A clinically interpretable convolutional neural network for the real time prediction of early squamous cell cancer of the esophagus; comparing diagnostic performance with a panel of expert European and Asian endoscopists

Introduction
Intrapapillary capillary loops (IPCLs) are microvascular structures that correlate with invasion depth of early squamous cell neoplasia (ESCN) and allow accurate prediction of histology. Artificial intelligence may improve human recognition of IPCL patterns and prediction of histology to allow prompt access to endoscopic therapy of ESCN where appropriate

Background and Aims
115 patients were recruited at two academic Taiwanese hospitals. ME-NBI videos of squamous mucosa were labelled as dysplastic or normal according to their histology and IPCL patterns classified by consensus of three experienced clinicians. A CNN was trained to classify IPCLs, using 67742 high quality ME-NBI by five-fold cross validation. Performance measures were calculated to give an average F1 score, accuracy, sensitivity and specificity. A panel of 5 Asian and 4 European experts predicted the histology of a random selection of 158 images using the JES IPCL classification – accuracy, sensitivity, specificity, positive and negative predictive values were calculated.

Results
Expert EU and Asian endoscopists attained F1 scores (a measure of binary classification accuracy) of 97.0% and 98% respectively. Sensitivity and accuracy of the EU and Asian clinicians was 97% and 98% and 96.9%, 97.1% respectively. The CNN average F1 score was 94%, sensitivity 93.7% and accuracy 91.7%. Our CNN operates at video rate and generates class activation maps that can be used to visually validate CNN predictions.

Conclusion

We report a clinically interpretable CNN developed to predict histology based on IPCL patterns, in real-time, using the largest reported dataset of images for this purpose. Our CNN achieved diagnostic performance comparable to an expert panel of endoscopists.

Background and aims

The application of artificial intelligence in diagnostic endoscopy, as well as other fields of medicine, is gathering pace. One such application of artificial intelligence is in the endoscopic diagnosis of early squamous cell neoplasia of the esophagus (ESCN). Esophageal cancer is the eighth most common cause of cancer worldwide; typically carrying a grim prognosis[1,2]. There is also a disproportionate geographic distribution of cases through Africa, the Middle East and into China and Japan[2,3].

Early detection and histologic diagnosis of ESCN is vital, in order to guide potentially curative therapy in patients who present prior to developing locally advanced or metastatic disease[4,5]. ESCN lesions confined to the mucosa exhibit low rates of local lymph node (LN) metastasis (<2%) compared to lesions which invade the submucosa (8-45.9%)[4,5]. As such these early lesions may be amenable to endoscopically delivered therapies such as endoscopic mucosal resection (EMR) or endoscopic submucosal dissection (ESD), both of which offer impressive 5 year survival[10–12].

Despite recent advances in endoscopic imaging technology early detection of ESCN remains a diagnostic challenge. ESCN lesions are subtle and can be easily missed during endoscopy; one study reported high miss rates for esophageal cancers on endoscopy undertaken in the three years preceding diagnosis[13–15]. These deficiencies are further compounded by variable clinician experience in assessing ESCN, as well as other human factors such as inattention and fatigue. A well validated endoscopic marker for the presence of squamous dysplasia and ESCN are intrapapillary capillary loops (IPCLs)[16,17]. IPCLs are microvessels first characterised on magnification endoscopy (ME)[18], which branch from the deep submucosal vessels and extend into the esophageal mucosa. As an ESCN develops there is a progressive distortion of IPCL patterns commensurate with the invasion depth of the lesion, hence the recognition of the changes in IPCL morphology allows clinicians to predict the histologic stage (**fig 1 and table 1**).

A number of classification systems for IPCL morphology and prediction of histology have been validated clinically [16,17,19]. For this study we used the Japanese Endoscopic Society (JES) IPCL classification; a recently developed, simplified system for classifying morphologic changes in IPCLs in order to predict histologic stage and invasion depth of ESCN[17]. As summarised in **table 1;** type A IPCLs correspond with normal mucosa or low grade dysplasia; type B1 with high grade dysplasia (HGD) or lamina propria invasion (LP); type B2 with invasion into the muscularis mucosa (MM) or first submucosal layer (SM1) and type B3 with invasion into the second submucosal layer (SM2) or beyond.

