

# Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application

Pranita Pradhan<sup>1,2</sup>, Tobias Meyer<sup>2</sup>, Michael Vieth<sup>3</sup>, Andreas Stallmach<sup>6</sup>, Maximilian Waldner<sup>4,5</sup>, Michael Schmitt<sup>1</sup>, Juergen Popp<sup>1,2</sup> and Thomas Bocklitz<sup>1,2</sup>

<sup>1</sup>*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University, Jena, Germany*

<sup>2</sup>*Leibniz Institute of Photonic Technology, Member of Leibniz Health Technologies Jena, Germany*

<sup>3</sup>*Institute of Pathology, Klinikum Bayreuth, Bayreuth, Germany*

<sup>4</sup>*Erlangen Graduate School in Advanced Optical Technologies (SAOT), Friedrich-Alexander University of Erlangen-Nuremberg, Germany*

<sup>5</sup>*Medical Department I, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany*

<sup>6</sup>*Department of Internal Medicine IV (Gastroenterology, Hepatology, and Infectious Diseases), Jena University Hospital, Jena, Germany*

**Keywords:** Semantic Segmentation, Non-linear Multimodal Imaging, Inflammatory Bowel Disease.

**Abstract:** Non-linear multimodal imaging, the combination of coherent anti-stokes Raman scattering (CARS), two-photon excited fluorescence (TPEF) and second harmonic generation (SHG), has shown its potential to assist the diagnosis of different inflammatory bowel diseases (IBDs). This label-free imaging technique can support the 'gold-standard' techniques such as colonoscopy and histopathology to ensure an IBD diagnosis in clinical environment. Moreover, non-linear multimodal imaging can measure biomolecular changes in different tissue regions such as crypt and mucosa region, which serve as a predictive marker for IBD severity. To achieve a real-time assessment of IBD severity, an automatic segmentation of the crypt and mucosa regions is needed. In this paper, we semantically segment the crypt and mucosa region using a deep neural network. We utilized the SegNet architecture (Badrinarayanan et al., 2015) and compared its results with a classical machine learning approach. Our trained SegNet model achieved an overall F1 score of 0.75. This model outperformed the classical machine learning approach for the segmentation of the crypt and mucosa region in our study.

## 1 INTRODUCTION

Histopathological examination represents the 'gold-standard' for diagnosing inflammatory bowel disease (IBD), where the quantification of colonic inflammation is based on the visual appearance of the tissue. However, histopathology delays the diagnosis due to a long sample preparation protocol that includes taking biopsies, tissue embedding, tissue sectioning and staining. In this regard, label-free imaging techniques like multiphoton microscopy (MPM) has been recognized as an *in vivo* imaging technique for IBD diagnostics (Schürmann et al., 2013) (Chernavskaia et al., 2016) (Waldner et al., 2017). These label-free techniques allow a non-destructive investigation of biomolecules in tissue with high tissue penetration depth and spatial resolution (Cicchi and Pavone, 2014) (Vogler et al., 2015).

In the past, MPM techniques like two-photon ex-

cited fluorescence (TPEF) and second harmonic generation (SHG) along with coherent anti-stokes Raman scattering (CARS) were used to visualize biomolecular changes associated with IBDs. Biomolecular information like changed CARS, TPEF and SHG signal intensity along with the crypt morphometries served as predictive marker for an inflamed colon tissue. Likewise, Chernavskaia et al. presented a predictive modelling of histological indexes associated with IBD based on the biomolecular changes of the crypt and mucosa region. Such an automatic predictive modelling of histological indexes is beneficial to accelerate IBD diagnosis. In the work of Chernavskaia et al., the crypt and mucosa region were manually segmented, so a full automatization of the predictive modelling of histological indexes requires a semantic segmentation of crypt and mucosa region without manual effort.

Semantic segmentation of the crypt and mucosa

region is challenging due to several reasons. First, shape irregularities of the crypts add a large biological variance to the data. For example, an inflamed colon tissue reveals crypt deformations and a loss of crypt density, whereas regularly shaped crypts can be found in healthy colon tissue. Second, the crypts are located within the mucosa and therefore the two regions overlap, making the classification even more difficult. Third, the identification of the crypt boundaries is complicated as they are closely located to each other. Lastly, there is a limited amount of annotated medical data, which captures various tissue structures of an inflamed colon. The above mentioned reasons lead to a high morphological variance of the tissue structures thereby making the semantic segmentation of the crypt and mucosa challenging. For this segmentation task machine learning algorithms can be utilized, either classical machine learning or deep learning. Due to this challenging segmentation task mentioned above a high domain-specific representation is needed, which is difficult to obtain using hand-crafted features in classical machine learning.

