Efficacy of a Comprehensive Binary Classification Model Using a Deep Convolutional Neural Network for Wireless Capsule Endoscopy

Sang Hoon Kim
Dongguk University Ilsan Hospital
Youngbae Hwang
Chungbuk National University
Dong Jun Oh
Dongguk University Ilsan Hospital
Ji Hyung Nam
Dongguk University Ilsan Hospital
Ki Bae Kim
Chungbuk National University Hospital
Junseok Park
Soonchunhyang University Hospital
Hyun Joo Song
Jeju National University Hospital
Yun Jeong Lim (drlimyj@gmail.com)
Department of Internal Medicine, Dongguk University Ilsan Hospital, Dongguk University College of Medicine, Goyang, Korea.

Research Article

Keywords: Manual reading of capsule endoscopy (CE), clinical practice, CNN-based binary classification, computer-aided reading methods

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-125525/v1

Abstract

Manual reading of capsule endoscopy (CE) video is a time-consuming process in diagnosing small bowel diseases. Although many algorithms have been introduced, multi-diagnosis has not been sufficiently validated. They are promising but still premature to be used in clinical practice. Therefore, we developed a practical binary classification model and tested it with unseen cases.

400,000 CE images were randomly selected from 84 cases. Among them, 240,000 were used to train an algorithm to categorize images binarily. The remaining images were utilized for validation and internal testing. The algorithm was externally tested with 256,591 unseen images.

Diagnostic accuracy was 98.067% when the trained model was applied to the validation set. It was 97.946% when applied to images for internal testing. When the model was applied to a dataset provided by an independent hospital not participated during training, its accuracy was 85.470%. The area under the curve was 0.922.

Our binary classification model showed excellent internal test results, and when tested in unseen external cases, misreadings were slightly increased while judging 'insignificant' images containing ambiguous substances. When we can get over this problem, CNN-based binary classification will become the most promising candidates for developing clinically ready computer-aided reading methods.

Introduction

The small intestine has been difficult to examine by traditional endoscopic and radiologic techniques.[1] Wireless capsule endoscopy (WCE) is widely used for small bowel imaging since its introduction by Iddan et al. in 2000.[2] It is a non-invasive, patient-friendly method that offers complete video-facilitating detection and monitoring of lesions within the entire small bowel. According to the European Society of Gastrointestinal Endoscopy (ESGE) recommendations, WCE is the first-line evaluation method in patients with obscure gastrointestinal (GI) bleeding. In patients with suspected Crohn's disease and negative ileo-colonoscopy findings, ESGE recommends WCE as the initial diagnostic modality for investigating the small bowel.[3] Various other small bowel diseases such as inherited polyposis syndromes can also be diagnosed and followed-up by capsule endoscopy.

Although 20 years have passed since its introduction, the way of reading WCE has not changed significantly. It is performed manually by expert gastroenterologists who check the entire recorded video. As a recorded video clip shows a long playing time of 10 to 13 hours, an average reading time for a single WCE session is 30 ~ 40 minutes.[4,5] Since most small bowel capsule images show normal mucosa, conventional manual reading is an inefficient and time-consuming method.[6,7] The fast video playback and extended reading time lead to loss of reader's concentration, which may increase the chance of missing lesions. In addition, since the number of gastroenterologists who have sufficient experience with small bowel diseases is limited, the way of manual reading of WCE images inhibits further propagation

of capsule endoscopy techniques. In the last few years, utilizing artificial intelligence (AI) as a core technology of the fourth industrial revolution for capsule endoscopy has gained significant attention for its potential to automatically detect diseases from the video and shorten WCE reading time.[8] The main deep learning algorithm for image analysis is the convolutional neural network (CNN). Several studies covering the use of CNN model in reading WCE images have been published. Some CNN based algorithms have been successful for detecting a variety of small bowel diseases, including ulcers, polyps, Crohn's disease, angioectasia, and bleeding.[9–17]

Despite advances in CNN-based reading algorithms, a recent meta-analysis with a systematic review of current literature pertaining to deep learning implementation in WCE has emphasized that the technology is still on the research stage.[18] The reason why it is still challenging to use in clinical practice because the majority of developed CNN-based automated models have only dealt with the detection of a single abnormality or have failed to perform validation in unseen images.

Researchers have been tried to develop CNN-based algorithms that can detect not only various small bowel lesions simultaneously but also demonstrate high accuracy within unseen images.[19,20] However, to date, this goal could not be accomplished despite various efforts such as increasing the number of images for AI learning. Therefore, creating a new CNN-based model that is practical and clinically ready-to-use is needed. As part of that effort, we developed a CNN-based binary classification algorithm that could categorize images into clinically meaningful lesions and those that are not. Since only selected images need to be read by gastroenterologists with such an algorithm, the reading time is expected to be drastically reduced. Hence, the objective of this study was to develop a CNN-based binary classification model and validate it with unseen outside WCE images.

Methods

Preparation of images for AI training, validation, and internal testing

To train and develop CNN-based AI, 84 capsule endoscopy videos from June 2019 to December 2019 were retrospectively acquired from a single institute (Dongguk University Ilsan Hospital). The study protocol was approved by Dongguk University Ilsan Hospital Institutional Review Board (No. 2018-00-009-002). In that period, WCE was performed using a Mirocam MC1600® device (Intromedic, Seoul, South Korea). Images were extracted in PNG format (320 × 320 pixel) using MiroView® 4.0, a dedicated software. After data anonymization, all small bowel images were reviewed and labeled manually into 7 subgroups (normal mucosa, bile predominant, air bubbles, debris, inflamed mucosa, atypical vascularity, and bleeding) by three experienced gastroenterologists. After initially labeling all images for learning and validation by two expert endoscopists, a senior gastroenterologist finally confirmed the classification of images to keep consistency across all data. Images were later binarily categorized as 'Significant' when reviewers judged the lesion to be of clinical significance and 'Insignificant' for the rest. Images that were

initially labeled as inflamed mucosa, atypical vascularity, or bleeding were classified as ‘Significant’. The ‘Insignificant’ group included images of normal mucosa, bile, air bubbles, or debris. A total of 400,000 images, including 200,000 clinically significant images and 200,000 insignificant images, were randomly selected for AI training, validation, and internal testing. In this study, even though capsule endoscopy images were collected retrospectively, we obtained informed consent in advance from all patients participating in the study. In particular, informed consent was obtained from legally authorized representatives for minors (age 18 or younger) participants. Besides, all research protocols were conducted under relevant guidelines and regulations.

