TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Predictive features for early cancer detection in Barrett's esophagus using volumetric laser endomicroscopy

Citation for published version (APA):
van der Sommen, F., Klomp, S. R., Swager, A. F., Zinger, S., Curvers, W. L., Bergman, J. J. G. H. M., Schoon, E. J., & de With, P. H. N. (2018). Predictive features for early cancer detection in Barrett's esophagus using volumetric laser endomicroscopy. *Computerized Medical Imaging and Graphics*, *67*, 9-20. https://doi.org/10.1016/j.compmedimag.2018.02.007

DOI:
10.1016/j.compmedimag.2018.02.007

Document status and date:
Published: 01/07/2018

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 13. May. 2025

Predictive features for early cancer detection in Barrett's esophagus using volumetric laser endomicroscopy

Fons van der Sommen[a,b,*], Sander R. Klomp[a], Anne-fré Swager[b], Svitlana Zinger[a], Wouter L. Curvers[b,c], Jacques J.G.H.M. Bergman[b], Erik J. Schoon[c], Peter H.N. de With[a]

[a] *Department of Electrical Engineering, Eindhoven University of Technology, P.O.Box 513, 5600 MB Eindhoven, The Netherlands*
[b] *Department of Gastroenterology, Academic Medical Center, Postbus 22660, 1100 DD Amsterdam, the Netherlands*
[c] *Department of Gastroenterology and Hepathology, Catharina Hospital, P.O.Box 1350, 5602ZA Eindhoven, The Netherlands*

Abstract

The incidence of Barrett cancer is increasing rapidly and current screening protocols often miss the disease at an early, treatable stage. Volumetric Laser Endomicroscopy (VLE) is a promising new tool for finding this type of cancer early, capturing a full circumferential scan of Barrett's Esophagus (BE), up to 3-mm depth. However, the interpretation of these VLE scans can be complicated, due to the large amount of cross-sectional images and the subtle grayscale variations. Therefore, algorithms for automated analysis of VLE data can offer a valuable contribution to its overall interpretation. In this study, we broadly investigate the potential of Computer-Aided Detection (CADe) for the identification of early Barrett's cancer using VLE. We employ a histopathologically validated set of ex-vivo VLE images for evaluating and comparing a considerable set of widely-used image features and machine learning algorithms. In addition,

*Corresponding author
Email addresses: F.v.d.Sommen@tue.nl (Fons van der Sommen), S.R.Klomp@student.tue.nl (Sander R. Klomp), A.Swager@amc.uva.nl (Anne-fré Swager), S.Zinger@tue.nl (Svitlana Zinger), Wouter.Curvers@catharinaziekenhuis.nl (Wouter L. Curvers), support@elsevier.com (Jacques J.G.H.M. Bergman), Erik.Schoon@catharinaziekenhuis.nl (Erik J. Schoon), P.H.N.de.With@tue.nl (Peter H.N. de With)

we show that incorporating clinical knowledge in feature design, leads to a superior classification performance and additional benefits, such as low complexity and fast computation time. Furthermore, we identify an optimal tissue depth for classification of 0.5–1.0 mm, and propose an extension to the evaluated features that exploits this phenomenon, improving their predictive properties for cancer detection in VLE data. Finally, we compare the performance of the CADe methods with the classification accuracy of two VLE experts. With a maximum Area Under the Curve (AUC) in the range of 0.90–0.93 for the evaluated features and machine learning methods versus an AUC of 0.81 for the medical experts, our experiments show that computer-aided methods can achieve a considerably better performance than trained human observers in the analysis of VLE data.
Keywords: Computer-aided detection and diagnosis, Endoscopy, Esophageal adenocarcinoma, Optical Coherence Tomography, Barrett's Esophagus

1. Introduction

Patients suffering from gastric reflux over an extended period of time are prone to developing Barretts Esophagus (BE). This is a condition in which the normal lining of the esophageal wall upwards from the gastroesophageal junction has been replaced by an acid-resistant cell type, which is similar to that of the small intestine (Shaheen & Richter, 2009). It has been estimated that 5.6% of the adult population of the US suffers from a BE (Hayeck et al., 2010) and with obesity and smoking as risk factors for its development (Cook et al., 2012; Lagergren, 2011), a strong increase in its incidence has been observed in recent years (van Soest et al., 2005). Patients with a BE have an over thirty-fold increased chance of developing Esophageal Adenocarcinoma (EAC) (Solaymani-Dodaran et al., 2004). If this type of cancer is detected at an early stage, it can be removed endoscopically, leading to an excellent prognosis (Ell et al., 2007). Typically, patients suffering from BE undergo regular endoscopic surveillance, to examine the BE segment and obtain random biopsies for detecting developing cancer (Reid et al., 2000). However, this surveillance protocol is not optimal,

since early cancer is often missed due to subtle appearance upon visual inspection and the biopsy sampling error (Peters et al., 2008; Corley et al., 2013). Hence, a considerable amount of early cancerous lesions are unnoticed so that the cancer is detected at a later stage, for which the prognosis is substantially worse.

Volumetric Laser Endomicroscopy (VLE) offers a very attractive solution which could efficiently find these early cancerous lesions in BE (Wolfsen et al., 2015). With VLE imaging, a balloon is inflated in the esophagus and a full circumferential scan of the esophageal wall is captured over a segment of 6 cm, up to a depth of 3 mm, using second generation Optical Coherence Tomography (OCT) (Gonzalo et al., 2010). This unique capability allows the physician to analyze the underlying tissue layers, theoretically enabling better detection of early cancer. However, due to the large volume and subtle nature of the greyscale cross-sectional data, the interpretation of the VLE images remains a challenging task for gastroenterologists. Although a recent clinical prediction model has shown reasonable detection accuracy (Swager et al., 2016d), the identification of early cancer on VLE images using current criteria remains very complex (Swager et al., 2016b,c). Hence, an automated system for the analysis of VLE scans would be highly desirable for supporting the physician. However, the applicability of computer-aided methods for the interpretation of VLE data remains to be discovered. Therefore, in this study, we investigate the basic feasibility and the potential of computer-aided methods for the analysis of VLE imagery. We establish a benchmark employing a considerable set of widely-used image features and classification methods on a histopathologically-validated dataset of ex-vivo VLE images. In addition, we propose three clinically-inspired features based on a recent clinical prediction model: two completely novel features and an additional one, derived from a widely-applied texture feature. To validate and compare the performance of the evaluated methods, we use a thorough validation procedure and compare the results to the classification performance of two VLE experts on the same set of VLE images.

The remainder of this paper is organized as follows. We first provide an

overview of related work in Section 2 and continue with a comprehensive description of the employed methods in Section 3, in which we elaborate on the data acquisition (Sections 3.1 and 3.2), the evaluated features and classification methods (Sec. 3.3 and 3.4) and a detailed description of our validation procedure (Sections 3.5 to 3.8). The results of this study are presented in Section 4 leading to our conclusions as outlined in Section 5.

2. Related work

Earlier work of Qi et al. (2010) has shown promising results on the detection of dysplasia using Endoscopic OCT (EOCT), achieving a detection accuracy of 84% and maximal Area Under the Curve (AUC) of 0.84. In contrast to VLE, EOCT is a probe-based system with a relatively small scanning surface (Rollins et al., 1998, 1999). Qi et al. used the working channel of the endoscope for the EOCT probe and a suction cap for taking a biopsy, ensuring a histopathology correlation between the scan and the tissue. This means only a small portion of the BE can be imaged at a time, whereas with VLE the complete circumferential and longitudinal Barrett segment is captured in a single scan. Hence, with EOCT it is infeasible to scan the complete BE and it is very hard to ensure no malignant lesions go unnoticed. Furthermore, the EOCT system in the work of Qi et al. employs a first-generation OCT imaging device, resulting in a considerably reduced image quality compared to the VLE system used in our study. However, the 18 features that have been developed by Qi et al. to distinguish between non-dysplastic tissue low-grade dysplasia and high-grade dysplasia might be applicable to the VLE images employed in this study. Hence, we have implemented these features and included them in our experiments.

More recently, Rodriguez-Diaz & Singh (2015) have presented an algorithm for computer-assisted image interpretation of VLE images that employs statistics on the Gray-Level Co-occurrence Matrices (GLCM) of the first wavelet components of the image, followed by a naive Bayes classifier. In this abstract, four tissue classes were separated and a set of 60 VLE images was employed for

validation. A sensitivity and specificity of 0.86 and 0.93, respectively, were reported for computer-aided classification between dysplastic and non-dysplastic Barretts tissue. Although the results of this study are promising, it accommodates a major drawback: the VLE images were not correlated one-to-one with histology, which is the gold standard for diagnosis. Therefore, the reliability of the ground truth used in this study is limited.

