International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-020-02127-w

ORIGINAL ARTICLE

Intrapapillary capillary loop classification in magnification endoscopy: open dataset and baseline methodology

Luis C. García-Peraza-Herrera[1,7] · Martin Everson[2,3] · Laurence Lovat[2,3] · Hsiu-Po Wang[4] · Wen Lun Wang[5] · Rehan Haidry[2,3] · Danail Stoyanov[6] · Sébastien Ourselin[7] · Tom Vercauteren[7]

Received: 21 January 2020 / Accepted: 17 February 2020

Abstract

Purpose Early squamous cell neoplasia (ESCN) in the oesophagus is a highly treatable condition. Lesions confined to the mucosal layer can be curatively treated endoscopically. We build a computer-assisted detection system that can classify still images or video frames as normal or abnormal with high diagnostic accuracy.

Methods We present a new benchmark dataset containing 68K binary labelled frames extracted from 114 patient videos whose imaged areas have been resected and correlated to histopathology. Our novel convolutional network architecture solves the binary classification task and *explains* what features of the input domain drive the decision-making process of the network.

Results The proposed method achieved an average accuracy of 91.7% compared to the 94.7% achieved by a group of 12 senior clinicians. Our novel network architecture produces deeply supervised activation heatmaps that suggest the network is looking at intrapapillary capillary loop patterns when predicting abnormality.

Conclusion We believe that this dataset and baseline method may serve as a reference for future benchmarks on both video frame classification and explainability in the context of ESCN detection. A future work path of high clinical relevance is the extension of the classification to ESCN types.

Keywords Early squamous cell neoplasia (ESCN) · Intrapapillary capillary loop (IPCL) · Class activation map (CAM)

Introduction

Oesophageal cancer is the sixth most common cause of cancer deaths worldwide [16] and a burgeoning health issue in developing nations from Africa along a 'cancer belt' to China. The current gold standard to investigate oesophageal cancer is gastroscopy with biopsies for histological analysis. Early squamous cell neoplasia (ESCN) is a highly treatable type of oesophageal cancer, with recent advances in endoscopic therapy meaning that lesions confined to the mucosal layer can be curatively resected endoscopically with a $< 2\%$ incidence of local lymph node metastasis [1]. The endoscopic appearances of ESCN lesions are subtle and easily missed, with significant miss rates on endoscopy within the 3 years preceding diagnosis [10]. Early cancers invading into the submucosa are likely to have local lymph node metastasis

This work was supported through an Innovative Engineering for Health award by Wellcome Trust (WT101957); Engineering and Physical Sciences Research Council (EPSRC) (NS/A00027/1) and a Wellcome/EPSRC Centre award (203145Z/16/Z and NS/A000050/1).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11548-020-02127-w) contains supplementary material, which is available to authorized users.

✉ Luis C. García-Peraza-Herrera
luis.herrera.14@ucl.ac.uk

1 Department of Medical Physics and Biomedical Engineering, UCL, London, UK

2 Division of Surgery and Interventional Science, UCL, London, UK

3 Department of Gastroenterology, University College Hospital NHS Foundation Trust, London, UK

4 Department of Internal Medicine, National Taiwan University, Taipei, Taiwan

5 Department of Internal Medicine, E-Da Hospital/I-Shou University, Kaohsiung, Taiwan

6 Wellcome/EPSRC Centre for Interventional and Surgical Sciences, UCL, London, UK

7 School of Biomedical Engineering and Imaging Science, KCL, London, UK

Published online: 12 March 2020

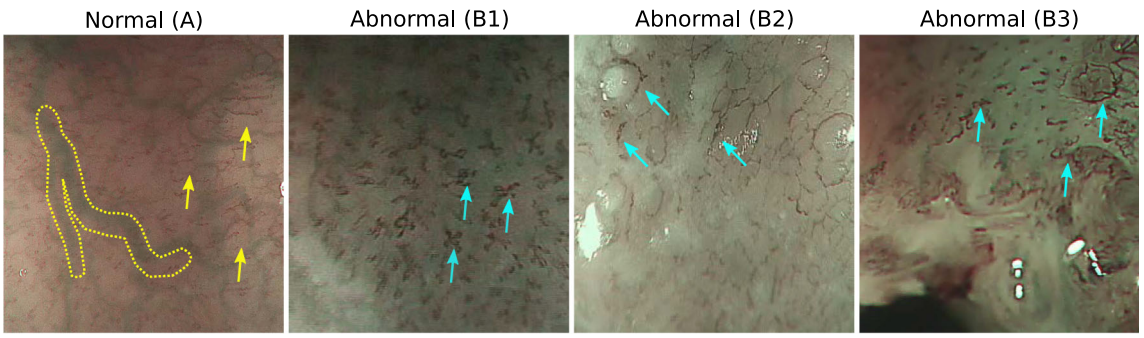

Fig. 1 Magnifying endoscopy (ME) frames extracted from videos of patients with different histopathology. Normal patients typically present a clear deep submucosal vasculature, large green-like vessels such as the one highlighted within the dashed yellow line are usually visible. Intrapapillary capillary loops (IPCLs) refer to the microvasculature (pointed by the arrows). Healthy patients tend to present thinner (yellow arrows) and less tangled IPCL patterns than those with abnormal tissue (blue arrows)

and should be referred promptly for consideration of surgical resection.