The JES classification is concise and has been widely adopted in centres that treat high volumes of ESCN. It offers high diagnostic accuracy and good interobserver agreement compared to other classifications – with an average 90.5% accuracy for predicting all neoplastic histology (type B1-B3 IPCLs). The accuracy of histology prediction was 91.9%, 93.4% and 95.9% for type B1, B2 and B3 IPCL patterns respectively[17]. Kim et al. also report excellent levels of interobserver agreement using the JES classification[20].

Computer-aided endoscopic diagnosis, using convolutional neural networks (CNNs) has the potential for use as an adjunct during endoscopy. CNNs used for this purpose typically need input data where specific visual features correspond with a classification. The stereotyped morphological changes in IPCL patterns seen with progressive dysplastic lesions, provide this data and so can be used to train a CNN. Repetitive training of the CNN allows the development of feature recognition, to allow it to make predictions on the histology of a lesion with increasing accuracy. A validated CNN, could provide a useful diagnostic adjunct for endoscopists assessing ESCN lesions, particularly in settings where experience or training may be limited.

We have previously reported a proof of concept study for the use of convolutional neural networks for the real time classification of ESCN lesions based on IPCL patterns[21–23]. Importantly our results were clinically interpretable, the use of class activation maps confirmed visually that the CNN was basing its classifications on IPCL patterns, as a clinician would during endoscopy. In this study we have expanded our dataset significantly in order to capture a wider spectrum of disease and variability in IPCL patterns; generating a CNN that can identify dysplastic oesophageal mucosa which is both clinically valid and interpretable. This study aims to build the foundation for more complex CNNs that can predict histologic invasion depth. Furthermore, little is known about the utility of the JES classification outside of an expert, high volume setting, predominantly Asian centres managing patients with ESCN. We therefore report a comparison between the diagnostic performance of European endoscopists, with Asian endoscopists and our convolutional neural network in predicting the histology of ESCN lesions based on their IPCL patterns.

Methods

Patient recruitment

Patients attending for endoscopic assessment at two ESCN referral centres in Taiwan were recruited (National Taiwan University Hospital and E-Da Hospital Kaosiung). In all included patients, pathological samples were acquired to confirm histologic diagnosis either by EMR, ESD or esophagectomy. Patients with active esophageal ulceration were excluded. Our study complied with the Declaration of Helsinki. The Institutional Review Board of E-Da Hospital approved this study (IRB number: EMRP-097-022. July 2017).

Endoscopic procedures and video acquisition

Gastroscopies were performed under conscious sedation or local anaesthesia by two expert endoscopists (WLW, HPW); an expert endoscopist was defined as a consultant gastroenterologist who has completed formal training and undertakes >50 ESCN assessments and resections per year. A solution of simethicone and water was applied to the esophageal mucosa prior to recording to remove mucus, food residue and blood and to facilitate clear visualisation of the esophageal mucosa and microvasculature. Endoscopies were performed using a high-definition magnification endoscope with narrow band imaging (HD ME-NBI) GIF-H260Z (Olympus, Japan), in combination with an Olympus Lucera CV-290 processor (Olympus, Japan).

Correlating imaged areas with histology

For patients with dysplastic lesions the endoscopist identified the lesion in overview, the lesion border was then marked by the endoscopist using cautery forceps, prior to resection or biopsies. Magnification was undertaken in NBI mode at 80-100x magnification in order to interrogate the IPCL patterns within the lesion border. IPCL patterns within the imaged area were then classified according to the JES IPCL classification system by consensus of three expert endoscopists (WLW, HPW, RJH). IPCLs were classified as type A, B1, B2 or B3 in order to predict the worst case histology for the entire lesion. Lesions with only type A IPCLs were classified as normal, those with >B1 lesions, dysplastic. Lesions were then resected by either endoscopic mucosal resection (EMR) or endoscopic submucosal dissection (ESD), formalin fixed and reviewed by a gastrointestinal histopathologist. The worst case lesion histology was then reported based on the worst histologic changes seen within the whole lesion. For patients with normal mucosa an area was selected, imaged under ME-NBI, IPCL patterns were classified. A forceps biopsy of the imaged area was then acquired to confirm the imaging findings.