In an attempt to autonomously segment and characterize the esophageal wall, Ughi et al. (2016) have proposed an algorithm for the analysis of Tethered Capsule OCT Endomicroscopy (TCE). TCE provides real-time three-dimensional imaging of the esophageal wall, after a capsule is swallowed by the patient and is slowly pulled back (Gora et al., 2013). In the study of Ughi et al., two tissue classes are characterized, namely Barretts and normal squamous tissue. The algorithm first creates an en face map for finding the contact between the surface tissue and the capsule. Next, the presence of a clearly layered structure in the signal is used as an indicator for normal squamous tissue, whereas for Barretts tissue, this clear layering is typically lacking. On 50 manually-annotated OCT images, the system demonstrated a sensitivity and specificity of 94% and 93%, respectively. In our study, we aim to distinguish between dysplastic and non-dysplastic Barretts tissue, both lacking the clear layering as it is present for normal squamous tissue. Although the proposed tissue classification algorithm of Ughi et al. is not suitable for early cancer detection, it could be applied as a pre-processing step for the analysis of complete VLE scans in order to separate the Barretts from the normal squamous tissue of the esophageal wall.

3. Materials and methods

3.1. Patients and Image Acquisition

The images used in this study are derived from a previously established database, consisting of correlated ex-vivo VLE images and histology slides (Swager et al., 2016a). In this subsection, we provide a short overview of the construction of this VLE-histology database (figures derived from (Swager et al.,

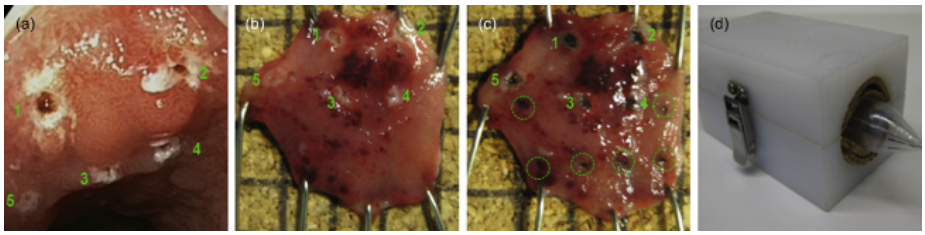

Figure 1: (a) In-vivo placement of reference Electrocoagulation Markers (ECMs). (b) Ex-vivo specimen pinned in in-vivo orientation on gridded cork with squares of 5 by 5 mm. (c) In-vivo ECMs (1-5) accentuated with ink, dashed circles indicate ink markers placed with 25 gauge needle. (d) Custom-designed tubularshaped fixture for VLE scanning of endoscopic resection specimen. *Reproduced with permission of John Wiley and Sons.*

2016a)). Two groups of patients were eligible for this study: patients with Non-Dysplastic Barretts Esophagus (NDBE) undergoing surveillance endoscopy, and patients referred for work-up and treatment of early neoplasia (High-Grade Dysplasia and/or early Esophageal Adenocarcinoma HGD/EAC).

Histology-VLE correlation. First, during high-definition endoscopy, the esophagus is examined according to the standard guidelines with white light and narrow-band imaging. Next, standard measurements are recorded for describing the Barretts segment and the neoplastic lesion (if present). In case of a lesion, it is delineated using electrocoagulation marks and additional electrocoagulation marks are placed within the delineated area. Subsequently, endoscopic resection is performed and the endoscopic resection specimens are pinned on cork with a 5 mm-squared grid (see Fig. 1 (b,c)). Both in-vivo placed markers (electrocoagulation) and ex-vivo placed marks (pins and ink by needle) are used as objective markers. Finally, the specimens are scanned with the VLE balloon using a custom-designed fixture (see Fig. 1 (d)).

Endoscopic procedure and resection specimens. To match the histopathology slides with VLE scans, the electrocoagulation, ink and pin marks are identified on both modalities. The markers are used to obtain one-to-one correlation between the VLE images of the ex-vivo endoscopic resection specimens and the corresponding digitized histopathology slides. During this process, the location of the histological transection plane was known, since the tissue sectioning was

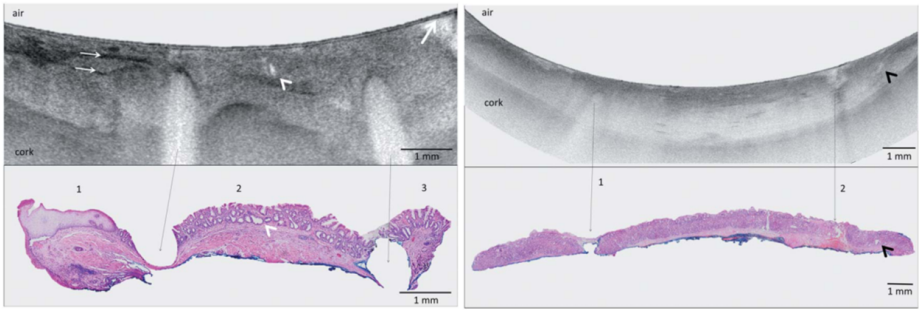

Figure 2: Left: histology-VLE match containing 2 pin markers. These are visible on histology (bottom panel) as pin holes (black arrows) and create low-backscattering, well demarcable structures on VLE (top panel). Area 1: squamous epithelium, characterized on VLE by layering (upper white arrow indicates transition of epithelium to lamina propria and lower white arrow to muscularis mucosa). Area 2 and 3 are non-dysplastic Barretts mucosa, which is visible by loss of layering and dilated Barretts glands (arrowheads). Bold white arrow indicates VLE balloon. Right: histology-VLE match with 1 pin marker (arrow 1) and 1 electrocoagulation marker (arrow 2), containing gastric mucosa. Arrowheads indicate dilated gastric gland. *Reproduced with permission of John Wiley and Sons.*

performed alongside the marks. The corresponding VLE plane is determined in the VLE scan based on the distance according to the gridlines (Fig. 1 (b,c)) and the marks. If at least two markers are visible on both modalities, it is considered a VLE-histology match (see Fig. 2).

VLE image data set. A total of 29 patients have been included for the construction of the data set resulting in 52 tissue specimens. From these specimens, 86 histology matches have been identified, resulting in a total of 200 matched VLE frames. Next, 125 frames have been excluded due a histopathogical diagnosis other than NDBE or HGD/EAC or due to insufficient image quality. From the remaining set, 10 frames (5 NDBE; 5 HGD/EAC) have been used for a clinical orientation phase and 60 frames (30 NDBE; 30 HGD/EAC) are included in the VLE image dataset.

3.2. Data normalization

For each matched VLE frame, a region of interest has been manually extracted, that is defined horizontally by the marker positions and vertically by the balloon edge (top) and the cork (bottom). To standardize the images, the number of horizontal lines is restricted to 400 pixels, corresponding to a depth

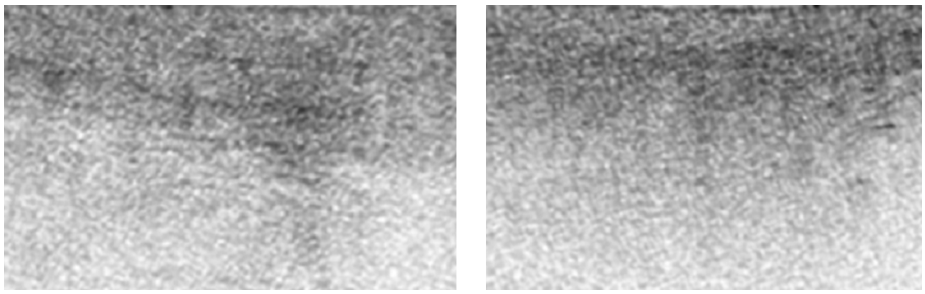

Figure 3: Normalized VLE images of non-dysplastic (left) and neoplastic tissue (right), which are cropped versions of the originals, showing only the region of interest. The horizontal and vertical regions of interest are defined by the markers and the distance between the balloon and the cork, respectively.

of approximately 2 mm. The amount of vertical scanlines is restricted by the markers that indicate the histology correlation. Hence, the resolution of the regions of interest is W 400 pixels, where W defines the width of the region of interest. Fig. 3 shows an example of two normalized VLE images.

3.3. Features for cancer detection

Based on a recently published clinical prediction model for the interpretation of VLE image data (Swager et al., 2016d), we have derived three image features for quantification of discriminative information, which have shown promising results in a preliminary study (Klomp et al., 2017). The clinical prediction model identifies three key aspects for scoring a VLE image: (1) lack of layering, (2) surface signal and (3) irregular glands. The first aspect captures the abnormal growth of early lesions, which disturbs the somewhat layered structure of Barretts epithelium. The second aspect describes the lack of surface maturation due to the presence of dysplasia (Odze, 2006), which results in a higher VLE surface signal, relative to the subsurface signal. The last aspect regards the presence of irregular shaped dysplastic glands, but as this aspect was only sparsely present in the data, we have focused on the first two clinical aspects for the derived image features.

To investigate if incorporating this clinical knowledge in the feature design leads to an improved performance, we have derived three features specifically

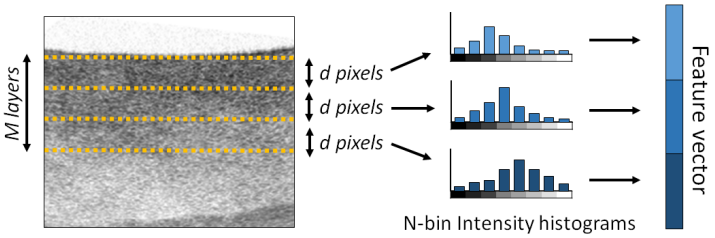

Figure 4: Computation of the Layer Histogram (LH) feature. First, the M top layers of d pixels are extracted from the VLE image. Next, for each layer an N-bin intensity histogram is computed. Finally, the resulting histograms are concatenated to obtain a feature vector.

based on this clinical prediction model: (1) Layer histogram, (2) Large-scale gray-level co-occurrence matrix features and (3) Bin median of pixel averages. For the remainder of this paper, these features are referred to as clinally-inspired features.