Intrapapillary capillary loops (IPCL) are a clinical microvascular feature recognised as an endoscopic marker for ESCN [7,12,13]. They have been classified by the Japanese Endoscopic Society (JES) in a simplified system aimed at improving the easy recognition of ESCN by endoscopists [12]. The type of IPCLs present also facilitates the accurate prediction of the lesion histology; Type A IPCLs (see Fig. 1) correlate with normal tissue. Type B1, B2, B3 IPCLs (see Fig. 1) demonstrate progressive morphologic abnormalities and correlate with the invasion of early neoplasia in the muscularis mucosa and submucosal tissue. Oyama et al. [12] demonstrate that the JES classification offers high diagnostic accuracy compared to other classifications for the prediction of dysplastic tissue—with the overall accuracy for histology prediction 90.5% across type B1-3. A computer-assisted detection (CADe) system that can classify still images or video frames as normal or abnormal with high diagnostic accuracy could provide a useful adjunct to both expert and inexpert endoscopists.

Contributions

We focus on the problem of classifying video frames as normal/abnormal. These frames are extracted from the magnification endoscopy (ME) recording of a patient. To the best of our knowledge, we introduce the first IPCL normal/abnormal open dataset[1] containing ME video sequences correlated with histopathology. Our dataset contains 68K video frames from 114 patients.

For a small and representative sample of 158 frames (IPCL types A, B1, B2, B3), we ask 12 senior clinicians to label them as normal/abnormal and report the inter-rater agreement as Krippendorff's α coefficient [8], achieving 76.6%. We also draw a comparison between raters and our *gold standard* histopathology results, achieving an average accuracy across raters of 94.7%.

We propose a novel convolutional network (CNN) architecture to solve the binary classification task with a particular focus on the explainability of predictions. Our proposed method achieved an average accuracy of 91.7%. In addition to a global classification estimation, our novel design produces activation maps and class scores at every resolution of the convolutional pyramid. The network has to *explain* where it is *looking at* prior to the generation of a class prediction. Looking at the activation maps for the abnormal class, we have observed that the network is *looking at* IPCL patterns when predicting abnormality. No conclusive evidence has been found that it is paying attention to large deep submucosal vessels to detect normal tissue. We believe that this baseline method may serve as a reference for future benchmarks on both video frame classification and explainability in the context of ESCN detection.

Related work

Computer-aided endoscopic detection and diagnosis could offer an adjunct in the endoscopic assessment of ESCN lesions; there has been a high level of interest in recent years in developing clinically interpretable models. The use of CNNs has shown potential across several medical specialties. In gastroenterology, considerable efforts have been devoted to the detection of malignant colorectal polyps [5,14,15] and upper gastrointestinal cancer [9]. However, its utility in endoscopic diagnosis of early oesophageal neoplasia remains in its infancy [2].

Guo et al. [4] propose a CNN that can classify images as dysplastic or non-dysplastic. Using a dataset of 6671 images, they demonstrate per-frame sensitivity of 98% for the detection of ESCN. Using a video dataset of 20 videos, they demonstrate per-frame sensitivity of 96% for the detection

[1] https://github.com/luiscarlosgph/ipcl.

of ESCN. Although the results are encouraging, the size of the patient sample is limited. Given the *black box* nature of CNNs this may represent a matter of concern with regards to generalization capability. Zhao et al. [17] have also reported a CNN for the classification of IPCL patterns in order to identify ESCN. Using 1383 images, although heavily skewed towards Type B1 IPCLs, they demonstrated overall accuracies of 87% for the classification of IPCL patterns. In this study, however the authors excluded type B3 IPCLs from the training and testing phase. The CNN also demonstrated only a 71% classification rate for normal IPCLs, indicating that it over-diagnoses normal tissue as containing type B1 IPCLs, and so representing dysplastic tissue.

Dataset details

This dataset will be made publicly available online upon publication and can thus serve as a benchmark for future work on detection of ESCN based on magnification endoscopy images.

Patient recruitment, endoscopic procedures and video acquisition

Patients attending for endoscopic assessment to two early squamous cell neoplasia (ESCN) referral centres in Taiwan (National Taiwan University Hospital and E-Da Hospital) were recruited with consent. Patients with oesophageal ulceration, active oesophageal bleeding or Barrett's oesophagus were excluded. Gastroscopies were performed by two expert endoscopists (WLW, HPW), either under conscious sedation or local anaesthesia. An expert endoscopist was defined as a consultant gastroenterologist performing $>$ 50 early squamous cell neoplasia (ESCN) assessments per year. All endoscopies were performed using an HD ME-NBI GIF-H260Z endoscope, with Olympus Lucera CV-290 processor (Olympus, Tokyo, Japan). A solution of water of simethicone was applied via the endoscope working channel to the oesophageal mucosa, in order to remove mucus, food residue or blood. This allowed good visualization of the oesophageal mucosa and microvasculature, including IPCLs.