Dataset overview

Video sequences from patients were assigned a class of normal or dysplastic. Frames were extracted at a rate of 30fps. Frames were individually quality controlled by a clinician with

experience in the endoscopic imaging of esophageal cancer (ME). Frames that were degraded by lighting or motion artefact, excessive mucus or blood were removed from the dataset where it was felt that it would not be possible for a clinician to make a decision based on IPCL patterns. Following quality control our total dataset comprised 114 patients (45 normal, 69 dysplastic). 67742 images were included in our dataset (28,078 normal and 39,662 dysplastic) with an average of 593 frames per patient. Five-fold cross validation was used. For each fold patients were randomly assigned into a training group, validation group (used for hyperparameter training) and testing group in the ratio of 80%, 10% and 10% respectively.

Convolutional neural network and class activation maps

ResNet-18 was chosen as the underlying model for our proposed CNN and was also used as a baseline for result comparison. Although state-of-the-art, ResNet-18 provides good classification performance, but is not a clinically interpretable model.

In our model[21], the fully connected layer is removed as described by Zhou et al[24]. The computation of the class score predictions is reformulated **(fig 2).** Such reformulation is mathematically equivalent to the solution proposed by Zhou et al, but produces class activation maps during the forward pass of the network. Our input images are 256x256, which leads to an 8x8 feature tensor at the output of the encoder; this low resolution does not permit clinical interpretability as the IPCL patterns which informed the classification cannot be identified. Hence, the feature demonstrated in (**fig 2)** is connected as a side output (connected to the loss) to all encoder resolutions[21] in order to produce class activation maps (CAMs). The entire workflow of our study is summarised in (**fig 3**).

Clinician classification of images and statistical analysis

A representative sample of 158 frames taken from all patients within the study were classified by a group of 9 expert endoscopists (5 Asian, 4 EU experts), none of whom were involved in the collection of videos. Diagnostic performance measures were calculated for individual endoscopists and for each geographical group. Accuracy, sensitivity and specificity were calculated; we report the average diagnostic performance for Asian endoscopists, EU endoscopists and our CNN. An F1 score, a measure of the diagnostic accuracy of binary classification algorithms, was also calculated. Interobserver agreement for clinicians was calculated using Krippendorf's alpha and assessed using a modified Likert scale (Landis and Koch).

We calculated the average per-frame diagnostic performance for our CNN (F1 score, accuracy, sensitivity and specificity) as well as an analysis of per-patient diagnostic performance. To establish a per-patient classification as normal or dysplastic, images were grouped by patient. Our network outputs a probability that each image shows dysplastic tissue, we set a threshold >0.5 as being positive that an image contains dysplasia. To establish a per-patient classification the probability outputs of each frame from that patient were used to compute an overall average probability of the presence of dysplasia. A threshold of >0.5 average probability was used to classify that patient as dysplastic and <0.5 as normal, CNN predictions were compared with the ground truth of histologic analysis.

Results

Patient characteristics

115 patients were included. 45 of these were determined to have a normal esophageal mucosa. 70 patients were determined to have dysplastic mucosa, with IPCL patterns ranging from type B1-B3 (**table 2**).

Comparative diagnostic performance of EU and Asian expert endoscopists

158 images were reviewed by a panel of 9 expert endoscopists, all blinded to the endoscopy procedure and histology results. Diagnostic performance of Asian and EU endoscopists were calculated respectively; F1 scores of 98% and 97%; accuracy of 97.1% and 96.9%; sensitivity of 96.9% and 98.9% and specificity of 97.6% and 91.5%. The pooled diagnostic performance of all expert endoscopists for F1 score, accuracy, sensitivity and specificity was 96.5%, 94.7%, 97% and 88% respectively. The interobserver agreement of all endoscopists was regarded as substantial with a Krippendorf's alpha of 76.7%. Performance measures are summarised in **table 3 .** Overall we demonstrate diagnostic performance using the JES classification system that either exceeds or is comparable to other reported work. Asian endoscopists had significantly higher specificity than EU endoscopists (97.6% vs 91.5% p=0.01) whereas EU endoscopists had significantly higher sensitivity than their Asian counterparts (98.9% vs 96.9% p=0.01).