Layer histogram. In order to capture the (lack of) layering present in the VLE image, we propose a simple image feature that computes the N-bin histograms of the first M layers of d pixels, starting from the top of the image. This results in a feature vector of dimensionality NM for each image. The motivation for this method is threefold: (1) the lack of layering is an indicator for dysplasia, (2) the top layers are described as the most informative for tissue classification (Swager et al., 2016d), and (3) the signal-to-noise ratio decreases for an increasing scanning depth. Figure 4 shows a schematic depiction of the computation of this feature.

Large-scale GLCM. As the clinical prediction model has a large emphasis on vertical structure, we have adapted the default Gray-Level Co-occurrence Matrix (GLCM) texture features (Haralick & Shanmugam, 1973), such that it captures the large-scale vertical tissue structures, like layering and surface maturation. In contrast to using a series of offsets capturing the close neighborhood of a pixel, we only use a single, relatively large offset in the vertical direction. Next, we compute the properties contrast, correlation, energy and homogeneity based on the obtained co-occurrence matrix. To distinguish this feature from

the traditional GLCM, in the remainder of the paper it will be referred to as Large-Scale GLCM (LS-GLCM).

Bin-median of pixel averages. As a computationally less demanding alternative, we have developed a very simple feature that also incorporates the key properties of the clinical prediction model. In order to capture the vertical signal gradient, we first compute the average signal intensity for each horizontal line of pixels and split them into N equally-sized bins. Next, to capture the tissue layering, we compute the medians of these bins. This results in an N-dimensional feature vector for each image. This feature is referred to as the Bin-Median of Pixel Averages (BMPA) in the remainder of the paper.

Features used for benchmarking. As VLE is a relatively new technology, no extensive evaluation of image features for tissue classification is readily available. In order to obtain a context for our results and to provide a benchmark for future studies on VLE tissue classification, we evaluate a broad set of commonly-used features for Computer-Aided Diagnosis / Detection (CAD / CADe). This set encompasses: Local Binary Patterns (LBP) (Ojala et al., 1996), statistics on the GLCM (Haralick & Shanmugam, 1973), Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005) and Gabor features (Fogel & Sagi, 1989). In addition, we have included the features of Rodriguez-Diaz & Singh (2015) and Qi et al. (2010). Finally, we have included image features extracted from deep Convolutional Neural Networks (CNNs) that have been pre-trained on large data sets. Typically, the network output of one of the intermediate network layers is extracted and the resulting features are referred to as CNN-codes (Orlando et al., 2017). This form of transfer learning has become quite popular in the field of medical image analysis and has been applied successfully to several CAD and CADe problems (Lu et al., 2016). In our experiments, we employ the output of the 6^{th} and 7^{th} neural network layers of the widely-used AlexNet, known as FC6 and FC7, respectively, which was pre-trained on the ImageNet data set (Deng et al., 2009). Prior to feeding the images to the CNN, a normalization step enforced the AlexNet input size of $227 \times 227 \times 3$ pixels, by cropping a square

in the horizontal middle and the vertical top of the image. The motivation for this choice is twofold: (1) we want to ensure a fixed aspect ratio of anatomical structures over the varying image sizes, which would be violated by applying a re-scaling operation, and (2) the observation of Swager et al. (2016d), that the upper layers reveal the most discriminative information for a clinical prediction model.

3.4. Image classification

For classification of the images, we have evaluated the following methods, using the features described in the previous subsection as input: Support Vector Machine (SVM) (Cortes & Vapnik, 1995), Random Forests (RF) (Breiman, 2001), Adaptive Boosting (AdaBoost) (Freund & Schapire, 1995), Neural Networks (NN), k-Nearest-Neighbors (kNN), Discriminant Analysis (DA) and Logistic Regression (LogReg). In addition, we have evaluated Convolutional Neural Networks (CNN) by retraining only the last couple of layers of a pre-trained network (Krizhevsky et al., 2012), which is a form of Transfer Learning (TL) (Lu et al., 2016), (Oquab et al., 2014). Obviously, for the latter experiment we are restricted to the features that are learned from the data that the original network was trained on and we cannot use this form of classification for the features presented in Section 3.3.

3.5. Hyperparameter optimization

The majority of the evaluated features and the classification methods include hyperparameters that affect the performance. In order to enable a fair comparison of the presented methods, prior to training the algorithm, the optimal values for these hyperparameters are estimated on the training set. For this, we employ 100 trials using randomly sampled values from a pre-defined search range. In each trial, the estimated performance for those parameter values is computed using a five-fold cross-validation. We prefer random search over the commonly-employed grid search, as recent work has shown that randomly chosen trials are more efficient for hyperparameter optimization (Bergstra &

Yoshua, 2012). Additionally, considering the amount of involved parameters in our evaluation, a grid-search of sufficient density would dramatically limit the feasibility of the experiments due to time- and computational constraints. Tables 1 and 2 provide an overview of the evaluated features and classification methods, respectively, including the involved hyperparameters for each of the methods and the employed search range for the random trials. These ranges have been chosen, based on commonly-encountered settings in literature and the type and dimensions of the input data. This typically results in a range around the default settings, of which values are uniformly sampled for each trial. In Tables 1 and 2, parameter ranges are indicated with $[\cdot]$ for real values and with $\{\cdot\}$ for integer values and non-numeric settings.

3.6. Performance Analysis

To analyze the classification performance of the different features and machine learning methods, we use Leave-One-Out Cross-Validation (LOOCV), where a score is generated for each image independently. Next, performance metrics are computed based on the predictions for all images. It should be noted that in this fashion, each image is classified by a slightly different model, as the training data is different for each test image. This yields the best available proxy for the generalization power of a certain algorithm, since any random choice of training and test data may lead to either too optimistic or too pessimistic results. From the individual prediction scores we compute the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) and we use the latter as our main figure of merit. In addition, we compute the Sensitivity and the Specificity which reflect the performance in the default point of operation.

3.7. Framework for comparative validation

In order to enable reproducibility and comparability of results presented in this study, we use the validation model of Jannin et al. (2006). We have slightly modified that framework, such that it applies to the methods used in this study

Feature	**Hyperparameters**	**Parameter selection**
Graylevel Co-occurrence Matrix (GLCM)	# levels, distance D, offsets	$\{4, 5, \ldots, 12\}$ $D \in \{1, 2, \ldots, 10\}$ $\begin{bmatrix} D & 0 \\ D & -D \\ 0 & -D \\ -D & -D \end{bmatrix}$
Local Binary Patterns (LBP)	radius, # neighbors, rotational invariance	$\{1, 2, \ldots, 8\}$ $\{4, 5, \ldots, 12\}$ $\{true, false\}$
Histogram of Oriented Gradients (HOG)	# horizontal levels, # vertical cells D, block size	$\{8, 9, \ldots, 20\}$ $\{8, 9, \ldots, 20\}$ $[b_h, b_v] \quad b_h, b_v \in \{1, 2, 3\}$
Gabor-based features	lower wavelength, high wavelength D, # scales P # orientations	$\lambda_{low} \in \{8, 9, \ldots, 30\}$ $\lambda_{high} = \lambda_{low} \cdot 2^{1-P}$ $P \in \{1, 2, 3, 4\}$ $\{2, 3, 4, 5, 6\}$
CNN codes of the FC6-layer of AlexNet (FC6)	-	-
CNN codes of the FC7-layer of AlexNet (FC7)	-	-
Qi et al. (2010) (Qi-PC5, Qi-F18)	IM: Filt. size, win. size, SD: block size, struct. element, threshold, CSAC: # bins TFCN: delta, offsets GLCM: # levels, offsets	from original publication
Rodriguez-Diaz & Singh (2015) (RD)	# levels, offsets	from original publication
Layer Histogram (LH)	# layers, layer size, offsets	$\{2, 3, \ldots, 8\}$ $\{30, 31, \ldots, 50\}$ $\{4, 5, \ldots, 12\}$
Large Scale GLCM (LS-GLCM)	# levels, distance D, offsets	$\{4, 5, \ldots, 12\}$ $D \in \{50, 51, \ldots, 70\}$ $[D, 0]$
Bin Median of Pixel Averages (BMPA)	# bins, bin size	$\{8, 9, \ldots, 16\}$ $\{15, 16, \ldots, 25\}$

Table 1: Employed features, hyperparameters and search range for optimization.