Correlating imaged areas with histology

Initially, a macroscopic assessment was made of the suspected lesion in an overview, with the borders of the lesion delineated by the endoscopist. The endoscopist then identified areas within the borders of the lesion on which to undertake magnification endoscopy. The IPCL patterns were interrogated using magnification endoscopy in combination with narrow-band imaging (ME-NBI). Magnification endoscopy was performed on areas of interest at $80-100$x magnification. Using the JES IPCL classification system, the IPCL patterns were classified by the consensus of three expert endoscopists (WW, HPW, RJH) as type A, B1, B2, B3, in order to give a prediction of the *worst-case* histology for the whole lesion. The entire lesion was then resected by either endoscopic mucosal resection (EMR) or endoscopic submucosal dissection (ESD). Resected specimens were formalin-fixed and assessed by an expert gastrointestinal histopathologist. As is the gold standard the *worst-case* histology was reported for the lesion as a whole, based on pathological changes seen within the resected specimen. Similarly to abnormal lesion areas, type A recordings (normal, healthy patients) were obtained by visual identification of healthy areas, magnification endoscopy, visual confirmation of normal vasculature and IPCL patterns, and biopsy to confirm the assessment.

Dataset description

Our IPCL dataset comprises a total of 114 patients (45 normal, 69 abnormal). Every patient has a ME-NBI video (30fps) recorded following protocol in "Correlating imaged areas with histology" section. Raw videos can present some parts where NBI is active. In this dataset, only ME subsequences are considered. All frames are extracted and assigned to the class *normal* or *abnormal* depending on the histopathology of the patient. They are quality controlled one-by-one (running twice over all the frames) by a senior clinician with experience in the endoscopic imaging of oesophageal cancer. Frames that are highly degraded due to lighting artifacts (e.g. blur, flares and reflections) up to the point where it is not possible (for the senior clinician) to make a visual judgement of whether they are normal or abnormal are marked as uninformative and not used. This curation process results in a dataset of 67742 annotated frames (28,078 normal, 39,662 abnormal) with an average of 593 frames per patient. For each fold, patients (not frames) are randomly split into 80% training, 10% validation (used for hyperparameter tuning), and 10% testing (used for evaluation). The statistics of each individual fold are presented in the supplementary material.

Evaluation per patient clip

Let $\{\hat{y}_{f,p}\}_{f=1}^{F_p}$ be the set of estimated probabilities for the frames f (out of F_p) belonging to patient clip p. Then, the estimated probability of abnormality for p is computed as an average of frame probabilities:

$$P\left(X = \text{abnormal} \middle| \{\hat{y}_{f,p}\}_{f=1}^{F_p}\right) = \frac{1}{F_p}\sum_{f=1}^{F_p}\hat{y}_{f,p} \tag{1}$$

Similarly to frame predictions, a threshold ($p = 0.5$) is applied to obtain a class label for p. As per our data collection protocol (see "Correlating imaged areas with histology" section), magnification endoscopy clips contain either normal or abnormal tissue. Hence, a correlation between $P(X = \text{abnormal}|\{\hat{y}_{f,p}\}_{f=1}^{F_p})$ and histopathology is expected. The analysis of clip classification errors facilitates the identification of worst cases, singling out patient-wide mistakes from negligible frame prediction errors.

Methods

In this section, we propose a reference method for IPCL binary classification with a particular focus on explainability that may serve as a baseline for future benchmarks. As it is common in data-driven classification, we aim to solve for a mapping f such that $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx \boldsymbol{y}$, where $\boldsymbol{x}$ is an input image, $\boldsymbol{y}$ the class label corresponding to $\boldsymbol{x}$, and $\boldsymbol{\theta}$ a vector of parameters. All the input images were preprocessed by downscaling them to a width of 256 pixels (height automatically computed from their original aspect ratio) so that we could fit a large batch of images into the GPU. To account for changes in viewpoint due to endoscope motion, random ($p = 0.5$) on-the-fly flips are applied to each image. Our baseline model is ResNet-18 [6]. The batch normalization moving average fraction is set to 0.7. Our batch size, momentum and weight decay hyperparameters are set to 256, 0.9, and 0.0005, respectively. The initial learning rate (LR) was tuned by grid search. It was set to $\lambda = 5e{-}3$ for training all folds, decaying it every 10K iterations ($\approx$ 40 epochs) by a factor of 0.5 until 45K iterations ($\approx$ 200 epochs). In our implementation, using an NVIDIA GeForce TITAN X Pascal GPU and Caffe 1.0 as deep learning framework, the inference time per-frame is 7.6ms [6.4ms, 9.9ms], enabling the algorithm for deployment as a real-time endoscopy solution.

Explaining network predictions, baseline without FC layer: ResNet-18-CAM

Explaining network predictions is of particular interest to draw a comparison between image features that clinicians employ in their clinical practice and those that might exist but be unknown to them. Conversely, adding attention to those image features that are known to be relevant but are not used by the network could potentially improve its performance. In the context of ESCN detection, this leads to investigate whether the network is actually looking at deep submucosal vessels and IPCL patterns to predict abnormality. The answer to this question typically comes in the form of a heatmap, with those parts relevant to the classification being highlighted.