On the other hand, deep convolutional neural networks (DCNNs) are capable of learning domain-specific representations of an image and have achieved successful results in image classification (Babaie et al., 2017) (Krizhevsky et al., 2012a), object recognition (Pathak et al., 2018) and semantic segmentation (Roth et al., 2015) (Long et al., 2014). Existing DCNNs like U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2015) have gained state-of-the-art results in biomedical image segmentation and in the field of digital pathology (Janowczyk and Madabhushi, 2016). In this study, we utilize DCNNs to semantically segment multimodal images into biologically significant regions for assisting the predictive modelling of histological indexes. Furthermore, we compare the segmentation results obtained by a DCNN with a classical machine learning approach.

The paper is organized as follows: In section 2 we introduce the previous work related to gland segmentation of histology images, in section 3 we outline our multimodal image dataset and our segmentation workflow. This is followed by a description of the evaluation metrics and a presentation of the results in section 4. We discuss and conclude our work in section 5 and 6, respectively.

## 2 RELATED WORK

Medical Image Segmentation (MIS) can be utilized for numerous applications like identifying tissue

structures, cell counting, lesion and tumour detection (Norouzi et al., 2014). The approaches for MIS can be categorized into three types. First, the segmentation using classical image processing techniques like thresholding, morphological operations and watershed transform (Wu et al., 2005). Second, training a classification model based on handcrafted image features (classical machine learning) like statistical features, grey level co-occurrence matrix features and local binary patterns (Farjam et al., 2007) (Doyle et al., 2007) (Naik et al., 2008) (Guo et al., 2018). And the third approach is the segmentation using high-level features obtained by a DCNN (Kainz et al., 2017) (Awan et al., 2017) (Chen et al., 2016).

Wu et al. utilized classical image processing algorithms including thresholding and seeded region growing for segmentation of the human intestinal glands. However, this method considered a prior knowledge of the morphological structures of the gland and was qualitatively evaluated (Wu et al., 2005). In another approach by Peng et al., k-means clustering and morphological operations were used to segment the prostate glandular structures. Based on these structures a linear classifier to distinguish normal and malignant glands was constructed (Peng et al., 2011). Peng et al. utilized a k-means clustering algorithm directly on the colour information. Therefore, the approach is not incorporating shape and texture features, which are important for crypt segmentation.

In the contribution by Farzam et al. and Doyle et al., texture, shape and graph-based features were extracted within a classical machine learning approach. Based on these features, a linear classifier to distinguish different pathological tissue sections of the prostate cancer patients was built (Farjam et al., 2007) (Doyle et al., 2007). In the work presented by Naik et al., a Bayesian classifier was used to identify true lumen areas and the false positive lumen areas were removed by applying size and structure constraints. Using the true lumen area, a level set curve (Li et al., 2005) was initialized and evolved until the interior boundary of the nuclei. Morphological features (like distance ratio, compactness, area overlap ratio) were calculated based on the boundaries of the detected lumen area and nuclei. This was followed by a manifold learning scheme called Graph Embedding algorithm (Shi and Malik, 2000) to reduce the dimension of the feature space. Based on the reduced feature space, a support vector machine (SVM) algorithm was used to classify the images into different Gleason grades of prostate cancer (Naik et al., 2008). The above-mentioned methods efficiently segmented regularly shaped gland structures but faced challenges in

segmenting irregularly shaped gland structure.

To tackle this problem, Gunduz-Demir et al. proposed an object-graph based approach that relies on decomposing an image into objects. Their approach used a three-step region growing algorithm, followed by boundary detection and false region elimination (Gunduz-Demir et al., 2010). In another work by Sirinukunwattana et al. (Sirinukunwattana et al., 2015), a Random Polygons Model (RPM) to segment glandular structure in human colon tissue was formulated. The glandular structures were modelled as polygons with random vertices that were located on the cell nuclei within the epithelium. Based on the spatial arrangement of the epithelial nuclei and neighbouring nuclei, an inference of the RPM was made via Reversible-Jump Markov Chain Monte Carlo simulation. False positive polygons were removed by post-processing procedures (Sirinukunwattana et al., 2015). While this technique is stochastic in nature, it can produce different results for the same image and thus a robust approach is needed.