Classifier	Hyperparameters	Parameter selection
Support Vector Machine (SVM)	regularization constant	$2^p,\ p \in [-5, 5]$
	kernel	$\{linear, RBF\}$
	kernel scale	$2^p,\ p \in [-5, 5]$
Random Forest (RF)	forest size	$\{50, 51, \ldots, 150\}$
	max. splits per tree	$\sqrt{\#dimensions}$
	randomness control	$r \cdot \#samples$
		$r \in [0.2, 0.3]$
AdaBoost	# learning cycles	$\{2, 3, \ldots, 8\}$
	weak learner type	Binary Split
k Nearest Neighbors (kNN)	# neighbors	$\{1, 2, \ldots, 15\}$
	distance metric	$\left\{ \begin{smallmatrix} Chebychev, \\ Euclidean, \\ Hamming \end{smallmatrix} \right\}$
Neural Network (NN)	# hidden layers	$\{5, 3, \ldots, 15\}$
Discriminant Analysis (DA)	discriminant type	$\left\{ \begin{smallmatrix} diagLinear, \\ diagQuadratic, \\ pseudoLinear, \\ pseudoQuadratic \end{smallmatrix} \right\}$
Naive Bayes (NB)	-	-
Logistic Regression (LogReg)	-	-
Transfer Learning using a Convolutional Neural Network (TL-CNN)	-	-

Table 2: Employed classification methods, hyperparameters and search range for optimization.

and it presents a clear overview of the validation procedure (as presented in Sections 3.5 and 3.6). Fig. 5 shows the modified framework, in which D_{TR} represents the training data after leaving sample $d^{(n)}$ out of the full data set D, G_M is a function that generates parameter values p for method M and $R_M^{(n)}$ is the resulting prediction score for image $d^{(n)}$ using method M. Reference method F_{ref} represents the acquisition protocol as described in Section 3.1 and R_{ref} is the resulting ground truth. Finally, F_C denotes a comparison function that compares the predictions scores $\hat{R}_M$ of method M with ground truth R_{ref} resulting in quality index Q_M, for which we use the AUC.

In this study, we use $J = 100$ random trials and five-fold cross-validation ($k = 5$) for hyperparameter optimization. As the training set is relatively small, the loss function can only be estimated coarsely in the parameter space. Therefore, using a very high number of trials will not lead to a better estimation of the optimal hyperparameter values, as this amounts to oversampling this coarse approximation of the loss function.

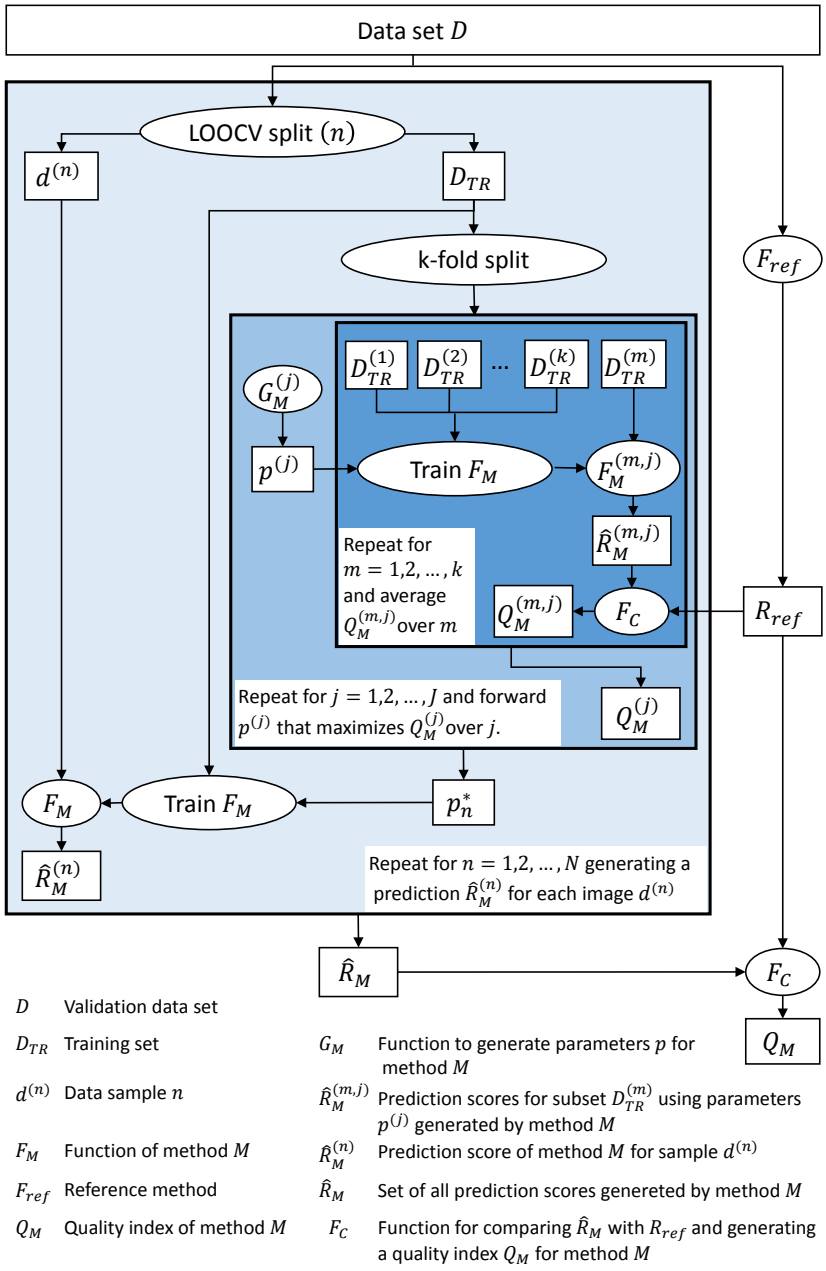

Figure 5: Framework for comparative validation, based on a validation model proposed by Jannin et al. (2006) and the procedures described in Sections 3.5 and 3.6. A prediction score is obtained for each image independently using Leave-One-Out Cross-Validation (LOOCV), where all hyperparameters are optimized over the training set.

3.8. Clinical validation

For clinical validation of the algorithm, we compare the results of the automated tissue classification to the performance of two VLE experts, obtained in an earlier study on the same set of images (Swager et al., 2016d). The experts involved in this study were blinded for the corresponding histopathology and scored 20 images in a learning phase and 40 images in a validation phase, while they were allowed to assess the 20 adjacent VLE frames (10 proximal/10 distal) of the matched frame. During the learning phase, the reviewers classified the images, using VLE features potentially predictive for Barretts neoplasia. After the learning phase, the histopathology of the 20 images was revealed during a consensus meeting. With the results of the learning phase a VLE prediction model was developed. Using this prediction model, the two experts scored the remaining 40 images. The combined results of both experts on this evaluation are used for comparison with the performance of the CAD system.

4. Results

4.1. Image features

Table 3 presents the classification performance for all combinations of features and classification methods, where results with an AUC of 0.80 or higher are printed in boldface. From this table, we observe that the proposed clinically inspired features show a superior performance to state-of-the-art alternatives over all classification methods, except for Neural Networks. The maximum AUC for the clinically inspired features is 0.90, which is achieved by LH features in combination with a linear SVM, while the maximum performance for the state-of-the-art features is 0.82 for FC6 and for FC7 features in combination with a neural network and a linear SVM classifier, respectively.

Considering the features presented in related work, a relatively poor performance is observed. The features proposed by Rodriguez-Diaz & Singh (2015) for VLE data show a maximum AUC of 0.73. While the features proposed by Qi et al. (2010) show a slightly better performance overall, they also indicate a

	GLCM	LBP	HOG	Gabor	FC6	FC7	RD	Qi-F18	Qi-PC5	LH	BMPA	LS-GLCM	**Average**
SVM *linear*	0.70	0.65	0.66	0.71	**0.80**	**0.82**	0.68	0.63	0.53	**0.90***	0.68	0.75	*0.71*
SVM *RBF*	0.72	0.56	0.76	0.65	0.78	0.78	0.61	0.61	0.45	**0.87***	**0.86**	0.79	*0.69*
DA	0.61	0.63	0.49	0.57	0.68	0.77	0.73	0.69	0.48	**0.84**	**0.86***	**0.85**	*0.70*
AdaBoost	0.70	0.64	0.51	0.60	0.70	0.72	0.59	0.66	0.36	**0.88***	**0.81**	**0.86**	*0.68*
RF	0.73	0.59	0.64	0.66	0.77	0.76	0.62	0.68	0.54	**0.86***	**0.81**	**0.84**	*0.71*
kNN	0.69	0.56	0.55	0.59	**0.82***	0.78	0.59	0.61	0.22	0.77	0.76	0.73	*0.64*
NN	0.73	0.58	0.58	0.65	0.73	0.72	0.58	0.72	0.50	0.76	**0.84***	**0.83**	*0.68*
NB	0.62	0.60	0.59	0.65	0.74	0.76	0.65	0.73	0.48	**0.83**	**0.84***	0.77	*0.69*
LogReg	0.66	0.62	0.48	0.65	0.54	0.64	0.67	0.67	0.61	**0.83**	0.67	**0.84***	*0.66*
Average	0.68	0.60	0.58	0.63	0.73	0.75	0.62	0.66	0.47	**0.84***	0.79	**0.81**	
Convolutional Neural Network (CNN) Transfer Learning													
TL-Retrain FC7							**0.84**						
TL-Retrain FC6+7							**0.81**						

Table 3: Features and machine learning methods for Barrett's cancer detection (see Tables 1 and 2 for acronym definitions). *Highest result for each machine learning method indicated with an asterisk.