Our baseline model (ResNet-18) may be formalized as $f_{\boldsymbol{\theta}} = r(h(g(T_{\boldsymbol{x}})))$ where $T_{\boldsymbol{x}} = T_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}^{H\times W\times K}$ is the feature tensor obtained after processing $\boldsymbol{x}$ at the deepest pipeline resolution, K represents the number of feature channels, $T_{\boldsymbol{x}}(k)$ is a matrix that represents the feature channel with index k, and g, h, and r represent the global average pooling (GAP), fully connected (FC), and final scoring convolution layers, respectively.

The FC layer h represents a challenge for explainability, as relevance is redistributed when gradients flow backwards, losing its spatial connection to the prediction being made [11]. Hence, inspired by [18], we stripped out the fully connected layer of 1000 neurons from the baseline model (ResNet-18), connecting the output of the GAP directly to the neurons that predict class score (those in layer r) and setting their bias to zero. Formally, this leads to $f_{\boldsymbol{\theta}} = r(g(T_{\boldsymbol{x}}))$, the output of the network before softmax being

$$\hat{\boldsymbol{y}}^{(c)} = \sum_{k\in K} w_{k,c} \left[\underbrace{\frac{1}{HW}\sum_{i,j}}_{\text{GAP}} \underbrace{T_{\boldsymbol{x}}(k)}_{\text{Feature tensor}} \right] \tag{2}$$

where $w_{k,c} \in \hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{y}}^{(c)}$ is the score predicted for class c. Following this approach, a heatmap per class can be generated obviating the GAP layer during inference, simply computing

$$\hat{\boldsymbol{y}}^{(c)}_{\text{CAM}} = \sum_{k\in K} w_{k,c} T_{\boldsymbol{x}}(k) \tag{3}$$

These heatmaps called class activation maps (CAMs) [18] keep a direct spatial relationship to the input, which is relevant for visual explanations. Although the architecture proposed in [18] requires removing the GAP layer to produce the CAMs, (2) can be reformulated as

$$\hat{\boldsymbol{y}}^{(c)} = \underbrace{\frac{1}{HW}\sum_{i,j}}_{\text{GAP}} \underbrace{\left[\sum_{k\in K} w_{k,c} T_{\boldsymbol{x}}(k) \right]}_{\text{CAM}} \tag{4}$$

in which case the CAMs are embedded within the network pipeline as a 1×1 convolution (as we have already shown in [3]). This leads to $f_{\boldsymbol{\theta}} = g(r(T_{\boldsymbol{x}}))$. We refer to this architecture as ResNet-18-CAM (as for the baseline, LR is set to $5e{-}3$ and decayed by 0.5 every 10K iterations until 45K iterations). The performance of this network is shown in Table 2. Although the accuracy of ResNet-18-CAM is comparable to the baseline network (ResNet-18), ResNet-18-CAM conveniently computes a heatmap per class as part of the network processing. However, the explainability in the context of our

classification problem remains very challenging due to the low resolution of the heatmaps produced.

Deeply supervised class activation maps: ResNet-18-CAM-DS

In the computer vision field, images tend to display one or a few large objects. This is, however, not the case in medical images such as the magnification endoscopy ones used to classify IPCL patterns. Due to their low resolution, it is very challenging to understand what the network is *looking at*, as abnormal microvasculature in an endoscopic image is not localized only in a single spot. In our clinical problem, two types of features could be expected to be highlighted, submucosal vessels and IPCLs, which represent endoscopic markers for ESCN [7,12,13]. The procedure to generate the CAM proposed in [18] employs the deepest feature maps as inputs to produce the attention heatmaps. For our input images of 256×256 pixels, these feature maps have a resolution of 8×8 pixels, leading to very low-resolution CAMs (also 8×8 pixels). This hinders the explanatory capability of the heatmaps, as small capillaries are the main clinically discriminating feature. It is of interest to know whether they are being *looked at* to predict abnormality. A trivial solution would be to reduce the depth of the network, but this could potentially hamper the learning of abstract features and decrease performance. In addition, the optimal amount of resolution levels for the given task to balance accuracy and explainability is a hyperparameter that would need to be tuned. Instead, we propose an alternative path modelling $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = (f_{\boldsymbol{\theta}_t} \circ f_{\boldsymbol{\theta}_{t-1}} \circ \cdots \circ f_{\boldsymbol{\theta}_2} \circ f_{\boldsymbol{\theta}_1})(\boldsymbol{x}) \tag{5}$$

where $f_{\boldsymbol{\theta}_t}$ represents the function that processes the input at resolution t, and whose output tensor has a width and height downsampled (strided convolution) by a factor of 0.5 with regards to its input tensor. In this formulation, given an $\boldsymbol{x}$ of size 256×256 pixels and $t = 5$, the output of $f_{\boldsymbol{\theta}_5}$ is 8×8 pixels.