Distribution of collected CE images

After labeling, images (n = 400,000) were classified into the following three categories (Fig. 1): training (n = 240,000, 60%), validation (n = 80,000, 20%), and internal testing (n = 80,000, 20%). For 80,000 (20%) images to be used for the validation process, it was necessary to set the best parameters showing optimal diagnostic accuracy. These images for validation were also used for additional training. Therefore, 320,000 images were used in the development of the algorithm. The remaining 80,000 (20%) images were exploited as data for internal testing as they were not used for training.

Collecting external hospital images for testing of the developed AI model

We retrospectively collected 30 capsule endoscopy (CE) videos from another medical hospital (Jeju National University Hospital) with the approval of its Institutional Review Board (Jeju National University Institutional Review Board, ‘JEJUNUH 2019-11-010’). A total of 258,256 images (significant: 179,246; insignificant: 79,010) were extracted from videos using the same capsule device (Mirocam MC1600®) and the same image extraction protocol. For the set of collected images, the three identical endoscopists who participated in the image classification for AI training arranged these data into two groups (‘significant’ and ‘insignificant’) according to the same criteria used for training. Through this, it was possible to evaluate the diagnostic accuracy of AI model for ‘unseen’ data acquired from a non-trained hospital.

Development of CNN-based AI for auxiliary reading model

For binary classification of CE images, an Inception-Resnet-V2 model in TensorFlow-Slim (TF-Slim) library renowned for its advantages such as light and efficient multi-level feature extraction of an inception module and deeper layers of a Resnet module to train an AI, was used. The training was done with 20 epochs, batch size of 24, and step size of 200,000. The elapsed time to transform image data into TFrecord (binary data rapidly readable in Tensorflow) for 80,000 test images was 2621.208 seconds. The processing time for evaluating transformed data was 431.258 seconds. Therefore, our AI showed a processing speed of 26.208 frames per second. To analyze which regions influenced the selection of final class, class activation maps (CAP) were drawn based on a channel-wise aggregation method (Fig. 2). This was done using a global average pooling (GAP) for each channel after the final convolutional layer

stage, demonstrating pixel-wise predicted values for the class. Using CAP, in which regions with clinical significance were indicated by red color, we were able to verify that predictions made by the trained deep learning network were similar to those of endoscopists.

A jet color map shows normalized prediction. Reddish and bluish colors refer to 1 and 0, respectively. This figure describes the approximate mechanism of the developed AI.

Outcomes and statistical analysis

The primary outcome was the area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, and accuracy of the CNN based AI model on determining whether each image possessed lesions of clinical significance. The trained model shaped the region of clinical significance and described the probability score of the lesion (range, 0–1). We verified the score threshold for the best output through the validation process. Data were analyzed with Statistical Package for Social Science, Ver 20.0 (SPSS Inc., Chicago, IL, USA).

Results

Study population

Clinical and demographic characteristics of images used for algorithm training and external testing are shown in Table 1. Cases used for training had an average age of 49.84 years (range, 16–92 years). Their indications for CE were obscure GI bleeding (51.2%), inflammatory bowel disease (35.7%), and small bowel tumors (11.9%). Significant and insignificant images were randomly selected (200,000 each). Relative proportions of inflammatory lesions and air bubbles were high. The average age of cases used for external testing was 52.31 years (range, 16–76 years). Their indications were not significantly different from those in the training dataset. Significant images with inflammatory lesions or bleeding were included relatively more than others. Therefore, there was a difference in image composition between data used for training and data used for external testing.

[table 1.] Clinical And Demographic Characteristics Of Cases

Characteristics	Training dataset (n = 84)	External test dataset (n = 30)
No. of images	400,000	256,591
Age (years), mean (± SD)	49.84 (± 19.17)	52.31 (± 19.08)
Sex, male	55 (65.5%)	18 (60.0%)
Indication of capsule endoscopy		
- Obscure GI bleeding	43 (51.2%)	19 (63.3%)
- Small bowel tumors	10 (11.9%)	3 (10.0%)
- Inflammatory bowel disease	30 (35.7%)	8 (26.7%)
- Others	1 (1.2%)	0 (0.0%)
Types of included images		
- Normal mucosa	42,729	21,317
- Bile	58,812	18,196
- Air bubbles	87,494	6,550
- Debris	10,965	32,947
- Vascular	32,116	4,733
- Inflammatory	92,297	85,387
- Bleeding	75,587	89,126
Binary classification of included images		
- Significant	200,000	179,246
- Insignificant	200,000	79,010

Internal test result: Binary classification capability through optimal threshold setting

Calculated probabilities of significance and results of class map activation for all images were reviewed by three gastroenterologists who contributed to the labeling of training images (Table 2). The difference in the classification ability for each cut-off probability score was reviewed by endoscopists who agreed to set the cut-off threshold at 0.5 after examining the concordance between manual classification results and calculated probability values for validation images. Based on this threshold, the AI system's sensitivity, specificity, and accuracy were 98.691%, 97.208%, and 97.946%, respectively, for the selected internal test set of images. The AUC value in internal testing was 0.998.

[Table 2.] Changes in the classification ability for each cut-off probability score

Cut-off value (Probability score)	Sensitivity (%)	Specificity (%)	Accuracy (%)
0.1	0.893	0.999	0.946
0.2	0.928	0.999	0.963
0.3	0.949	0.997	0.973
0.4	0.963	0.994	0.978
0.5	0.973	0.988	0.980
0.541*	0.977	0.985	0.981
0.6	0.982	0.978	0.980
0.7	0.989	0.957	0.973
0.8	0.994	0.911	0.953
0.9	0.998	0.784	0.891
** Estimated value according to Youden index.*			

External test result: Accuracy drop for new & unseen cases

When the developed algorithm was tested with an external set of images, its accuracy was dropped to 85.470%, with a sensitivity of 89.684% and a specificity of 75.994%. The AUC value was 0.922 in external testing. The ROC for external testing compared to the internal validation process is demonstrated in Fig. 3.