	GLCM★	LBP★	HOG★	Gabor★	FC6★	FC7★	RD★	Qi-F18★	Qi-PC5★	LH†	BMPA†	LS-GLCM★	**Avg.**
SVM *linear*	**0.87**	**0.80**	0.68	0.75	**0.80**	0.76	0.72	0.79	0.79	**0.90***	0.68	**0.86**	*0.78*
SVM *RBF*	0.71	**0.89***	0.75	0.74	0.74	0.77	0.59	0.70	0.70	**0.87**	**0.86**	**0.83**	*0.76*
DA	0.76	0.71	0.49	**0.83**	**0.82**	0.73	0.64	0.69	0.77	**0.84**	**0.86***	**0.83**	*0.74*
AdaBoost	0.60	**0.85**	0.56	**0.81**	0.74	0.69	0.63	0.53	0.65	**0.88***	**0.81**	0.78	*0.70*
RF	0.64	**0.83**	0.67	**0.82**	0.74	0.73	0.57	0.68	0.68	**0.86**	**0.81**	**0.88***	*0.75*
kNN	0.78	**0.83***	0.53	0.64	**0.83***	0.72	0.53	0.78	0.77	0.77	0.76	0.67	*0.73*
NN	0.75	0.75	0.68	**0.84***	0.74	0.67	0.58	0.68	0.68	0.76	**0.84***	**0.81**	*0.72*
NB	0.65	**0.80**	0.69	0.77	0.74	0.76	0.61	0.75	0.75	**0.83**	**0.84***	0.75	*0.74*
LogReg	0.64	0.76	0.69	0.76	0.54	0.51	0.51	0.75	0.75	**0.83**	0.67	**0.90***	*0.68*
Avg.	0.71	**0.80**	0.64	0.77	0.74	0.71	0.60	0.72	0.73	**0.84***	0.79	**0.81**	

Table 4: Detection results after applying automated rows-of-interest selection (modified features indicated with a ★ symbol). *Highest result for each machine learning method indicated with an asterisk. †Results copied from Table 3 for comparison.

maximal AUC of only 0.73. This observation can be explained by the fact that these features have been developed for EOCT, which has a far smaller scanning surface and a lower signal-to-noise ratio. For our experiments, the full features showed a slightly better performance than the first five principal components, which was proposed in the original publication.

Traditional texture features, such as HOG, GLCM, LBP and Gabor-based features show a relatively poor performance in Table 3. As VLE images exhibit no consistent edges, the poor classification performance when using HOG is expected. With an average AUC of 0.58 over all evaluated classification methods, HOG achieved the worst performance. When using LBP, only a slightly better average classification of 0.60 is observed with a maximal AUC of 0.65 when using a Linear SVM classifier. From the set of traditional texture features, GLCM showed the best performance with an average AUC of 0.68 and a maximal AUC of 0.73 for classification methods kNN and Random Forests. Overall, the performance of the traditional texture features is slightly disappointing, for which a possible explanation will be discussed in Section 4.3.

In contrast, the results obtained using transfer learning are relatively good. The lower two rows of Table 3 present the AUC that is achieved when the last layer(s) of AlexNet are retrained using the VLE images. When retraining only the last or the last two layers, an impressive AUC of 0.81 and 0.84, respectively, can be observed. Using the AlexNet neuron responses of the mid-level image representations as features, i.e. the output of Fully Connected (FC) layers, a maximal AUC of 0.82 is observed for both the FC6 and FC7 features. Interestingly, both forms of transfer learning achieve relatively good results even though the used CNN was trained on ImageNet, which contains images of a completely different nature. This observation is in line with earlier findings on transfer learning applied to medical data (Lu et al., 2016).

The results obtained using the proposed clinically-inspired features showed the highest classification performance, with a maximal AUC of 0.90, 0.86 and 0.86, for LH, BMPA and LS-GLCM, respectively. For all classification methods except neural networks, the highest AUC (indicated with an asterisk in Table 3)

was achieved using one of the clinically inspired features. With an average AUC of 0.81, 0.79 and 0.81, these features also show a robust detection performance for different classification methods.

4.2. Scanlines of interest

Earlier clinical studies hypothesized that the top layers of the tissue contain the most discriminative information regarding the presence of dysplasia (Swager et al., 2016d). This hypothesis is further strengthened by the results presented in Table 3, which clearly indicate that the features focusing on the upper part of the image, i.e. LH and BMPA, generally achieve a superior performance. These two features both include parameters directly affecting the number of scanlines that are used for analysis (from the top down). Using the distribution over these parameters, resulting from hyperparameter optimization, the number of effective scanlines can be computed. Fig. 6 shows the distribution over the number of scanlines used for computing the LH and BMPA features for the four respective best performing machine learning methods. Clearly, the selected values for the optimal hyperparameters result in an effective number of scanlines that is considerably lower than the full 400-pixel image height. More specifically, the optimal number of scanlines ranges from 50 to 250 pixels in 84% and 90% of the experiments for LH and BMPA, respectively. With a lateral resolution of 4.7 m, this translates to approximately the top 0.2–1.2 mm of the tissue.

The observation that the two best performing features focus on the upper tissue layers, fuels the expectation that other feature methods can also achieve a better performance when only the upper part of the VLE image is used for analysis. Therefore, in an additional experiment, we have included the number of scanlines as a free parameter for all other evaluated features. This means that during hyperparameter optimization, each feature has the option to use a limited number of effective scanlines. For each trial, the number of scanlines is randomly sampled, using a uniform distribution ranging from 50 to 250 pixels. Table 4 shows the result of this experiment, including the results for LH and BMPA from Table 3, which have been added for ease of comparison. This table

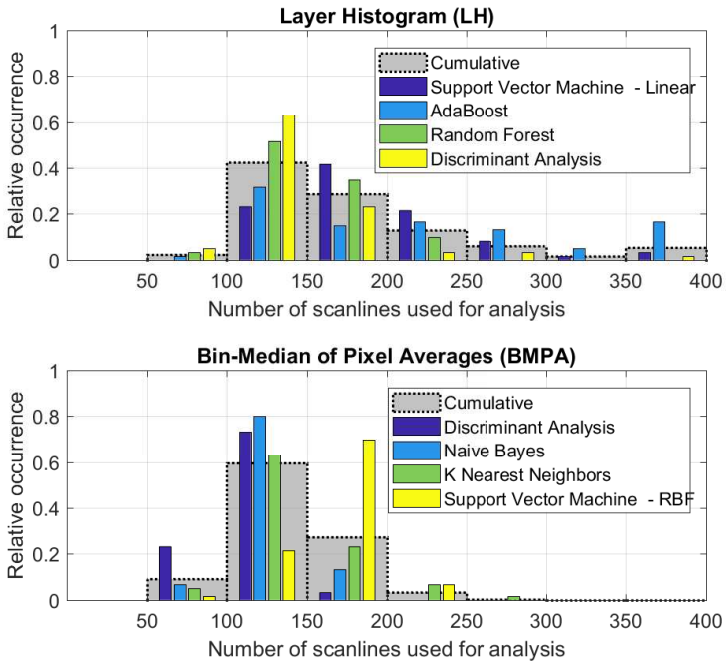

Figure 6: Histograms of the effective number of scanlines (from the top down) used by the LH feature (top) and the BMPA feature (bottom) for the four best performing classification methods. During cross-validation, the optimal values are determined for the number of layers and the number of scanlines per layer. The histograms are computed using the product of these two parameters for each of the 60 validation iterations.

clearly demonstrates that the performance of certain features can be considerably improved by adding the effective number of scanlines for analysis as a free parameter. More specifically, allowing the methods to limit the analysis to the upper 0.2–1.2 mm of the tissue boosts the average performance for all evaluated traditional texture features and the ones proposed by Qi et al. (2010).

Considering the features proposed for similar applications, the features of Rodriguez-Diaz et al. show a comparable performance. However, the features of Qi et al. exhibit a relatively strong increase. This performance boost can be explained by the fact that these features have been originally developed for EOCT, a modality that yields a considerably worse signal-to-noise ratio for lower tissue layers than VLE. Hence, these features have been specifically designed for a range of only 1–1.5 mm below the surface tissue and it is not surprising that they achieve a higher performance when the analysis is limited to the top part of the VLE images.

To measure the performance of the CNN codes in this experiment, the images are rescaled after cropping to the randomly sampled number of scanlines. This choice was motivated by the varying width of the matched VLE images, limited by the markers for histopathology correlation, which inhibited the use of a square crop as was applied in the previous experiment (see Section 4.1). Furthermore, this allowed the network to use all the available vertical scanlines of each image. Interestingly, the CNN codes from the FC6 and FC7 layers of AlexNet show a slight decrease in performance in Table 4. Firstly, in the latter case, the morphological tissue structures are deformed by rescaling the images only in the vertical direction and secondly, it is likely that the original AlexNet input size of 227 scanlines is already in the optimal range of effective scanlines. Hence, the FC6 and FC7 features already enjoyed the advantage of limiting the analysis to the top layers in the results presented in Table 3.

Overall, the clinically inspired features still show a superior performance to all the state-of-the-art alternatives, except for LBP. However, the performance gap has been narrowed considerably. Hence, this experiment further reinforces the hypothesis that for the classification of early Barretts cancer in VLE imagery,

the top tissue layers contain the most discriminative information.