Given (5), let $T_{\boldsymbol{x},t}$ be the output tensor produced by $f_{\boldsymbol{\theta}_t}$. Then, similarly to (4), we propose to generate a class score prediction at each resolution t as follows

$$\hat{\boldsymbol{y}}_t^{(c)} = \underbrace{\frac{1}{HW}\sum_{i,j}}_{\text{GAP}} \underbrace{\left[\sum_{k\in K} w_{k,c} T_{\boldsymbol{x},t}(k)\right]}_{\text{CAM at resolution } t} \tag{6}$$

and final class scores are obtained as a sum over scores at different resolutions:

$$\hat{\boldsymbol{y}}^{(c)} = \sum_t \hat{\boldsymbol{y}}_t^{(c)} \tag{7}$$

As indicated by (6), prior to generating a class prediction, a CAM at resolution t is produced. This heatmap contains both positive and negative contributions from the input image towards class c. However, for the sake of heatmap clarity, we consider *only* the positive contributions towards each class when generating our CAMs. That is, we want to see what part of the image *contributes* to normality/abnormality, as opposed to what part of the image *does not* contribute to normality/abnormality. Thus, our CAMs are generated as follows

$$\hat{\boldsymbol{y}}_{\text{CAM}_t}^{(c)} = \left[\sum_{k\in K} w_{k,c} T_{\boldsymbol{x},t}(k)\right]^+ \tag{8}$$

where $z^+ = \max(0, z)$. A loss based just on this final score alone would not force the network to produce meaningful CAMs at every resolution level. Therefore, we also propose to deeply supervise the side predictions in our proposed loss:

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}}, \{\hat{\boldsymbol{y}}_t^{(c)}\}_{c=1,t=1}^{C,T}\right) &= \mathcal{L}_f\left(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}}, \{\hat{\boldsymbol{y}}_t^{(c)}\}_{c=1,t=1}^{C,T}\right) \\ &\quad + \sum_t \mathcal{L}_s^t\left(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}}, \{\hat{\boldsymbol{y}}_t^{(c)}\}_{c=1}^{C}\right)\end{aligned} \tag{9}$$

where $\boldsymbol{x}$ is the input image, y the ground truth class label, $\hat{\boldsymbol{\theta}}$ the network parameters, and $\{\hat{\boldsymbol{y}}_t^{(c)}\}_{c=1,t=1}^{C,T}$ represent the score predictions for each class c at resolution t. Both $\mathcal{L}_f(\cdot)$ and $\mathcal{L}_s^t(\cdot)$ are denoted $\mathcal{L}_f$ and $\mathcal{L}_s^t$ for a simplified notation. $\mathcal{L}_f$ is defined as

$$\begin{aligned}\mathcal{L}_f = &-y \log\left[\sigma\left(\sum_t \hat{\boldsymbol{y}}_t^{(c)}\right)_{c=1}\right] \\ &-(1-y)\log\left[\sigma\left(\sum_t \hat{\boldsymbol{y}}_t^{(c)}\right)_{c=0}\right]\end{aligned} \tag{10}$$

where $\sigma(\cdot)_c$ represents the softmax function for class index c. $\mathcal{L}_s^t$ is the side loss for the prediction at each different resolution t, defined as:

$$\begin{aligned}\mathcal{L}_s^t = &-y \log\left[\sigma\left(\hat{\boldsymbol{y}}^{(c)}\right)_{c=1}\right] \\ &-(1-y)\log\left[\sigma\left(\hat{\boldsymbol{y}}^{(c)}\right)_{c=0}\right]\end{aligned} \tag{11}$$

In addition to the network generating CAMs at every resolution prior to generating the scores as part of the prediction pipeline, the combined loss $\mathcal{L}$ proposed allows for the validation of the accuracy at each resolution depth of the network. We refer to the architecture that implements the model in (5) with embedded CAMs at different resolutions following (6) and loss (9) as ResNet-18-CAM-DS (see Fig. 2).