Diagnostic performance of convolutional neural network

We report both per-frame analysis and per-patient analysis. On a per-frame basis the diagnostic performance results of our CNN are summarised in **table 4**. Across all folds we report an average F1 score, accuracy, sensitivity and specificity of 94.0%, 91.7%, 93.7% and 92.4% respectively. The AUC of our system is 95.8%.

We also assessed the per patient diagnostic performance of our CNN. Using five-fold cross validation we used 12 patients per fold for testing, using five folds gave a total of 60 independent patients that were used to test the CNN classification. A true positive was determined if a patient with dysplasia was classified as such in >50% of the CNN predictions for images from that patient. We observed that for this CNN iteration our system failed to classify correctly overall in only one patient (patient 158), reporting a false positive result in two independent folds. A receiver operating characteristic curve demonstrates an AUC of 0.96, suggesting a high level of diagnostic accuracy **(fig 4)**. **Figure 5** illustrates, for each fold, the images that were classified with highest probability by our network as either true positive (TP), true negative (TN), false positive (FP) or false negative (FN) and reflects how our network performs with diagnostically challenging images.

Class activation maps and clinical interpretability

Our network is able to classify, in real-time IPCL patterns as normal or abnormal. We demonstrate representative examples of class activation maps **(fig 6).** These are

representations of what the CNN 'sees' when it classifies a frame of an endoscopy video. The output CAM is clinically interpretable, highlighting to the clinician which areas of IPCLs informed the classification.

Discussion

We report a comparison in diagnostic performance between a cohort of expert endoscopists based in Europe and Asia, with a clinically interpretable CNN designed to predict the histology of early squamous cell neoplasia of the esophagus based on IPCL patterns.

Early identification and assessment of ESCN stage is vital, since lesions confined to the mucosa have low rates of metastasis to local lymph nodes and so may be curatively resected endoscopically. As discussed, ESCN lesions are subtle and can be easily missed; to aid clinicians several classification systems exist to assist the identification and characterisation of ESCN based on IPCL morphology[16,17,19]. To our knowledge almost all studies assessing the use of IPCL classifications focus on clinicians within high-volume referral centres, predominantly in Asia; little is known about the diagnostic performance in Western healthcare settings where clinician experience with ESCN is likely to be reduced.

Artificial intelligence, using CNNs, may offer an adjunct to improve clinician recognition of IPCL patterns and thereby improve their identification of ESCN. Furthermore a validated system may improve triage of lesions that are either normal, require endoscopic resection, or that are not amenable to endoscopic treatment and require either surgical intervention or palliation. Such a system could improve the speed at which patients with ESCN receive the most appropriate therapy, as well as reduce the burden on histopathology services of processing normal biopsies.

We have previously reported a proof of concept study that outlines the use of a CNN to classify esophageal mucosa as squamous or dysplastic[22,23]. This study aimed to further develop this CNN using an expanded dataset with greater variability of IPCL patterns. The diagnostic performance of our CNN is promising; demonstrating accuracy for the prediction of dysplasia of 91.7%, an F1 score of 94% and sensitivity of 93.7%. This compares favourably with an analysis of the diagnostic performance of expert clinicians using the JES classification by Oyama et al., which reported accuracy of 91.9 - 95.9% and sensitivities of 55% - 97.5%. Our CNN is able to classify images at video rate and was trained using consecutive, segmented frames from endoscopy videos, so has the potential for real-time use. Given that the use of CNNs in diagnostic endoscopy for this purpose is in its infancy there are few reported studies in the literature. Guo et al. propose a CNN capable of classifying dysplastic compared to normal tissue. Using a dataset of 6671 images they report sensitivities of 98%. As noted in our previous studies, while promising, relatively small datasets such as this may struggle to achieve such diagnostic performance when trained and tested on larger, more variable datasets[25]. Similarly Zhao et al. report a smaller dataset of 1383 images, with accuracies of 87% for the detection of dysplastic IPCL patterns[26]. We note that this dataset was heavily skewed towards images containing type B1 IPCL patterns, which may again affect how well the CNN can generalise. The CNN we propose has been tested and trained on a larger dataset, with a balance of IPCL subtypes and maintains a high level of diagnostic accuracy.