4.3. Optimal number of scanlines

To further investigate the optimal range of tissue depth, we have computed the classification performance of the most promising feature methods (AUC > 0.8) from Table 4 for an increasing Depth Of Interest (DOI). As it is not feasible to carry out this experiment while using hyperparameter optimization, we have fixed the involved hyperparameters to the optimal values as obtained in the previous experiment (Sec. 4.2). Fig. 7 shows the AUC for an increasing DOI for various features and classification methods, where classification methods that did not surpass an AUC of 0.8 have been excluded. From these plots, it is clear that each feature generally has a unique optimal DOI, where most classification methods achieve the highest classification performance. Typically, this optimal DOI is well within the earlier determined range of 0.2–1.2 mm, as observed in Section 4.1, except for the FC6 features, which achieve optimal performance for a slightly higher DOI of approximately 1.3 mm.

Although the optimal value for the DOI is pretty stable over the different classification methods for each individual feature, it considerably varies over the different feature extraction methods. While the FC6 features show an optimum AUC for a DOI of roughly 1.3 mm, the clinically inspired features generally show an optimal performance for a DOI ranging from 0.5–1 mm. For LS-GLCM, a second optimum DOI can be observed around 1.2–1.3 mm for two classification methods, coinciding with the optimal DOI for the FC6 features. The presence of these two distinct optima might indicate that there are anatomical structures at these depths that contain discriminative information about the histopathology.

At approximately 0.3–0.6 mm depth, the transitions between the epithelium, lamina propria and the muscularis mucosa occur. The (lack of) visibility of these layer transitions in the VLE image can be indicative for the presence of cancer (Swager et al., 2016d). This leads to the hypothesis that the first optimum in Fig. 7 is related to the transitions or boundaries between these tissue layers. For the second optimum in Fig. 7, we could not assign any obvious underlying

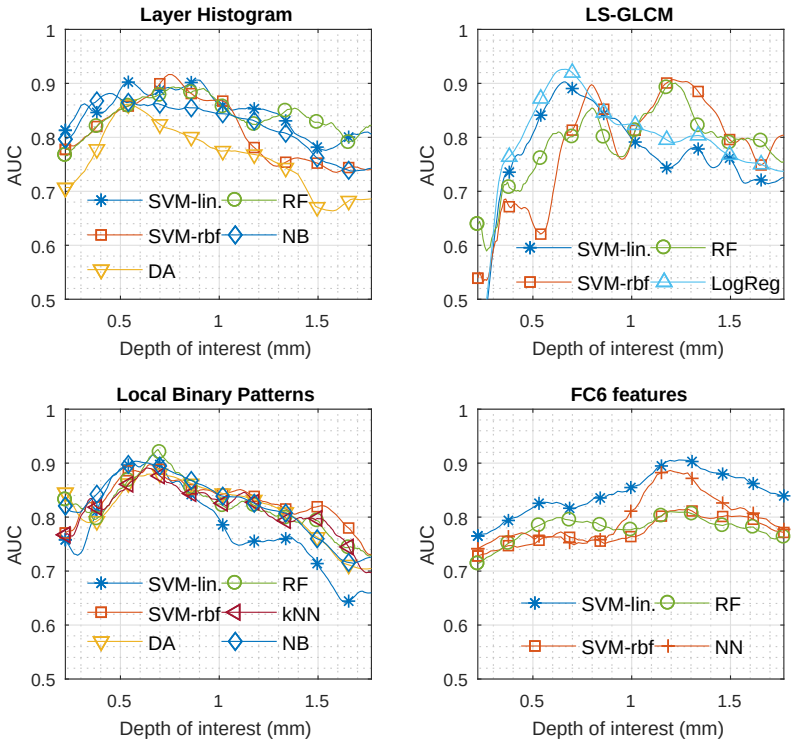

Figure 7: Classification performance of the most promising features from Table 4 for various machine learning methods over an increasing Depth Of Interest (DOI). Each feature generally shows a unique optimum for which most classification methods achieve the highest AUC. The maximal AUCs are 0.92, 0.93, 0.93 and 0.91. for LH, LS-GLCM, LBP and FC6 features, respectively.

anatomical structures. Signals at this depth could potentially originate from the submucosal layer and submucosal structures, such as blood vessels. However, further research is necessary to investigate this phenomenon.

4.4. Feature computation time

In order to obtain an overview of the computation time for each of the evaluated features, we compute the features for the complete data set 1,000 times and derive statistics on the resulting execution times. As hyperparameters can have a large influence on the execution time, we sample these values from their optimal distributions (acquired during hyperparameter optimization for generating Table 4) in order to obtain representative numbers. Note that the number of scanlines used for analysis was a free parameter in this experiment, which considerably reduces the computation time with respect to the default implementations, especially for the state-of-the-art features. The execution time experiments have been carried out using the software package MATLAB 2016a (Mathworks Inc., Natick, Massachusetts, USA) on a desktop PC (hexa-core @3.3 GHz CPU, 16GB RAM, 2GB GPU).

The bar graph shown in Fig. 8 displays the resulting median computation time for each of the features, including the Interquartile Range (IQR) (red bars) and the maximum and minimum computation time for each feature (red triangles). Red asterisks are used for maxima that are outside the range of the plot, which are 1.93 and 3.84 seconds for GLCM and Gabor, respectively. The features from Qi et al. are excluded in Fig. 8, since the computation time for these features is several orders of magnitude larger, i.e. approximately 6 minutes on average for the full set.

From the plot, it is clear that the clinically inspired features generally execute considerably faster than the state-of-the-art alternatives. The proposed BMPA feature shows a median computation time of 24 ms (IQR 22–25), which is remarkably stable over different settings with a minimum and maximum computation time of 20 and 91 ms, respectively. The LH feature exhibits similar properties, with a median computation time of 78 ms (IRQ: 70–88, min-max:

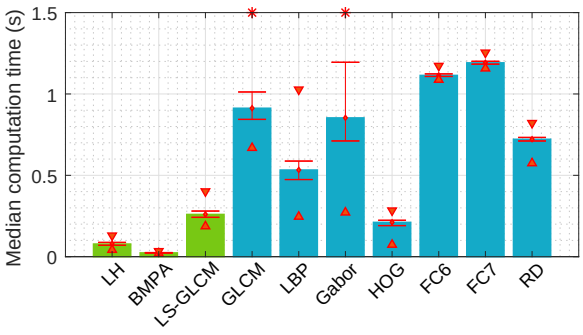

Figure 8: Median computation time for the full set of 60 VLE images for the investigated features, where the proposed features are represented in green and the state-of-the-art alternatives are depicted in blue. The interquartile ranges are indicated by the red lines and the minimum and maximum computation times are indicated by red up- and down-pointing triangles.

42–149). The third clinically inspired feature, namely the LS-GLCM, shows a slightly less stable performance when the min-max range of 187 to 475 ms is considered. This effect can be explained by the varying number of gray levels used for computing the GLCMs. However, with a median computation time of 260 ms (IQR: 242–280), it still computes in a relatively fast and stable range.

The features extracted from the fully connected layers of AlexNet show a considerably longer execution time of 1,111 and 1,191 ms for FC6 and FC7, respectively. As the inputs images are normalized prior to feeding it to the CNN, the highly stable performance is expected for these features. The Gabor-based features show the least stable performance, which is explained by the varying optimal values for the number and size of the filters that were selected during hyperparameter optimization. The evaluated traditional texture features show a reasonable and relatively stable performance in execution time: 913 ms (IQR 844–1,012) for GLCM, 533 ms (IQR 474–588) for LBP and 211 ms (IQR 191–224).

4.5. CAD system vs. clinical experts

For clinical validation, we compare our results to the classification performance of two VLE experts on the same set of images. For a comparison of both

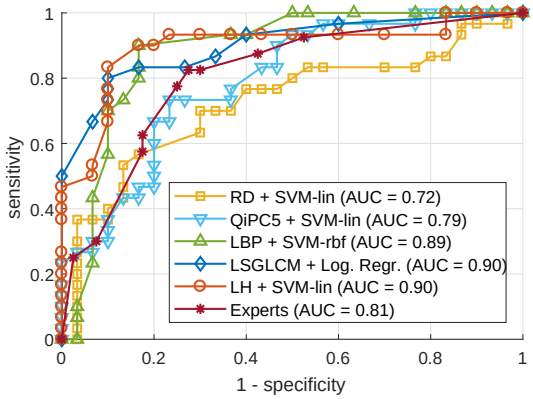

Figure 9: Receiver Operating Characteristic (ROC) curves for some of the most promising combinations (i.e. LBP + non-linear SVM, LSGLCM + log. regression and LH + linear SVM), the best results for state-of-the-art features (i.e. QiPC5 + linear SVM and RD + linear SVM) versus two medical experts.

the sensitivity and specificity, Fig. 9 shows the ROC curves of three of the most promising methods from Table 4 versus the ROC curve for the combined result of the experts. Note that the systems used for this plot result from our second experiment (see Sec. 4.3), in which the number of scanlines used for analysis is included as a free parameter. In addition, we have also included the features from Qi et al. (2010) and the ones proposed by Rodriguez-Diaz & Singh (2015) in this plot.