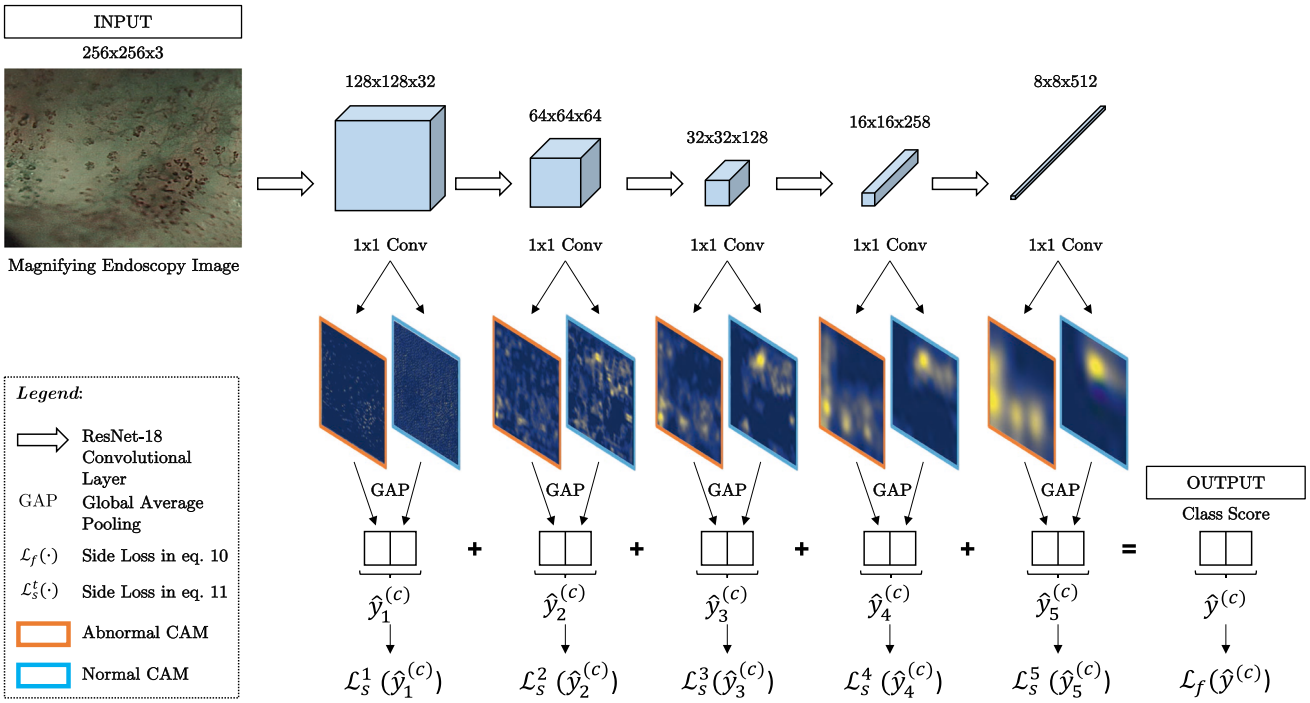

Fig. 2 Proposed model ResNet-18-CAM-DS with embedded positive class activation maps at all resolutions

Results

Our recording protocol (see "Correlating imaged areas with histology" section) enforces that areas recorded in the short patient clips are biopsied. Histopathology labels (normal/abnormal) corresponding to the biopsied specimen are propagated to all the frames of the clip. It is then of interest to evaluate the agreement between the label assigned to each individual frame (based on patient's histopathology) and its correlation to the assessment made by visual inspection of IPCL patterns. A team of 12 senior clinicians with experience in endoscopic imaging of oesophageal cancer labelled 158 images from the dataset (randomly picked across patients and manually filtered so that quasi-identical images are not included). A 25% per IPCL pattern class (normal, B1, B2, B3) is kept across the sample (leading to an imbalance 25% normal, 75% abnormal). The inter-rater agreement was evaluated using the Krippendorff's α coefficient, where values 0% and 100% represent extreme disagreement and perfect agreement, respectively, $\alpha \geq 80\%$ indicates reliable agreement, and $\alpha \geq 66.7\%$ tentative agreement [8]. The Krippendorff's α obtained for the senior clinicians was 76.7%. The labels of each clinician were also compared to the histopathology, obtaining an average sensitivity, specificity, accuracy, and F_1 score (given in %, with a 95% confidence interval) across the 12 clinicians of 97.0 [92.1, 1.0], 88.0 [49.6, 1.0], 94.7 [83.9, 99.7], and 96.5 [89.7, 99.8], respectively.

We report the quantitative classification results for ResNet-18, ResNet-18-CAM, and ResNet-18-CAM-DS in Tables 1, 2, and 3, respectively. ResNet-18-CAM-DS achieved an average sensitivity, specificity, accuracy, and F_1 score of 93.7%, 92.4%, 91.7%, and 94.0%, respectively, all of them better than those achieved by ResNet-18 and ResNet-18-CAM. Accuracy is only three percentage points away from the average of clinical raters. Across all folds, a total of 60 patient clips (12 per fold) are predicted to be normal/abnormal. The binary class estimation for each clip is computed following (1). Each patient in the dataset folder has a unique identification number. We will refer to them in this section to facilitate

Table 1 Results for ResNet-18 (baseline model) on frame classification over the testing set of each fold of the IPCL dataset

Measure (%)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Sensitivity	99.1	96.6	96.7	99.5	64.3	91.2
Specificity	87.9	74.7	62.1	84.9	100.0	81.9
Accuracy	94.8	90.0	77.2	92.8	66.8	84.3
F_1 score	95.8	93.1	79.0	93.7	78.2	88.0

Table 2 Results for ResNet-18-CAM on frame classification over the testing set of each fold of the IPCL dataset

Measure (%)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Sensitivity	98.6	94.6	95.4	97.6	75.9	92.4
Specificity	91.7	89.8	65.9	89.4	100.0	87.4
Accuracy	95.9	93.1	78.8	93.8	77.6	87.8
F_1 score	96.7	95.0	79.8	94.4	86.3	90.4

Table 3 Results for ResNet-18-CAM-DS on frame classification over the testing set of each fold of the IPCL dataset

Measure (%)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Sensitivity	99.6	91.3	98.3	98.9	80.5	93.7
Specificity	81.3	95.1	96.6	89.3	99.8	92.4
Accuracy	92.5	92.4	97.4	94.5	81.9	91.7
F_1 score	94.1	94.4	97.0	95.1	89.2	94.0