In order to assess the diagnostic performance of our CNN with clinicians, we assessed the performance of a cohort of Asian endoscopists with EU endoscopists on a representative sample of images from the same patients. Our endoscopists achieved high diagnostic accuracy using the JES system. The Asian endoscopists achieved an accuracy of 97.1%, F1 score 98%, sensitivity of 96.9% and specificity of 97.6%. This was comparable with EU endoscopists who achieved an accuracy of 96.9%, F1 score 97%, sensitivity of 98.9% and specificity of 91.5%. The higher sensitivity of EU endoscopists for identifying dysplasia is likely related to them over-diagnosing dysplasia, as evidenced by the lower specificity. Our CNN achieves lower but comparable diagnostic performance when compared with the expert endoscopists in this study, we note that in some instances it performed better than some individual clinicians and that the diagnostic performance of our expert panel was higher than that reported in other studies using both the JES and other IPCL classifications. We suggest that our CNN shows promise and with further development could provide a useful diagnostic adjunct to clinicians involved in the endoscopic management of ESCN. Given that this study looked at expert clinicians we propose that compared to less experienced endoscopists our CNN may offer a further improvement in diagnostic accuracy.

Although we used the JES IPCL classification system for this study, consideration should also be given to the performance of our CNN relative to other studies using a range of diagnostic classifications. Oyama et al. demonstrated an overall accuracy for histology prediction of 90.5% using the JES classification[17]. In a retrospective analysis of patients with histologically confirmed ESCN who underwent ME-NBI assessment, Mizumoto et al. report diagnostic accuracy of 82% for the differentiation of lesions superficial to the LPM compared to those invading deeper than the MM using the JES classification[27]. Kim et al. report an overall accuracy for identifying dysplastic lesions using the JES system of 78.6%[20]. In our study, both the EU and Asian clinicians exceed this, with an overall diagnostic accuracy of 94.7 [CI: 83.9-99.7]. Our CNN also demonstrated higher average accuracy of 91.7% for the prediction of dysplastic tissue.

We recognise that for a CNN used for this application to have clinical utility it must be interpretable in real-time. To facilitate this we selected ResNet-18 as the underlying model for our proposed CNN. ResNet-18 provides good classification performance, but lacks an intrinsic mechanism that highlights what image features inform the class prediction estimated for a given input image. Class activation maps (CAMs) highlight areas of the input image that are considered to inform the class prediction (normal or dysplastic) by the CNN.

The CAMs generated by our CNN serve two functions. First, they provide a key safety feature; by demonstrating informative features in the input data, we can ensure that successful predictions are not based on spurious features such as reflections, lighting or other features that may be biased in our dataset. Secondly CAMs may also facilitate the identification of new image features and endoscopic markers for ESCN that might exist but are not yet recognised by clinicians. This CNN model falls into a growing body of work which aims to produce CNNs that are able to both classify/predict and also explain the rational for their predictions in an interpretable manner.

Further work should focus on training CNNs with datasets of increasing size and variability. While this remains in the future, in order to be used in clinical settings CNNs need to be rigorously validated against clinicians; although our CNN operates at video rate, further work will be needed to make it usable in real-time endoscopic assessment with unprocessed videos. The logical extension of this CNN will be to develop a system capable of predicting histologic invasion depth of lesions based on IPCL patterns; such a system, if carefully validated, could be used to undertake 'virtual biopsies' and speed up triage to appropriate therapy of patients presenting with either normal or abnormal mucosa. We also envisage that a developed version of this system could be used to improve detection of abnormal oesophageal mucosa and thereby reduce the miss rate of oesophageal cancer on endoscopy. We suggest that further studies should aim to characterise the diagnostic performance of non-expert endoscopists based in high and low volume centres for the treatment of ESCN. This could provide a benchmark against which to assess the potential utility of our CNN as an adjunct to endoscopic assessment of ESCN by non-experts. A clinically validated

Word count 3398

Acknowledgements

This work was supported by Wellcome Trust [WT101957, 203145Z/16/Z], EPSRC [NS/A000027/1, NS/A000050/1], NIHR BRC UCLH/UCL High Impact Initiative and a UCL EPSRC CDT Scholarship Award [EP/L016478/1]. The authors would like to thank NVIDIA Corporation for the donated GeForce GTX TITAN X GPU.