Approaches using DCNNs like AlexNet (Krizhevsky et al., 2012b), VGGNet (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2014), U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2015) have achieved promising results in MIS. The recent *MICCAI 2015 Gland Segmentation Challenge* presented several innovative algorithms for segmentation of colon glands in histology images (Sirinukunwattana et al., 2015). Chen et al. achieved state-of-the-art performance on the Warwick-QU colon adenocarcinoma dataset by integrating multi-level feature representation with Fully Convolutional Network (FCN) (Chen et al., 2016). Likewise, Kainz et al. used two DCNN that were inspired by the LeNet-5 architecture (LeCun et al., 1998) (Kainz et al., 2017). The first DCNN was used to separate the closely located gland structures and the second DCNN was used to distinguish gland and non-gland regions (Kainz et al., 2017). In Awan et al., a DCNN was used to mark gland boundaries and based on the glandular shape, a two-class and three-class classification model for colorectal adenocarcinoma using histology image was designed (Awan et al., 2017).

In this contribution, we intend to use a SegNet model (Badrinarayanan et al., 2015) for the semantic segmentation of non-linear multimodal images into four distinct regions. Our method is different to the described previous works in the following ways:

- This work is the first to implement semantic segmentation of crypts and mucosa region in non-linear multimodal images. All the above methods have been implemented on H&E (Hematoxylin

and Eosin) stained image which needs a long sample preparation time and leads to sample destruction. In contrast, label-free non-linear multimodal imaging can be used as an *in vivo* technique and its automatic tissue classification can provide a real-time histological index prediction.

- Our method is adapted to multimodal images that show low SNR and are hard to analyze (Vogler et al., 2015).
- Unlike other machine learning methods, we perform a four-class semantic segmentation of multimodal images. In addition to the crypt region we also segment the mucosa region that can be used to assign a histological index.

## 3 MATERIAL AND METHODS

### 3.1 Dataset

For this study, we utilized an already published dataset composed of twenty multimodal images sampled from twenty IBD patients. Each multimodal image was converted to an RGB image, which was constructed based on the three modalities, CARS at 2850  $\text{cm}^{-1}$  (red channel), TPEF at 458 nm (green channel) and SHG at 415 nm (blue channel). We followed the same image pre-processing steps as explained by Chernavskaia et al. which included downsampling of the multimodal image followed by median filtering, uneven illumination correction (Legesse et al., 2015), background estimation and contrast adjustment (Chernavskaia et al., 2016). The dataset was randomly divided into 11 training, 5 validation and 4 test images. The training dataset was augmented using a rotation angle 60° and 90°.

A histological index between 0 (healthy) and 2 (severe disease) based on crypt architecture, mucosal chronicity and activity was assigned to every image by a trained pathologist. In addition, manually annotated crypt and mucosa regions were obtained as a false-colour image (as shown in figure 1). The manually annotated image is partitioned into four subregions  $R_l$ ,  $l = \{0, 1, 2, 3\}$ : mucosa without crypt ( $R_0$ ) labelled as 0, crypt ( $R_1$ ) labelled as 1, non-mucosa ( $R_2$ ) labelled as 2 and background ( $R_3$ ) labelled as 3.

Table 1: Overview of the dataset.

Dataset	# images	Rotation angle	Total # patches	# selected patches
Train	11	0°, 60°, 90°	9.228	3.990
Validation	5	0°	1.168	1.168
Test	4	0°	880	880