The performance of the CNN model at each validation level is described in Table 3. At the preset cut-off value of 0.5, sensitivities of the model for significant lesions were as follows: 97.3% for internal validation, 97.2% for internal testing, and 89.7% for external validation.

[Table 3.] Performance of binary classification model at each validation level

	Internal validation	Internal testing	External testing
No. of images	80,000	80,000	256,591
AUC	0.99794	0.99775	0.92194
Accuracy, %	98.067	97.946	85.470
Sensitivity, %	97.322	97.208	89.684
Specificity, %	98.817	98.691	75.994
- All performance outcomes were calculated when a cut-off value of 0.5 was applied.			

Difference in classification performance according to image type

We compared how well the CNN model classified images into significant and insignificant groups compared to manual classification by endoscopists (Table 4). Of a total of 179,246 images that were classified as ‘significant’ by endoscopists, 160,849 images were classified correctly by CNN, leading to a false negative rate of about 10.3%. On the other hand, out of a total of 79,010 images classified as ‘insignificant’ by endoscopists, AI accurately classified 60,770 images as insignificant, showing a false positive rate of 23.1%. It was found that specificity was a more challenging goal to achieve than sensitivity in the binary classification process by the CNN model.

Among subtypes of insignificant images, which subtype was more prone to misreading was then analyzed. As a result, we found that when air bubbles occupied most of the image, the image was more likely to be classified as significant in 40.3% of cases (Table 5). For example, as shown in Fig. 4, even when only air bubbles and normal mucosa in small bowel lumen were present, the CNN class activation map pointed out that the area without significance possessed a lesion with a probability score above the threshold.

[Table 4.] Classification results of manually labeled images by the convolutional neural network

CNN classification (External testing)	Classification by endoscopists		Total	
	Significant	Insignificant		
Significant	160,849	18,240	179,089	**PPV 89.8%**
Insignificant	18,397	60,770	79,167	**NPV 76.8%**
Total	179,246	79,010		
	Sensitivity 89.7%	**Specificity 75.9%**		

[Table 5.] False positive rate by the CNN model according to subtypes of insignificant images

Subtypes of insignificant images					
	Mucosa	**Bile**	**Air bubbles**	**Debris**	**Total**
Total No.	21,317	18,196	6,550	32,947	79,010
False (+) No.	5,051	3,172	2,637	7,380	18,240
False (+) %	23.69%	17.43%	40.26%	22.40%	23.09%
** In external testing dataset.*					

Discussion

In this study, we developed a practical CNN based AI model performing comprehensive binary classification and tested its accuracy using images from both an institution where AI learning was conducted and an independent external hospital. To the best of our knowledge, this is the first AI model developed and tested using Mirocam® capsule image-set. The diagnostic accuracy during validation and internal testing showed values above 98%, with an excellent AUC of 0.99. Unfortunately, when it was tested using external hospital data, it showed an accuracy reduction of about 10% and an AUC drop of 0.07.

It is essential to detect various gastrointestinal pathologies for automated AI reading in CE. However, a broad spectrum of diseases can be present in the small intestine. Therefore, it is difficult for an AI algorithm to classify them correctly in a multi-class manner. Moreover, in AI learning, it is not easy to develop a successful model due to problems such as data imbalance between normal and lesion images. In addition, when two or more lesions are mixed in a single image without individually annotating them, it may lead to incorrect learning and inappropriate reading. In this regard, a binary classification CNN based model is considered to be a practical and achievable concept. Since this binary model aims to select meaningful photos and provide them to ‘human’ endoscopists, sensitivity and positive predictive value are important parameters when evaluating such a model. Our algorithm showed a high sensitivity of 89.7% and a PPV of 89.8% even for unseen external cases. It is expected to save reading time and reduce the rate of missing lesions for non-experts.

Since reading capsule endoscopic video is a very tedious and time-consuming process, AI-assisted image recognition and classification technology is auspicious. In the last two years, various studies on the automated reading of CE have been published. These include a study showing very high sensitivity and specificity by detecting angioectasia with a CNN model published by Leenhardt et al.[14], a model published by Aoki et al.[12] that judged erosions and ulcerations with an accuracy of 90%, and a research using weakly supervised learning AI model for diagnosing a variety of pathologies by Iakovidis et al.[21] In addition, Ding et al.[17] have collected over 100 million CE images from 77 hospitals and trained AI with 158,235 images from 1970 cases and evaluated whether an accurate diagnosis of various types of lesions is possible.

Nevertheless, most AI models for CE are considered insufficient for clinical use, as most models can only detect one or two specific disease entities. Despite training with a relatively large number of images, their diagnostic accuracy is still insufficient for clinical use. Indeed, a study design of testing with images from the same institution where training images are gathered (even though it was multi-center based) raises questions whether the developed AI model will still show high accuracy when it is tested with images from external hospitals.

WCE is taken as a motion video. At least 50,000 images are saved per session. Looking at capsule endoscopic images one by one, each image may contain either pure significant (or insignificant) content or significant lesions mixed with some insignificant materials. Most AI models are specialized in detecting 'lesions' only. They get quite confused when the lesion is widely covered by insignificant content. In this regard, our model also showed that insignificant images were more likely to be false positively classified as 'Significant' than the opposite way. To overcome these limitations, AI model should be trained to also recognize 'insignificant' contents correctly and show an accurate probability value even with images of mixed significance. This suggests that an improved algorithm that can separate groups of consecutive low-quality images and exclude these groups from AI reading may not only enhance the reading accuracy but also reduce the presentation of unnecessary images to 'human' interpreter since the AI's false positive rate is relatively high for images with poor bowel preparation (which contains bile, bubbles, and debris). Therefore, promoting the ability to judge the quality of bowel preparation by the CNN-based model is important.