The authors would like to thank Dr David Graham, Dr Vinay Seghal, Dr Mohamed Hussein, Dr Yezen Sammaraiee, Dr Stephen Mitrasinovic for their contributions to this study and manuscript

References

1. Eslick GD. Epidemiology of Esophageal Cancer. *Gastroenterol Clin North Am*. 2009;38(1):17-25. doi:10.1016/j.gtc.2009.01.008

2. Zhang Y. Epidemiology of esophageal cancer. *World J Gastroenterol WJG*. 2013;19(34):5598-5606. doi:10.3748/wjg.v19.i34.5598

3. Epidemiology of Esophageal Cancer in Japan and China. *J Epidemiol*. 2013;23(4):233-242. doi:10.2188/jea.JE20120162

4. Nagami Y, Ominami M, Shiba M, et al. The five-year survival rate after endoscopic submucosal dissection for superficial esophageal squamous cell neoplasia. *Dig Liver Dis*. 2017;49(4):427-433.

5. Ono S, Fujishiro M, Niimi K, et al. Long-term outcomes of endoscopic submucosal dissection for superficial esophageal squamous cell neoplasms. *Gastrointest Endosc*. 2009;70(5):860-866.

6. Cho JW, Choi SC, Jang JY, et al. Lymph Node Metastases in Esophageal Carcinoma: An Endoscopist's View. *Clin Endosc*. 2014;47(6):523-529. doi:10.5946/ce.2014.47.6.523

7. Bollschweiler E, Baldus SE, Schröder W, et al. High rate of lymph-node metastasis in submucosal esophageal squamous-cell carcinomas and adenocarcinomas. *Endoscopy*. 2006;38(02):149-156.

8. Shimada H, Okazumi S, Matsubara H, et al. Impact of the Number and Extent of Positive Lymph Nodes in 200 Patients with Thoracic Esophageal Squamous Cell Carcinoma after Three-field Lymph Node Dissection. *World J Surg*. 2006;30(8):1441-1449. doi:10.1007/s00268-005-0462-6

9. Sgourakis G, Gockel I, Lyros O, Hansen T, Mildenberger P, Lang H. Detection of lymph node metastases in esophageal cancer. *Expert Rev Anticancer Ther*. 2011;11(4):601-612. doi:10.1586/era.10.150

10. Shimizu Y, Tsukagoshi H, Fujita M, Hosokawa M, Kato M, Asaka M. Long-term outcome after endoscopic mucosal resection in patients with esophageal squamous cell carcinoma invading the muscularis mucosae or deeper. *Gastrointest Endosc*. 2018;56(3):387-390. doi:10.1016/S0016-5107(02)70043-6

11. Inoue H, Minami H, Kaga M, Sato Y, Kudo S. Endoscopic Mucosal Resection and Endoscopic Submucosal Dissection for Esophageal Dysplasia and Carcinoma. *Gastrointest Endosc Clin N Am*. 2010;20(1):25-34. doi:10.1016/j.giec.2009.08.005

12. Oyama T, Tomori A, Hotta K, et al. Endoscopic Submucosal Dissection of Early Esophageal Cancer. *Clin Gastroenterol Hepatol*. 2018;3(7):S67-S70. doi:10.1016/S1542-3565(05)00291-0

13. Menon S, Trudgill N. How commonly is upper gastrointestinal cancer missed at endoscopy? A meta-analysis. *Endosc Int Open*. 2014;2(2):E46-E50. doi:10.1055/s-0034-1365524

14. Cheung D, Menon S, Trudgill N. OC-012 How Commonly is Oesophageal Cancer Missed at Endoscopy (A UK Primary Care Based Study)? *Gut*. 2013;62(Suppl 1):A5 LP-A6.

15. Cheung D, Menon S, Hoare J, Dhar A, Trudgill N. Factors associated with upper gastrointestinal cancer occurrence after endoscopy that did not diagnose cancer. *Dig Dis Sci*. 2016;61(9):2674-2684.

16. Inoue H, Kaga M, Ikeda H, et al. Magnification endoscopy in esophageal squamous cell carcinoma: a review of the intrapapillary capillary loop classification. *Ann Gastroenterol Q Publ Hell Soc Gastroenterol*. 2015;28(1):41-48.

