun 2019, doi: 10.1016/j.ijmedinf.2019.03.012.

[60] R. Valori *et al.*, "Performance measures for endoscopy services: A European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative," *United European Gastroenterol J,* vol. 7, no. 1, pp. 21-44, Feb 2019, doi: 10.1177/2050640618810242.

[61] J. L. Teh *et al.*, "Longer examination time improves detection of gastric cancer during diagnostic upper gastrointestinal endoscopy," *Clin Gastroenterol Hepatol,* vol. 13, no. 3, pp. 480-487 e2, Mar 2015, doi: 10.1016/j.cgh.2014.07.059.

[62] T. Kawamura *et al.*, "Examination time as a quality indicator of screening upper gastrointestinal endoscopy for asymptomatic examinees," *Dig Endosc,* vol. 29, no. 5, pp. 569-575, Jul 2017, doi: 10.1111/den.12804.

[63] J. J. Sung *et al.*, "An updated Asia Pacific Consensus Recommendations on colorectal cancer screening," *Gut,* vol. 64, no. 1, pp. 121-32, Jan 2015, doi: 10.1136/gutjnl-2013-306503.

[64] R. Jover *et al.*, "Endoscopist characteristics that influence the quality of colonoscopy," *Endoscopy,* vol. 48, no. 3, pp. 241-7, Mar 2016, doi: 10.1055/s-0042-100185.

[65] H. H. Yen and Y. C. Hsu, "Changing from two- to one-operator colonoscopy insertion technique is feasible with similar quality outcomes," *JGH Open,* vol. 3, no. 2, pp. 159-162, Apr 2019, doi: 10.1002/jgh3.12124.

[66] S. Beg *et al.*, "Quality standards in upper gastrointestinal endoscopy: a position

doi:10.6342/NTU202203401

statement of the British Society of Gastroenterology (BSG) and Association of Upper Gastrointestinal Surgeons of Great Britain and Ireland (AUGIS)," *Gut,* vol. 66, no. 11, pp. 1886-1899, Nov 2017, doi: 10.1136/gutjnl-2017-314109.

[67] W. Januszewicz and M. F. Kaminski, "Quality indicators in diagnostic upper gastrointestinal endoscopy," *Therap Adv Gastroenterol,* vol. 13, p. 1756284820916693, 2020, doi: 10.1177/1756284820916693.

[68] E. J. Lai, A. H. Calderwood, G. Doros, O. K. Fix, and B. C. Jacobson, "The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research," *Gastrointest Endosc,* vol. 69, no. 3 Pt 2, pp. 620-5, Mar 2009, doi: 10.1016/j.gie.2008.05.057.

[69] I. Pillai, G. Fumera, and F. Roli, "Designing multi-label classifiers that maximize F measures: State of the art," (in English), *Pattern Recognit.,* Article vol. 61, pp. 394-404, Jan 2017, doi: 10.1016/j.patcog.2016.08.008.

[70] H. M. Chiu *et al.*, "Long-term effectiveness of faecal immunochemical test screening for proximal and distal colorectal cancers," *Gut,* Jan 25 2021, doi: 10.1136/gutjnl-2020-322545.

[71] M. F. Kaminski *et al.*, "Quality indicators for colonoscopy and the risk of interval cancer," *N Engl J Med,* vol. 362, no. 19, pp. 1795-803, May 13 2010, doi: 10.1056/NEJMoa0907667.

[72] L. Yao *et al.*, "A Gastrointestinal Endoscopy Quality Control System Incorporated With Deep Learning Improved Endoscopist Performance in a Pretest and Post-Test Trial," *Clin Transl Gastroenterol,* vol. 12, no. 6, p. e00366, Jun 15 2021, doi: 10.14309/ctg.0000000000000366.

[73] F. M. Aslinia *et al.*, "Anatomic classification of the endoscopic appearance of the normal appendiceal orifice: a novel tool for recognition and documentation

doi:10.6342/NTU202203401

of cecal intubation," *Clin Anat,* vol. 25, no. 4, pp. 496-502, May 2012, doi: 10.1002/ca.21276.

[74] V. de Jonge *et al.*, "Quality evaluation of colonoscopy reporting and colonoscopy performance in daily clinical practice," *Gastrointest Endosc,* vol. 75, no. 1, pp. 98-106, Jan 2012, doi: 10.1016/j.gie.2011.06.032.

[75] S. Ouazzani *et al.*, "Implementation of colonoscopy quality monitoring in a Belgian university hospital with integrated computer-based extraction of adenoma detection rate," *Endosc Int Open,* vol. 9, no. 2, pp. E197-E202, Feb 2021, doi: 10.1055/a-1326-1179.

[76] T. D. Gohel *et al.*, "Polypectomy rate: a surrogate for adenoma detection rate varies by colon segment, gender, and endoscopist," *Clin Gastroenterol Hepatol,* vol. 12, no. 7, pp. 1137-42, Jul 2014, doi: 10.1016/j.cgh.2013.11.023.

[77] E. A. Holzwanger *et al.*, "Benchmarking definitions of false-positive alerts during computer-aided polyp detection in colonoscopy," *Endoscopy,* Nov 2 2020, doi: 10.1055/a-1302-2942.

[78] S. M. Milluzzo, P. Cesaro, L. M. Grazioli, N. Olivari, and C. Spada, "Artificial Intelligence in Lower Gastrointestinal Endoscopy: The Current Status and Future Perspective," *Clin Endosc,* vol. 54, no. 3, pp. 329-339, May 2021, doi: 10.5946/ce.2020.082.

[79] S. S. Deliwala *et al.*, "Artificial intelligence (AI) real-time detection vs. routine colonoscopy for colorectal neoplasia: a meta-analysis and trial sequential analysis," *Int J Colorectal Dis,* May 1 2021, doi: 10.1007/s00384-021-03929-3.

[80] M. Cho, J. H. Kim, K. S. Hong, J. S. Kim, H. J. Kong, and S. Kim, "Identification of cecum time-location in a colonoscopy video by deep learning analysis of colonoscope movement," *PeerJ,* vol. 7, p. e7256, 2019, doi:

doi:10.6342/NTU202203401

10.7717/peerj.7256.

[81] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *arXiv e-prints,* p. arXiv:2201.03545, 2022. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2022arXiv220103545L.

[82] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *MACHINE LEARNING,* vol. 109, no. 2, pp. 373-440, FEB 2020, doi: 10.1007/s10994-019-05855-6.

[83] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *arXiv e-prints,* p. arXiv:1911.05722, 2019. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv191105722H.

[84] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv e-prints,* p. arXiv:2002.05709, 2020. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2020arXiv200205709C.

[85] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big Self-Supervised Models are Strong Semi-Supervised Learners," *arXiv e-prints,* p. arXiv:2006.10029, 2020. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2020arXiv200610029C.

[86] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," *arXiv e-prints,* p. arXiv:1904.05862, 2019. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190405862S.

doi:10.6342/NTU202203401