Our results showed that the diagnostic accuracy of AI with unseen external data decreased by approximately 10%. This might be due to the demographic difference of patients undergoing capsule endoscopy at each hospital. The trained AI model might have also faced completely new endoscopic images that were not experienced during training. Increasing training images by ten times is known to improve the diagnostic accuracy by 3% in the field of WCE analysis according to Segui et al.[22] As diagnostic accuracy improves when there are more images for learning, if additional supervised learning is done using unseen external data, the following accuracy test with CE images of two hospitals is expected to improve over 95%. However, some engineers concern a problem of 'catastrophic forgetting' that the accuracy of diagnosis for the dataset of initial training hospital may be reduced when AI is additionally trained with a large number of images from other institutions. In addition, manually annotating massive data for CE images is also a laborious, error-prone, and expensive process that may lead to huge label bias. Therefore, the belief that only multi-center based large number of training images will lead us to a clinically ready AI model may not be necessarily correct. Instead, we believe that in order to advance to a stage of multi-class reading, it is necessary to form an optimized training image set best suitable for the health environment of a regional community where capsule endoscopy is performed. By properly adjusting 'Significant' and 'Insignificant' ratios of images as well as compositions of lesions (for example, the ratio among inflammatory, vascular, bleeding, and small bowel tumors), we can expect a more rational AI reading outcome. Regarding the optimal image composition ratio, additional studies are needed.

This study has some limitations. First, AI training was based on images of a single institution. Second, the labeling of images was done by three gastroenterologists. There might be some inter-observer differences between them. In addition, it was a study using retrospectively collected data without utilizing clinical data during learning and feedback processes. We believe that the need for reinforcement learning through a combination of images with prospective clinical data, including final diagnosis and prognosis, is high to improve diagnostic accuracies of future AI models dramatically.

In summary, our study demonstrates the practical applicability of a CNN-based comprehensive binary classification model in small bowel capsule endoscopy. We believe that this model is a promising tool that can be used in everyday practice in the near future.

Abbreviations

WCE, Wireless capsule endoscopy

CNN(s), Convolutional neural network(s)

TF-Slim, TensorFlow-Slim

AUC, Area under the curve

ESGE, European Society of Gastrointestinal Endoscopy

GI, Gastrointestinal

AI, Artificial intelligence

CE, Capsule endoscopy

CAP, Class activation map

GAP, Global average pooling

ROC, Receiver operating characteristic

Declarations

Acknowledgement

This research was supported by a grant (grant number: HI19C0665) from the Korean Health

Technology R & D project through the Korean Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea.

Author contributions

Conception and design: Yun Jeong Lim

Data collection and interpretation: Hyun Joo Song, Junseok Park, Ki Bae Kim, Dong Jun Oh and Ji Hyung Nam

Data analysis and drafting of the article: Sang Hoon Kim and Youngbae Hwang

Critical revision: Yun Jeong Lim and Sang Hoon Kim

* All authors approved the final version of this article.

Competing interests.

The authors declare no competing interests.

Data Availability

The datasets analyzed during the current study are available from the corresponding authors on reasonable request.

References

1. Aktas, H. & Mensink, P. B. Small bowel diagnostics: Current place of small bowel endoscopy. *Best Practice & Research Clinical Gastroenterology* **26**, 209–220, doi:https://doi.org/10.1016/j.bpg.2012.03.007 (2012).
2. Iddan, G., Meron, G., Glukhovsky, A. & Swain, P. Wireless capsule endoscopy. *Nature* **405**, 417, doi:10.1038/35013140 (2000).
3. Pennazio, M. *et al.* Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. *Endoscopy* **47**, 352–376, doi:10.1055/s-0034-1391855 (2015).
4. Mishkin, D. S. *et al.* ASGE Technology Status Evaluation Report: wireless capsule endoscopy. *Gastrointestinal endoscopy* **63**, 539–545, doi:10.1016/j.gie.2006.01.014 (2006).
5. Koulaouzidis, A., Iakovidis, D. K., Karargyris, A. & Plevris, J. N. Optimizing lesion detection in small-bowel capsule endoscopy: from present problems to future solutions. *Expert review of gastroenterology & hepatology* **9**, 217–235, doi:10.1586/17474124.2014.952281 (2015).
6. Lee, N. M. & Eisen, G. M. 10 years of capsule endoscopy: an update. *Expert review of gastroenterology & hepatology* **4**, 503–512, doi:10.1586/egh.10.44 (2010).
7. Rondonotti, E. *et al.* Can we improve the detection rate and interobserver agreement in capsule endoscopy? *Digestive and liver disease: official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver* **44**, 1006–1011, doi:10.1016/j.dld.2012.06.014 (2012).

8. Hricak, H. 2016 New Horizons Lecture: Beyond Imaging-Radiology of Tomorrow. *Radiology* **286**, 764–775, doi:10.1148/radiol.2017171503 (2018).
9. *Conference proceedings: … Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society.Annual Conference* **2016**, 639–642, doi:10.1109/embc.2016.7590783 (2016).
10. Fan, S., Xu, L., Fan, Y., Wei, K. & Li, L. Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. *Physics in medicine and biology* **63**, 165001, doi:10.1088/1361-6560/aad51c (2018).
11. Alaskar, H., Hussain, A., Al-Aseem, N., Liatsis, P. & Al-Jumeily, D. Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images. *Sensors (Basel, Switzerland)* **19**, doi:10.3390/s19061265 (2019).
12. Aoki, T. *et al.* Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointestinal endoscopy* **89**, 357–363.e352, doi:10.1016/j.gie.2018.10.027 (2019).
13. Klang, E. *et al.* Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. *Gastrointestinal endoscopy* **91**, 606–613.e602, doi:10.1016/j.gie.2019.11.012 (2020).
14. Leenhardt, R. *et al.* A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy. *Gastrointestinal endoscopy* **89**, 189–194, doi:10.1016/j.gie.2018.06.036 (2019).
15. Aoki, T. *et al.* Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network. *Journal of gastroenterology and hepatology*, doi:10.1111/jgh.14941 (2019).
16. Tsuboi, A. *et al.* Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images. *Digestive endoscopy: official journal of the Japan Gastroenterological Endoscopy Society* **32**, 382–390, doi:10.1111/den.13507 (2020).
17. Ding, Z. *et al.* Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. *Gastroenterology* **157**, 1044–1054.e1045, doi:10.1053/j.gastro.2019.06.025 (2019).
18. Soffer, S. *et al.* Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointestinal endoscopy*, doi:10.1016/j.gie.2020.04.039 (2020).
19. Hwang, Y., Park, J., Lim, Y. J. & Chun, H. J. Application of Artificial Intelligence in Capsule Endoscopy: Where Are We Now? *Clinical endoscopy* **51**, 547–551, doi:10.5946/ce.2018.173 (2018).
20. Park, J. *et al.* Recent Development of Computer Vision Technology to Improve Capsule Endoscopy. *Clinical endoscopy* **52**, 328–333, doi:10.5946/ce.2018.172 (2019).
21. Iakovidis, D. K., Georgakopoulos, S. V., Vasilakakis, M., Koulaouzidis, A. & Plagianakos, V. P. Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. *IEEE transactions on medical imaging* **37**, 2196–2210, doi:10.1109/tmi.2018.2837002 (2018).