17. Oyama T, Inoue H, Arima M, et al. Prediction of the invasion depth of superficial squamous cell carcinoma based on microvessel morphology: magnifying endoscopic classification of the Japan Esophageal Society. *Esophagus*. 2017;14(2):105-112. doi:10.1007/s10388-016-0527-7

18. INOUE H, HONDA T, NAGAI K, et al. Ultra-high magnification endoscopic observation of carcinoma in situ of the esophagus. *Dig Endosc*. 1997;9(1):16-18.

19. Arima M, Tada M, Arima H. Evaluation of microvascular patterns of superficial esophageal cancers by magnifying endoscopy. *Esophagus*. 2005;2(4):191-197.

20. Kim SJ, Kim GH, Lee MW, et al. New magnifying endoscopic classification for superficial esophageal squamous cell carcinoma. *World J Gastroenterol*. 2017;23(24):4416-4421. doi:10.3748/wjg.v23.i24.4416

21. García-Peraza-Herrera LC, Everson M, Lovat L, et al. Intrapapillary capillary loop classification in magnification endoscopy: open dataset and baseline methodology. *Int J Comput Assist Radiol Surg*. 2020;15(4):651-659. doi:10.1007/s11548-020-02127-w

22. Everson M, Herrera LGP, Li W, et al. Artificial intelligence for the real-time classification of intrapapillary capillary loop patterns in the endoscopic diagnosis of early oesophageal squamous cell carcinoma: A proof-of-concept study. *United Eur Gastroenterol J*. 2019;7(2):297-306.

23. Garcia-Peraza-Herrera L, Everson M, Li W, et al. Interpretable Fully Convolutional Classification of Intrapapillary Capillary Loops for Real-Time Detection of Early Squamous Neoplasia. *Arxiv*. Published online 2018. https://arxiv.org/abs/1805.00632

24. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2016:2921-2929.

25. Guo L, Xiao X, Wu C, et al. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). *Gastrointest Endosc*. Published online 2019.

26. Zhao Y-Y, Xue D-X, Wang Y-L, et al. Computer-assisted diagnosis of early esophageal squamous cell carcinoma using narrow-band imaging magnifying endoscopy. *Endoscopy*. 2019;51(04):333-341.

27. Mizumoto T, Hiyama T, Oka S, et al. Diagnosis of superficial esophageal squamous cell carcinoma invasion depth before endoscopic submucosal dissection. *Dis Esophagus*. 2018;31(7):dox142-dox142.

IPCL pattern	Neoplastic	Typical histology
A	No	Normal/LGD
B1	Yes	HGD/LP
B2	Yes	MM/SM1
B3	Yes	SM2 or deeper

Table 1: Summary of the JES classification of IPCL patterns

Patient histology	Number of patients in dataset
Normal	45
HGD/LPM	35
MM/SM1	17
>SM2	18

Table 2: Breakdown of patient numbers recruited to the study by histologic stage of ESCN

Endoscopist	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 score (%)
EU	96.9	98.9	91.5	97.0
Asian	97.1	96.9	97.6	98.0
Pooled average	94.7	97.0	88.0	96.5

Table 3: Summary of expert endoscopist performance statistics for detection of abnormal IPCL patterns

Fold	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 score (%)
1	92.5	99.6	81.3	94.1
2	92.4	91.3	95.1	94.4
3	97.4	98.3	96.6	97.0
4	94.5	98.9	89.3	95.1
5	81.9	80.5	99.8	89.2
Average	91.7	93.7	92.4	94.0

Table 4: Summary of CNN performance statistics for per frame detection of abnormal IPCL patterns

Figure legends

Figure 1: Representative ME-NBI images used in this study of different IPCL patterns seen in each of the JES (Japanese Endoscopic Society) subtypes. Green arrows show normal IPCLs. Red circumscribed areas denote abnormal IPCL patterns.

Figure 2: Summary of the CNN side output incorporated to allow clinical interpretability of class activation maps (CAMs).

Figure 3: Schematic representation of study workflow

Figure 4: ROC curve for the diagnostic performance of our CNN (adapted from [21])

Figure 5: Images the CNN classified with the highest certainty for each fold, that were subsequently identified as true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for dysplasia).

Figure 6: Representative images of normal and abnormal IPCL patterns seen at endoscopy (left) with the corresponding CAMs generated by our CNN (right).