From Fig. 9 it is clear that the proposed LH and LS-GLCM achieve a superior performance to the experts over the complete range of sensitivity/specificity operating points. This is also reflected in the aggregated performance over this range with an AUC of 0.81 for the experts versus an AUC of 0.90 for both proposed features, respectively. With an AUC of 0.89, the edited LBP shows a comparable classification performance, however, a specificity of 1.0 cannot be achieved using this feature in combination with a SVM and a RBF kernel function. Other classification methods, such as Random Forest or AdaBoost, do achieve maximum specificity for LBP, but these methods show a lower overall AUC (see Table 4).

With a maximum AUC of 0.79 and 0.72, the state-of-the-art features for

EOCT of Qi et al. and for VLE of Rodriguez-Diaz et al., respectively, both demonstrate an inferior classification performance compared to that of the experts. While this poorer performance is mainly expressed in lower AUC scores, it is also reflected in the ROC curves in Fig. 9 of these methods, which are below that of the experts over almost the complete range.

5. Discussion and conclusions

In this study, we have investigated the use of computer-aided methods for the automated analysis of Volumetric Laser Endomicroscopy (VLE) to detect early Barretts cancer. We have evaluated commonly used image analysis features (e.g. Local Binary Patterns, Histogram of Oriented Gradients and Gray-Level Co-occurrence Matrix features), in combination with popular classification methods (e.g. Support Vector Machine, Random Forest and Neural Nets). In this evaluation, we have included features that have been proposed in two recent studies on cancer detection in VLE and Endoscopic OCT. In addition, we have proposed two clinically-inspired features that capture information from a clinical prediction model for scoring VLE images, namely (1) Layer Histogram (LH) and (2) Bin-Median of Pixel Averages (BMPA). Furthermore, we have used this clinical knowledge to adapt the Gray-Level Co-Occurrence Matrix (GLCM) features in order to make it better suitable for cancer detection using VLE data. Finally, we have investigated the use of pre-trained Convolutional Neural Networks (CNNs), for the classification of VLE images.

To the best of our knowledge, this is the first study using a histopathologically validated set of ex-vivo VLE images for the evaluation of computer-aided methods for cancer detection in BE. We employ 60 VLE images (30 dysplastic, 30 non-dysplastic) for validation of the presented methods. Leave-one-out cross-validation is used to generate a prediction score for each image, where the involved hyperparameters are optimized over the training set by means of random trials and five-fold cross-validation. We provide a clear overview of this validation method, using an earlier presented validation framework, in order to

ensure comparability of the results presented in this study with future studies on the classification of VLE images.

Our results show that the proposed clinically-inspired features generally achieve a considerably higher classification performance than state-of-the-art alternatives, over a wide range of classification methods. In particular, a maximal AUC of 0.90, 0.86 and 0.90 is observed over all classification methods for LH, BMPA and LS-GLCM, respectively. This demonstrates that for some problems, relatively simple solutions can lead to an optimal overall performance, which has also been observed for similar problems (Iakovidis & Koulaouzidis, 2014). Both evaluated forms of transfer learning with pre-trained CNNs offered the only alternative with an AUC above 0.8, showing a maximal AUC ranging from 0.81–0.84. This performance is remarkable, given the pre-training with non-VLE data. Traditional shape and texture features like LBP, GLCM, HOG and Gabor features, demonstrated poor classification results with an AUC ranging between 0.50 and 0.73.

While investigating the optimal hyperparameters for the clinically-inspired features, we have found that the top layers of the image generally contain the most discriminative information, thereby confirming a clinical hypothesis that was reported in several medical studies. An additional experiment showed that including the number of scanlines that is used for analysis or Depth of Interest (DOI) - as a free parameter, significantly increased the detection performance for some of the evaluated features. In particular for LBP, the maximal AUC was elevated from 0.64 to 0.89. To further investigate this phenomenon, we carried out an additional experiment, in which we gradually increased the DOI and evaluated the classification performance of the most promising features from the previous experiment. The results showed that there is an optimal DOI for each individual feature typically in the range of 0.5 to 1 mm, for which AUCs in the range of 0.90–0.93 were observed. For the FC features derived from AlexNet, the optimum occurred at a slightly larger DOI of approximately 1.3 mm.

Considering execution time, the proposed clinically-inspired features clearly outperformed the state-of-the art alternatives. With a full-dataset median com-

putation time of 24 ms and 78 ms, respectively, the proposed BMPA and LH features demonstrated an over six-fold speed-up with respect to most state-of-the-art alternatives. Although LBP, FC6-features and Gabor-based features exhibited a comparable classification performance for some classifiers, with median computation times of 533 ms, 853 ms and 1,114 ms, respectively, these features are considerably less attractive.

For clinical validation, we have compared our results to the classification performance of two VLE experts that scored the same set of VLE images. This comparison showed that the proposed features demonstrate a considerably better classification performance over the complete range of possible operating points, with a maximal AUC of 0.90 for the computer-aided methods versus an AUC of 0.81 for the clinical experts. In contrast, the two evaluated methods proposed in related work yielded an AUC of 0.72–0.79 in this comparison, showing a poorer performance than both the experts and the methods proposed in this paper.

With this evaluation of various commonly-used and novel features and classification methods for the classification of VLE images, we hope to present an exhaustive overview of promising methods for this purpose. However, this study also knows some important limitations. As this is the first work in which pathologically validated VLE scans are used, only a limited number of images were available for evaluation of the investigated methods. Using a carefully developed framework for validation and hyperparameter optimization, we have tried to address the limitation of using only a small number of data samples. While empirically-determined hyperparameter values will most likely yield better results, usage of the employed validation framework will almost certainly lead to results that are more robust over different sets of data. A second limitation is that the overview we have presented can never be fully exhaustive. Inevitably, there will be adaptations of the evaluated classification methods that lead to slightly higher scores and given the momentum of the field of deep learning, alternatives to AlexNet as a basis for transfer learning are rapidly emerging. However, we hope that the overview provided in this paper will invoke complementary studies on cancer detection in VLE scans. An additional limitation

arises from the use of ex-vivo data, in which some structures might exhibit a different appearance than on in-vivo VLE data. These differences can arise from e.g. different mechanical properties of the resected tissue vs. the intact organ, or a different interaction between the tissue and the balloon in the employed fixture versus inside the esophagus. Further research is required to confirm the presented results on in-vivo VLE. Finally, the number of vertical scanlines that are employed for analysis varied per image and was limited by the markers used for histopathology correlation. Hence, although we have identified the optimal number of *horizontal scanlines* in our experiments, the optimal number of *vertical* scanlines for analysis is still open for further investigation. This experiment will reveal useful clues for implementation of a CAD system for VLE, where a sliding window approach could be employed for analysis of the complete circumferential scan of typically 4,096 vertical lines.

In conclusion, we present a large scale, thorough evaluation of a broad set of existing and novel features in combination with various popular classification methods for early cancer detection using VLE images. We demonstrate that computer-aided interpretation of VLE scans is feasible and it can clearly outperform human experts in distinguishing early cancerous tissue from non-dysplastic tissue based on ex-vivo VLE images. Our results show that the use of clinical prediction models for the development of such methods can be of great benefit for both classification accuracy and execution time. This observation is especially true, when only a limited amount of data is available. Furthermore, we identify an optimal range of approximately 0.5–1 mm scanning depth in the tissue for the classification of neoplasms, which can be linked to the presence of anatomical structures like the transitions between tissue layers, as well as a degrading SNR for a deeper scanning depth. Therefore, we propose to include the *scanlines of interest* as a free parameter during hyperparameter optimization for VLE CAD systems, and we demonstrate that this can significantly boost the classification performance. Future studies should further exploit and confirm this depth range and use information along the axial dimension in order to achieve a better detection performance. Furthermore, a large in-vivo dataset

with histological correlates should be constructed to expand this work to full in-vivo VLE scans.

Acknowledgments

The authors gratefully acknowledge the support from NinePoint Medical (NinePoint Medical Inc., Bedford, MA, USA), who granted us the permission to use the VLE scans for this work.

References

Bergstra, J., & Yoshua, B. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, *13*, 281–305.

Breiman, L. (2001). Random forests. *Mach. Learn.*, *45*, 5–32. doi:10.1023/A:1010933404324.

Cook, M. B., Shaheen, N. J., Anderson, L. A., Giffen, C., Chow, W. H., Vaughan, T. L., Whiteman, D. C., & Corley, D. A. (2012). Cigarette smoking increases risk of barrett's esophagus: An analysis of the barrett's and esophageal adenocarcinoma consortium. *Gastroenterology*, *142*, 744–753. doi:10.1053/j.gastro.2011.12.049.

Corley, D. A., Mehtani, K., Quesenberry, C., Zhao, W., De Boer, J., & Weiss, N. S. (2013). Impact of endoscopic surveillance on mortality from barrett's esophagus-associated esophageal adenocarcinomas. *Gastroenterology*, *145*, 312–319.e1. doi:10.1053/j.gastro.2013.05.004. arXiv:NIHMS150003.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, *20*, 273–297. doi:10.1007/BF00994018.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 886–893). IEEE volume 1. doi:10.1109/CVPR.2005.177.

Deng, J. D. J., Dong, W. D. W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, (pp. 2–9). doi:10.1109/CVPR.2009.5206848.