22. Seguí, S. *et al.* Generic feature learning for wireless capsule endoscopy analysis. *Computers in biology and medicine* **79**, 163–172, doi:10.1016/j.compbiomed.2016.10.011 (2016).

Figures

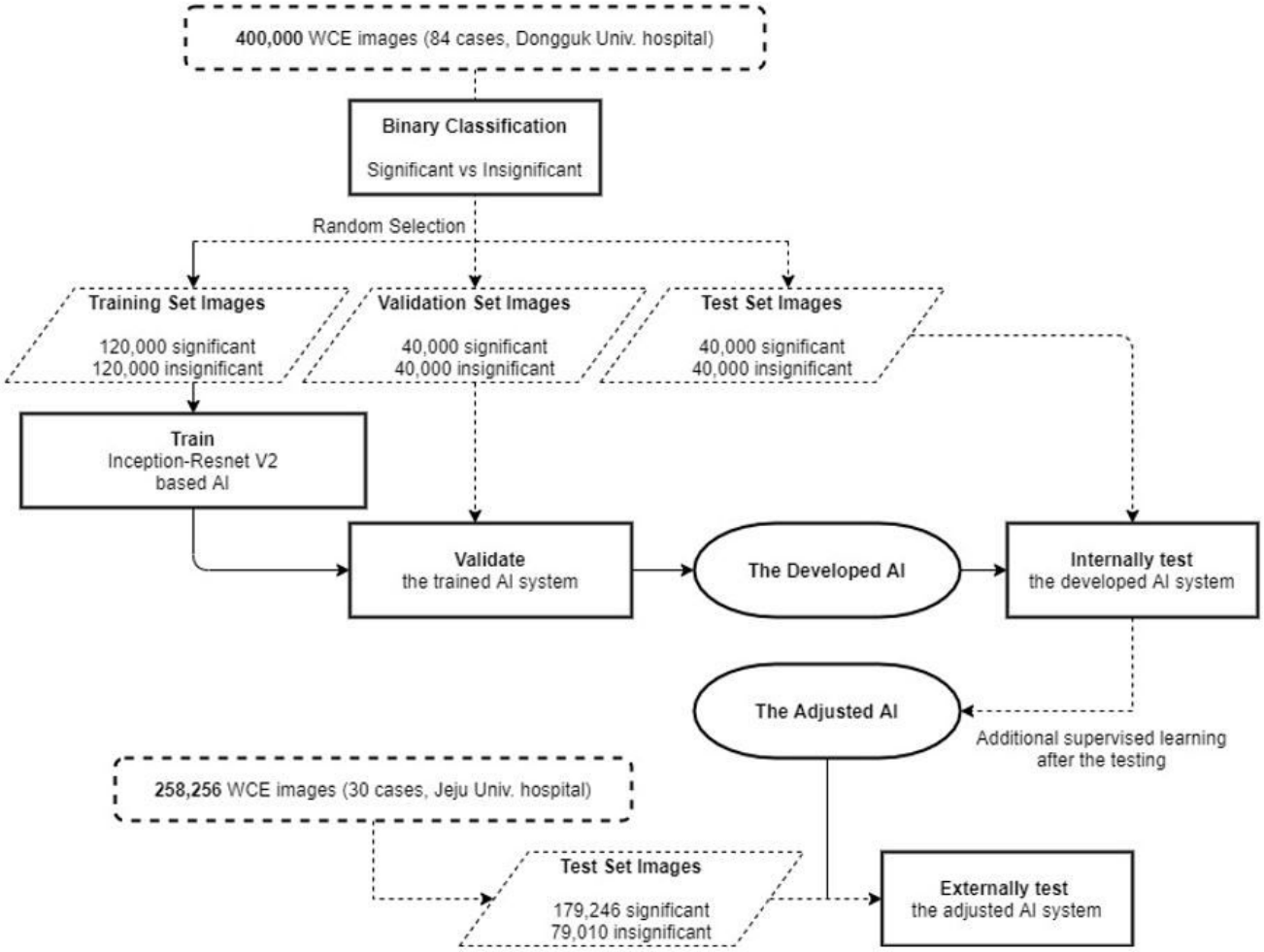


Figure 1

Flowchart of the Study Design.