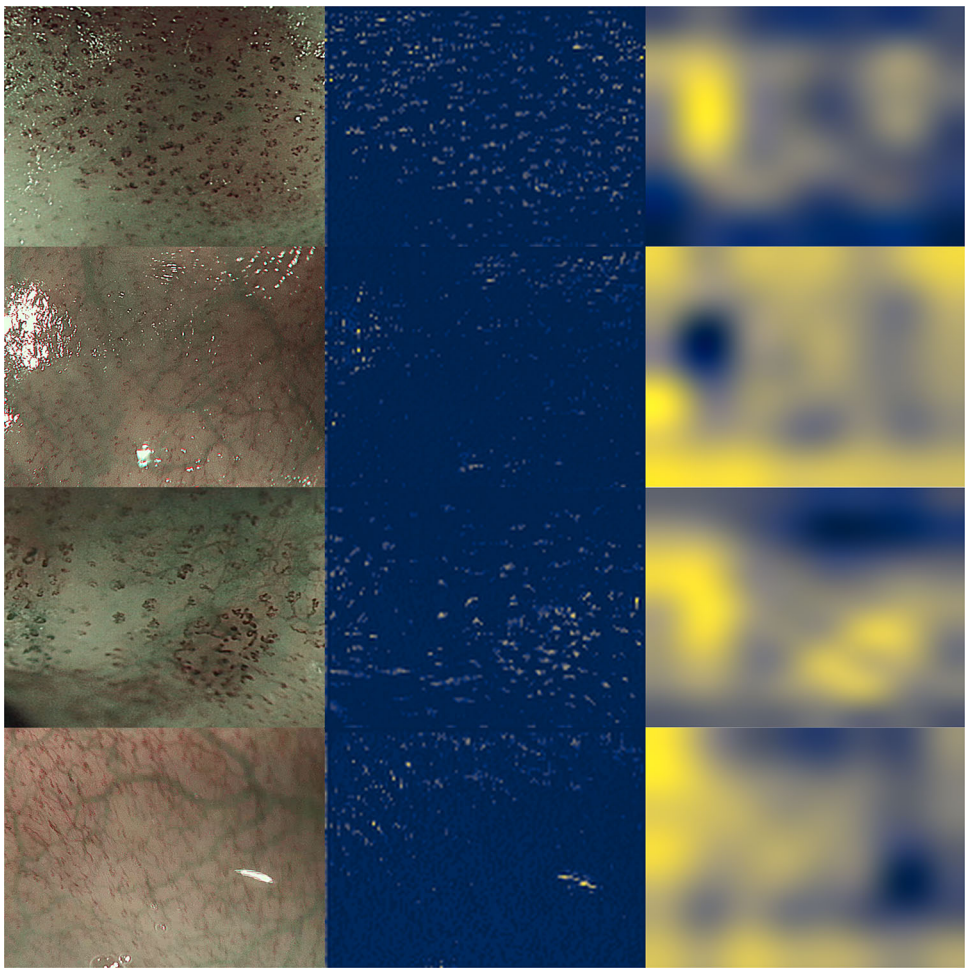

Fig. 3 Representative images from the testing set of fold 1 (left). Highest resolution CAM generated by ResNet-18-CAM-DS for the abnormal class (better viewed in the digital version). That is, $\hat{y}^{(c)}_{\text{CAM}_t} = \hat{y}^{(1)}_{\text{CAM}_1}$ (centre). Class activation maps generated by ResNet-18-CAM [18] (right). In contrast to traditional CAMs generated by ResNet-18-CAM (right), ours (centre) suggest that our network is *looking at* IPCLs to predict abnormality

the search of these patients in the dataset folder. Following (1) to estimate the class of a patient clip, ResNet-18 fails on three patients. Folds 1, 2, and 4 fail on patient 158 (false positive), fold 3 fails on patient 143 (false positive), and fold 5 fails on patient 66 (false negative). ResNet-18-CAM fails on two patients, 143 (false positive) on fold 3, and 66 (false negative) on fold 5. ResNet-18-CAM-DS fails only on folds 1 and 4 in patient 158 (see supplementary material for some frames of these problematic patients). In Fig. 3 a qualitative comparison is shown between the class activation maps produced for the abnormal class by ResNet-18-CAM-DS (at its highest resolution) and the standard class activation maps proposed by Zhou et al. [18]. As our system is designed as a CADe, we have computed the ROC curve (see supplementary material)

to inform the consequences that several choices of sensitivity have on specificity. The AUC of the system is 95.8%.

Discussion and conclusion

Our proposed method ResNet-18-CAM-DS achieves slightly higher average accuracy (91.7%) across folds than our baseline ResNet-18 (84.3%). Although the automated classification accuracy (91.7%) is still below the average achieved by the clinicians (94.7%), it performs better than some of them (their CI low value is 83.9%). It is also encouraging to see that accuracy did not decrease at the expense of an improved explainability. More data and further methodological refinements will most likely lead to improved accuracy. Qualitative results in Fig. 3 seem to indicate that the network is *looking at* IPCL patterns to assess abnormality, which aligns with the clinical practice. However, we have not observed high activations over the large green submucosal vessels in the heatmaps for the normal class. This suggests that they may not be used by the network as an aid to solving the classification problem. Future work could concentrate on adding an attention mechanism to the network in order to consider such vessels as a feature of normal images.