Ell, C., May, A., Pech, O., Gossner, L., Guenter, E., Behrens, A., Nachbar, L., Huijsmans, J., Vieth, M., & Stolte, M. (2007). Curative endoscopic resection of early esophageal adenocarcinomas (barrett's cancer). *Gastrointest. Endosc.*, *65*, 3–10. doi:10.1016/j.gie.2006.04.033.

Fogel, I., & Sagi, D. (1989). Gabor filters as texture discriminator. *Biol. Cybern.*, *61*, 103–113.

Freund, Y., & Schapire, R. E. (1995). Computational learning theory. *Comput. Learn. theory*, *904*, 23–37. doi:10.1007/3-540-59119-2.

Gonzalo, N., Tearney, G. J., Serruys, P. W., van Soest, G., Okamura, T., García-García, H. M., Jan van Geuns, R., van der Ent, M., Ligthart, J., Bouma, B. E., & Regar, E. (2010). Second-generation optical coherence tomography in clinical practice. high-speed data acquisition is highly reproducible in patients undergoing percutaneous coronary intervention. *Rev. Esp. Cardiol.*, *63*, 893–903. doi:10.1016/S1885-5857(10)70183-3.

Gora, M. J., Sauk, J. S., Carruth, R. W., Gallagher, K. A., Melissa, J., Nishioka, N. S., Kava, L. E., Rosenberg, M., Bouma, B. E., & Tearney, G. J. (2013). Tethered capsule endomicroscopy enables imaging of gastrointestinal tract microstructure. *Nat. Med.*, *19*, 238–240. doi:10.1038/nm.3052.Tethered.

Haralick, R. M., & Shanmugam, K. (1973). Textural features for image classification. *IEEE Trans. Syst., Man, Cybern., Syst*, *3*, 610 – 621.

Hayeck, T. J., Kong, C. Y., Spechler, S. J., Gazelle, G. S., & Hur, C. (2010). The prevalence of barrett's esophagus in the us: Estimates from a simulation model confirmed by seer data. *Dis. Esophagus*, *23*, 451–457. doi:10.1111/j.1442-2050.2010.01054.x.

Iakovidis, D. K., & Koulaouzidis, A. (2014). Automatic lesion detection in wireless capsule endoscopya simple solution for a complex problem. In *Image Processing (ICIP), 2014 IEEE International Conference on* (pp. 2236–2240). IEEE.

Jannin, P., Grova, C., & Maurer, C. R. (2006). Model for defining and reporting reference-based validation protocols in medical image processing. *International Journal of Computer Assisted Radiology and Surgery*, *1*, 63–73.

Klomp, S. R., van der Sommen, F., Swager, A.-f., Zinger, S., Schoon, E. J., Curvers, W. L., Bergman, J. J. G. H. M., & de With, P. H. N. (2017). Evaluation of image features and classiffication methods for barrett's cancer detection using vle imaging. In *Proc. SPIE 10134, Medical Imaging 2017, Computer-Aided Diagnosis* (p. 101340D).

Krizhevsky, A., Hinton, G. E., & Sutskever, I. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, (pp. 1–9). doi:`http://dx.doi.org/10.1016/j.protcy.2014.09.007`.

Lagergren, J. (2011). Influence of obesity on the risk of esophageal disorders. *Nature Reviews Gastroenterology and Hepatology*, *8*, 340–347.

Lu, L., Shin, H.-c., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection : Cnn architectures , dataset characteristics and transfer learning. *IEEE Trans. Med. Imag.*, *35*, 1285–1298. doi:`10.1109/TMI.2016.2528162`.

Odze, R. D. (2006). Diagnosis and grading of dysplasia in barrett's oesophagus. *J. Clin. Pathol.*, *59*, 1029–38. doi:`10.1136/jcp.2005.035337`.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.*, *29*, 51–59. doi:`10.1016/0031-3203(95)00067-4`.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. *Comput. Vis. Pattern Recognit. (CVPR), 2014 IEEE Conf.*, (pp. 1717–1724). doi:10.1109/CVPR.2014.222.

Orlando, J. I., Prokofyeva, E., del Fresnob, M., & Blaschko, M. (2017). Convolutional neural network transfer for automated glaucoma identification. In *12th International Symposium on Medical Information Processing and Analysis* (pp. 101600U–101600U). International Society for Optics and Photonics.

Peters, F. P., Curvers, W. L., Rosmolen, W. D., de Vries, C. E., ten Kate, F. J. W., Krishnadath, K. K., Fockens, P., & Bergman, J. J. G. H. M. (2008). Surveillance history of endoscopically treated patients with early barrett's neoplasia: Nonadherence to the seattle biopsy protocol leads to sampling error. *Dis. Esophagus*, *21*, 475–479. doi:10.1111/j.1442-2050.2008.00813.x.

Qi, X., Pan, Y., Sivak, M. V., Willis, J. E., Isenberg, G., & Rollins, A. M. (2010). Image analysis for classification of dysplasia in barrett's esophagus using endoscopic optical coherence tomography. *Biomed. Opt. Express*, *1*, 825–847.

Reid, B. J., Blount, P. L., Feng, Z., & Levine, D. S. (2000). Optimizing endoscopic biopsy detection of early cancers in barrett's high-grade dysplasia. *The American journal of gastroenterology*, *95*, 3089–3096.

Rodriguez-Diaz, E., & Singh, S. K. (2015). 422 computer-assisted image interpretation of volumetric laser endomicroscopy in barrett's esophagus. *Gastroenterology*, *148*, S91–92. doi:10.1016/S0016-5085(15)30316-4.

Rollins, A. M., Ung-Arunyawee, R., Chak, A., Wong, R. C., Kobayashi, K., Sivak, M. V., & Izatt, J. a. (1999). Real-time in vivo imaging of human gastrointestinal ultrastructure by use of endoscopic optical coherence tomography with a novel efficient interferometer design. *Opt. Lett.*, *24*, 1358–1360. doi:10.1364/OL.24.001358.

Rollins, A. M., Yazdanfar, S., Kulkarni, M., Ung-Arunyawee, R., & Izatt, J. (1998). In vivo video rate optical coherence tomography. *Opt. Express*, *3*, 219–229. doi:10.1364/OE.3.000219.

Shaheen, N. J., & Richter, J. E. (2009). Barrett's oesophagus. *The Lancet*, *373*, 850–861. doi:10.1016/S0140-6736(09)60487-6.

van Soest, E. M., Dieleman, J. P., Siersema, P. D., Sturkenboom, M. C. J. M., & Kuipers, E. J. (2005). Increasing incidence of barrett's oesophagus in the general population. *Gut*, *54*, 1062–6. doi:10.1136/gut.2004.063685.

Solaymani-Dodaran, M., Logan, R. F. A., West, J., Card, T., & Coupland, C. (2004). Risk of oesophageal cancer in barrett's oesophagus and gastro-oesophageal reflux. *Gut*, *53*, 1070–4. doi:10.1136/gut.2003.028076.

Swager, A., Boerwinkel, D. F., de Bruin, D. M., Weusten, B. L., Faber, D. J., Meijer, S. L., van Leeuwen, T. G., Curvers, W. L., & Bergman, J. J. (2016a). Volumetric laser endomicroscopy in barrett's esophagus: a feasibility study on histological correlation. *Dis Esophagus*, *29*, 505–512. doi:10.1111/dote.12371.

Swager, A., van Oijen, M., Tearney, G., Leggett, C., Meijer, S., Bergman, J., & Curvers, W. (2016b). How good are experts in identifying early barrett's neoplasia in endoscopic resection specimens using volumetric laser endomicroscopy? *Gastroenterology*, *150*, S628. doi:10.1016/S0016-5085(16)32158-8.

Swager, A., van Oijen, M., Tearney, G., Leggett, C., Meijer, S., Bergman, J., & Curvers, W. (2016c). How good are experts in identifying endoscopically visible early barrett's neoplasia on in vivo volumetric laser endomicroscopy? *Gastrointest Endosc*, *83*, AB573. doi:10.1016/j.gie.2016.03.1180.

Swager, A., Tearney, G., Leggett, C., van Oijen, M., Meijer, S., Weusten, B., Curvers, W., & Bergman, J. (2016d). Identification of volumetric laser endomicroscopy features predictive for early neoplasia in barrett's esophagus

using high-quality histological correlation. *Gastrointest Endosc*, (p. in press). doi:10.1016/j.gie.2016.09.012.

Ughi, G. J., Gora, M. J., Swager, A.-f., Soomro, A., Grant, C., Tiernan, A., Rosenberg, M., Sauk, J. S., Nishioka, N. S., & Tearney, G. J. (2016). Automated segmentation and characterization of esophageal wall in vivo by tethered capsule optical coherence tomography endomicroscopy. *Biomed. Opt. Express*, *7*, 660–665. doi:10.1364/BOE.7.000409.

Wolfsen, H. C., Sharma, P., Wallace, M. B., Leggett, C., Tearney, G., & Wang, K. K. (2015). Safety and feasibility of volumetric laser endomicroscopy in patients with barrett's esophagus (with videos). *Gastrointest Endosc*, *3*, 1–10. doi:10.1016/j.gie.2015.03.1968.