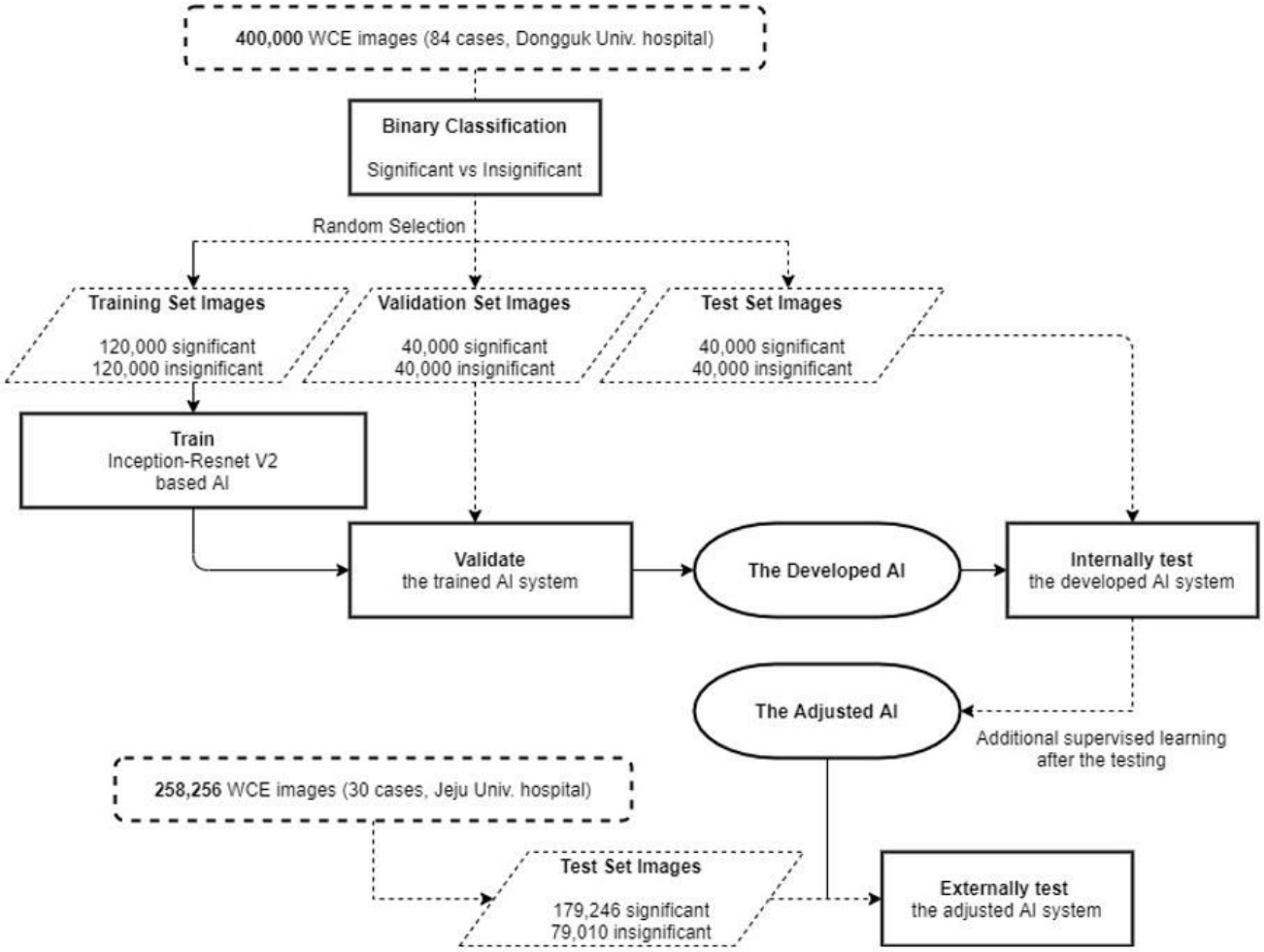


Figure 1

Flowchart of the Study Design.

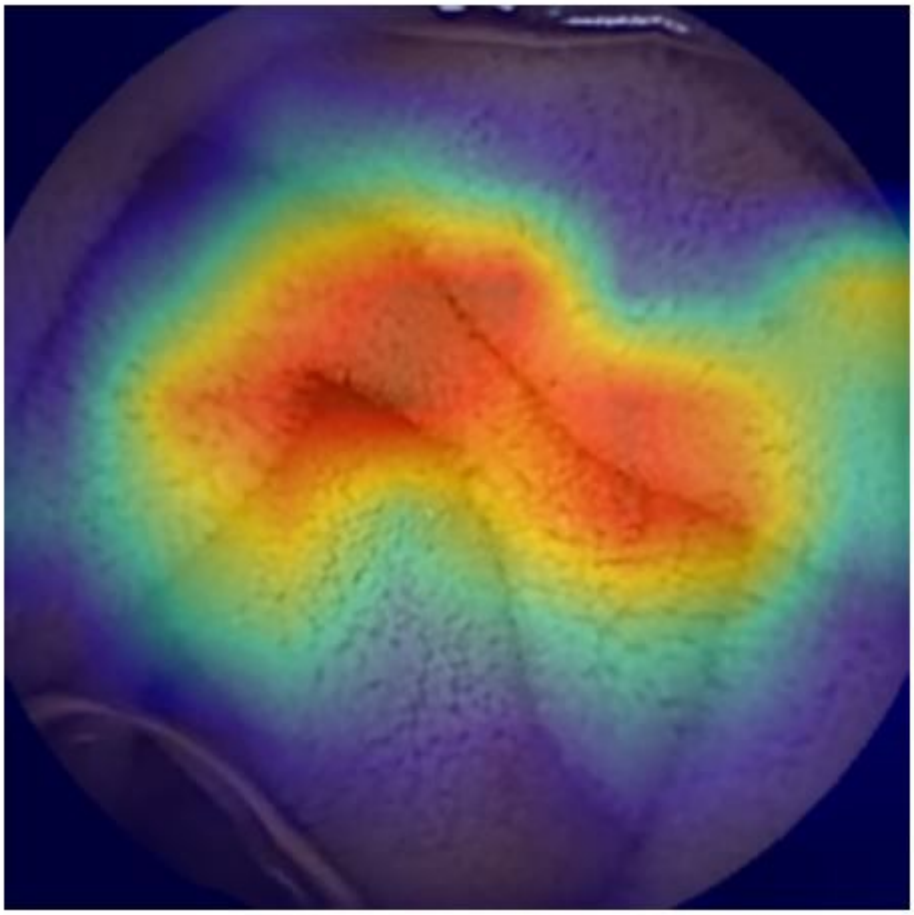


Figure 2

Class Activation Map of Significant Lesions.