Compliance with ethical standards

Conflict of interest R. J. H. has received research grant support from Pentax Medical, Cook Endoscopy, Fractyl Ltd, Beamline Ltd and Covidien plc to support research infrastructure. T. V. owns shares from Mauna Kea Technologies, Paris, France. The other authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The Institutional Review Board of E-Da Hospital approved this study (IRB number: EMRP-097-022. July 2017).

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Cho JW, Choi SC, Jang JY, Shin SK, Choi KD, Lee JH, Kim SG, Sung JK, Jeon SW, Choi IJ, Kim GH, Jee SR, Lee WS, Jung HY (2014) Lymph node metastases in esophageal carcinoma: an endoscopist's view. Clin Endosc 47(6):523. https://doi.org/10.5946/ce.2014.47.6.523
2. Everson M, Herrera L, Li W, Luengo IM, Ahmad O, Banks M, Magee C, Alzoubaidi D, Hsu H, Graham D, Vercauteren T, Lovat L, Ourselin S, Kashin S, Wang HP, Wang WL, Haidry R (2019) 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 7(2):297–306. https://doi.org/10.1177/2050640618821800
3. Garcia-Peraza-Herrera LC, Everson M, Li W, Luengo I, Berger L, Ahmad O, Lovat L, Wang HP, Wang W.L, Haidry R, Stoyanov D, Vercauteren T, Ourselin S (2018) Interpretable fully convolutional classification of intrapapillary capillary loops for real-time detection of early squamous neoplasia. arXiv:1805.00632
4. Guo L, Xiao X, Wu C, Zeng X, Zhang Y, Du J, Bai S, Xie J, Zhang Z, Li Y, Wang X, Cheung O, Sharma M, Liu J, Hu B (2020) Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest Endosc 91(1):41–51. https://doi.org/10.1016/j.gie.2019.08.018
5. Hassan C, Wallace MB, Sharma P, Maselli R, Craviotto V, Spadaccini M, Repici A (2019) New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut. https://doi.org/10.1136/gutjnl-2019-319914
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
7. Inoue H, Honda T, Yoshida T, Nishikage T, Nagahama T, Yano K, Nagai K, Kawano T, Yoshino K, Tani M, Takeshita K, Endo M (1996) Ultra-high magnification endoscopy of the normal esophageal mucosa. Dig Endosc 8(2):134–138. https://doi.org/10.1111/j.1443-1661.1996.tb00429.x
8. Krippendorff K (2004) Content analysis: an introduction to its methodology. Sage Publications, Thousand Oaks
9. Luo H, Xu G, Li C, He L, Luo L, Wang Z, Jing B, Deng Y, Jin Y, Li Y, Li B, Tan W, He C, Seeruttun SR, Wu Q, Huang J, Huang DW, Chen B, Lin SB, Chen QM, Yuan CM, Chen HX, Pu HY, Zhou F, He Y, Xu RH (2019) Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol 20(12):1645–1654. https://doi.org/10.1016/S1470-2045(19)30637-0
10. Menon S, Trudgill N (2014) How commonly is upper gastrointestinal cancer missed at endoscopy? A meta-analysis. Endosc Int Open 02(02):E46–E50. https://doi.org/10.1055/s-0034-1365524
11. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Proc 73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011
12. Oyama T, Inoue H, Arima M, Momma K, Omori T, Ishihara R, Hirasawa D, Takeuchi M, Tomori A, Goda K (2017) Prediction of the invasion depth of superficial squamous cell carcinoma based on microvessel morphology: magnifying endoscopic classification of the Japan Esophageal Society. Esophagus 14(2):105–112. https://doi.org/10.1007/s10388-016-0527-7
13. Sato H, Inoue H, Ikeda H, Sato C, Onimaru M, Hayee B, Phlanusi C, Santi E, Kobayashi Y, Kudo SE (2015) Utility of intrapapillary capillary loops seen on magnifying narrow-band imaging in estimating invasive depth of esophageal squamous cell carcinoma. Endoscopy 47(02):122–128. https://doi.org/10.1055/s-0034-1390858

14. Su JR, Li Z, Shao XJ, Ji CR, Ji R, Zhou RC, Li GC, Liu GQ, He YS, Zuo XL, Li YQ (2020) Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc 91(2):415–424.e4. https://doi.org/10.1016/j.gie.2019.08.026
15. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, Liu P, Li L, Song Y, Zhang D, Li Y, Xu G, Tu M, Liu X (2019) Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 68(10):1813–1819. https://doi.org/10.1136/gutjnl-2018-317500
16. Zhang Y (2013) Epidemiology of esophageal cancer. World J Gastroenterol 19(34):5598. https://doi.org/10.3748/wjg.v19.i34.5598
17. Zhao YY, Xue DX, Wang YL, Zhang R, Sun B, Cai YP, Feng H, Cai Y, Xu JM (2019) Computer-assisted diagnosis of early esophageal squamous cell carcinoma using narrow-band imaging magnifying endoscopy. Endoscopy 51(04):333–341. https://doi.org/10.1055/a-0756-8754
18. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2921–2929. https://doi.org/10.1109/CVPR.2016.319