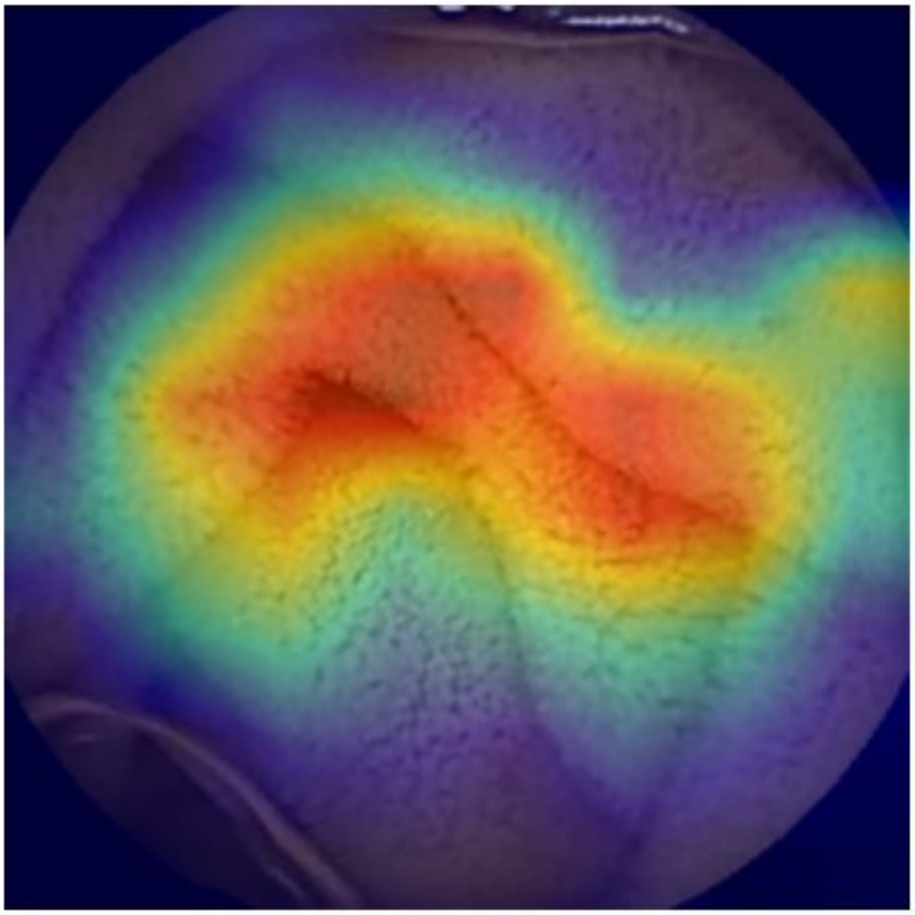


Figure 2

Class Activation Map of Significant Lesions.

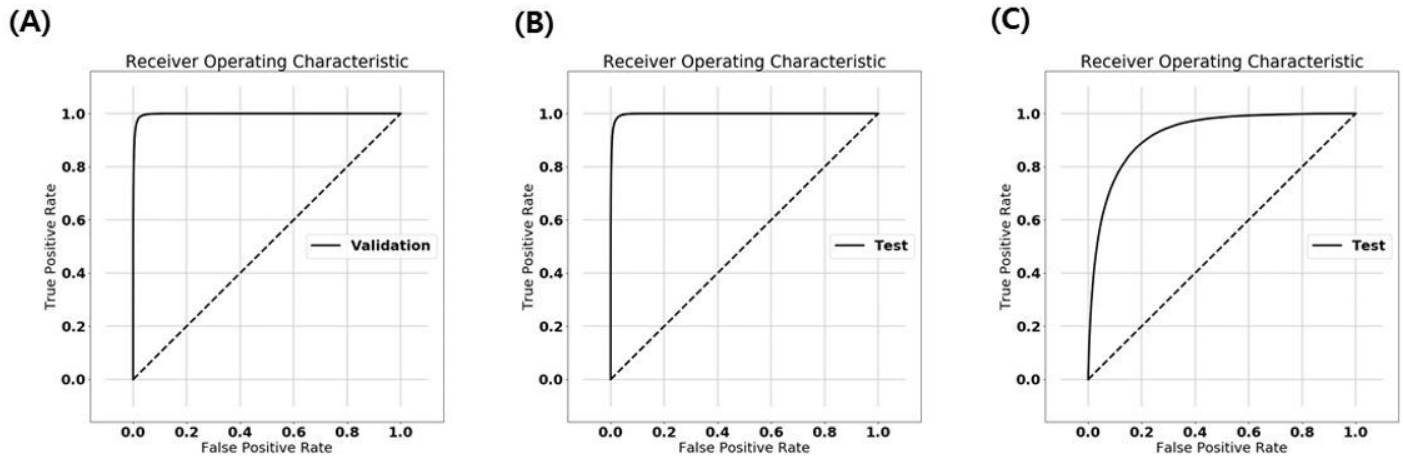


Figure 3

Receiver Operating Characteristic Curves of (A) Internal validation, (B) Internal testing, and (C) External testing.

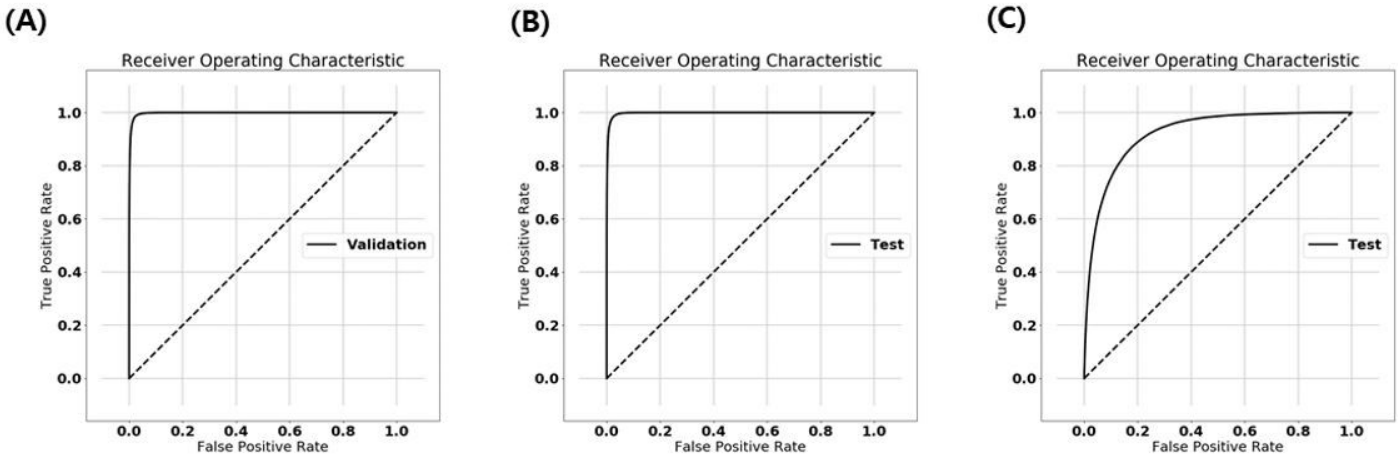


Figure 3

Receiver Operating Characteristic Curves of (A) Internal validation, (B) Internal testing, and (C) External testing.

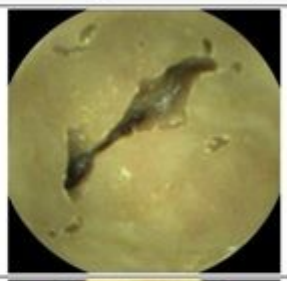
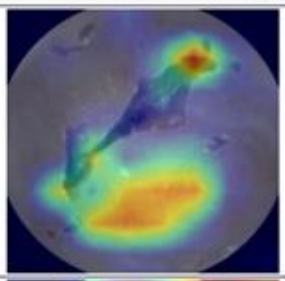
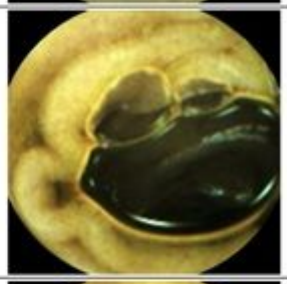
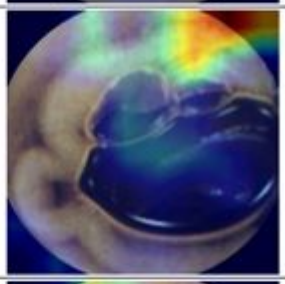
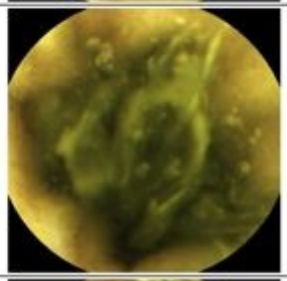
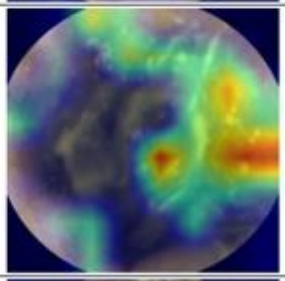
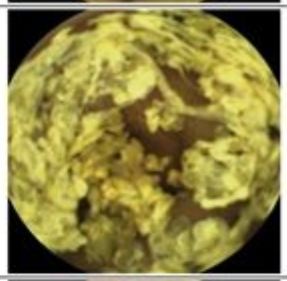
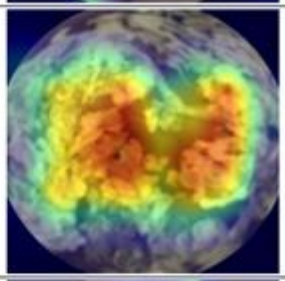
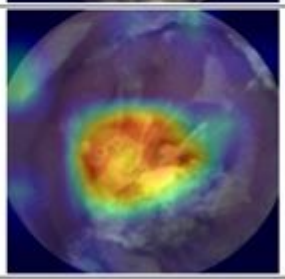
Image No.	Type	CE image in white light	Class Activation Map
JM1053_02_47_00_3	False positive		
JM3068_02_10_24_0	False positive		
JM3064_04_13_45_0	False positive		
JM5034_04_09_48_9	False positive		
JM3071_05_51_55_5	False negative		

Figure 4

Examples of Several Class Activation Maps of Misclassified Images.

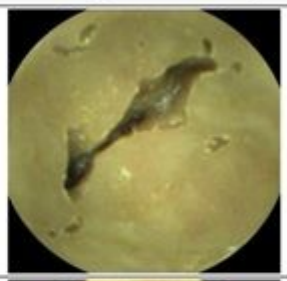
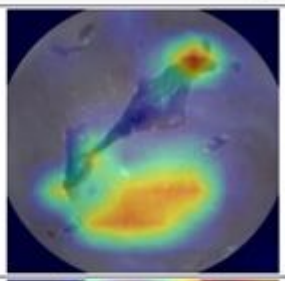
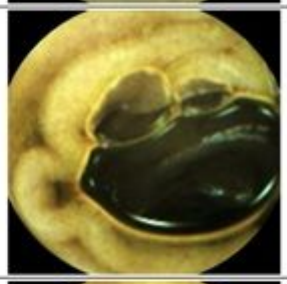
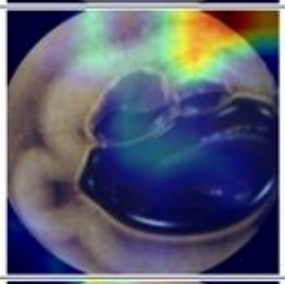
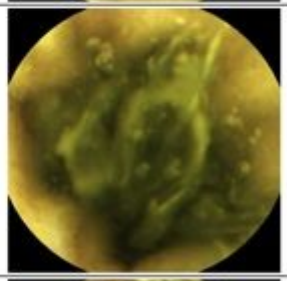
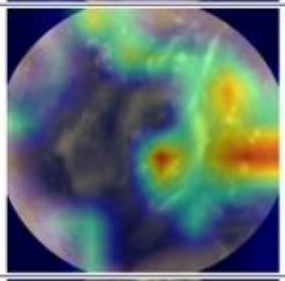
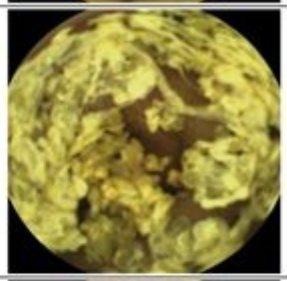
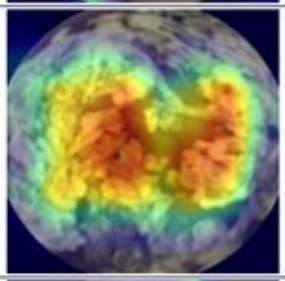
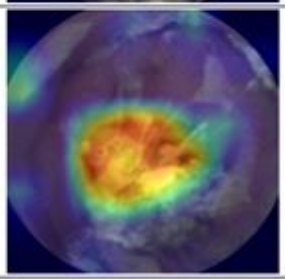
Image No.	Type	CE image in white light	Class Activation Map
JM1053_02_47_00_3	False positive		
JM3068_02_10_24_0	False positive		
JM3064_04_13_45_0	False positive		
JM5034_04_09_48_9	False positive		
JM3071_05_51_55_5	False negative		

Figure 4

Examples of Several Class Activation Maps of Misclassified Images.