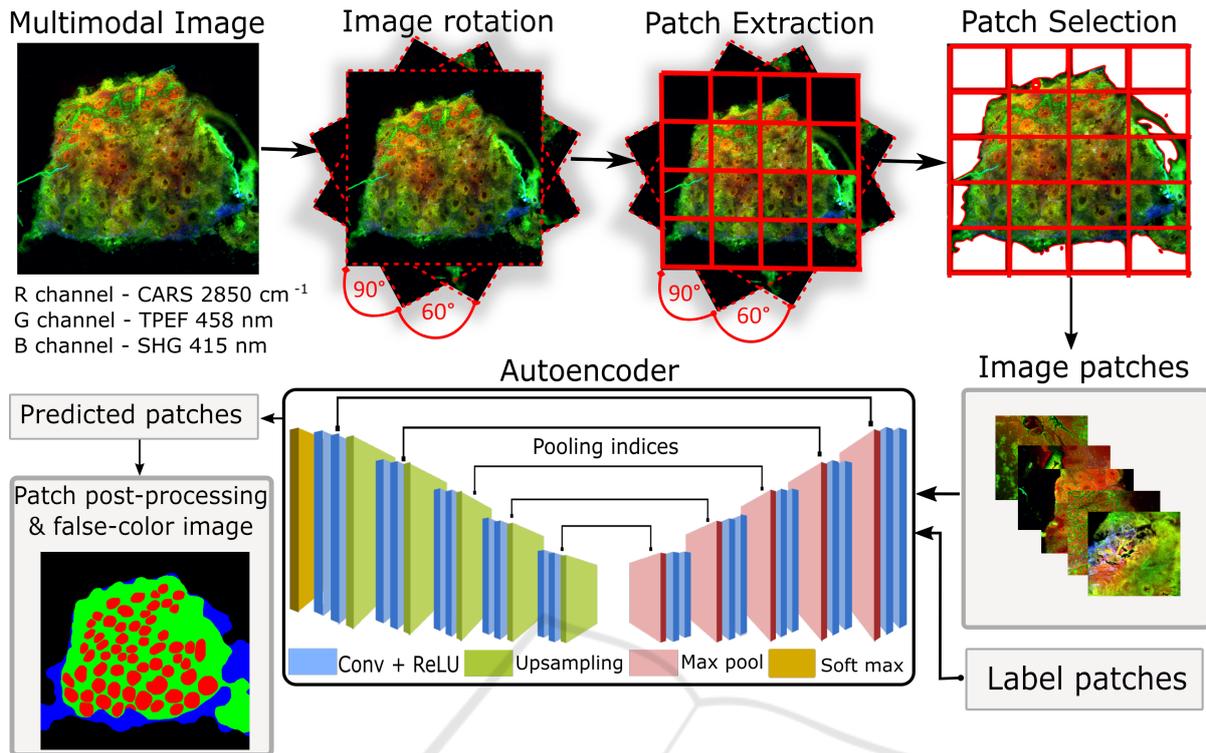


Figure 1: Overview of our proposed segmentation workflow. (1) Data augmentation using random rotations of the multimodal image. (2) Extracting patches of size  $256 \times 256$ . (3) Selecting patches including only tissue regions. (4) Training a SegNet model using the patches and their label map. (5) Combining the predicted patches into a segmentation map. The segmentation map is depicted as false-colour image where green, red, blue and black represent  $R_0$  (mucosa without crypt),  $R_1$  (crypt),  $R_2$  (non-mucosa),  $R_3$  (background).

### 3.2 Classical Machine Learning Approach

We formulate the tissue segmentation task as a pixel classification problem. For the pixel classification we extracted texture features locally using first-order statistical moments of the histogram (Guo et al., 2018). These features give an intuitive understanding of the underlying texture in different tissue regions locally and are easy to calculate.

A set of 11 texture features (mean, standard deviation, skewness, kurtosis, median, energy, entropy, RMS, variance, minimum, maximum) was calculated using a window of  $(5 \times 5)$  around each pixel for every channel of the multimodal images. This led to 33 texture features (11 features  $\times$  3 channels) per pixel. To reduce the computational complexity, the multimodal images were resized by a factor of 3 and features were calculated for every fifth pixel. Background pixels were excluded with the help of a binary mask which was generated using k-means clustering ( $k = 2$ ) and morphological operations. Morphological operations composed of dilation (kernel size:  $9 \times 9$ , iterations: 2), closing (kernel size:  $9 \times 9$ , iterations: 2) and open-

ing (kernel size:  $3 \times 3$ , iterations: 2) were used. A linear classification model based on a principal component analysis in combination with a linear discriminant analysis (PCA-LDA) was built using the 33 texture features. The model was trained using texture features extracted from 11 training images and its performance was evaluated on the remaining images.

All the computations were performed using Python machine learning library Scikit-learn (Pedregosa et al., 2011), Numpy (Travis E, 2006) and Scipy (Jones et al., 2001). The total execution time was approximately 20 hours on a commercially available PC system Intel® Core™ i5-7500 CPU, 3.40 GHz, 16GB RAM.

### 3.3 Deep Learning Approach

#### 3.3.1 Encoder-decoder Architecture

We used the SegNet architecture proposed by Badrinarayanan et al. (Badrinarayanan et al., 2015). This network proceeds with an encoder and a decoder with 13 convolutional layers in each. The input to the first layer of the encoder is an image of size  $M \times N$ .

An activation map of the  $(m + 1)^{th}$  encoder layer is given as:

$$x_{m+1} = [\text{MAX}\{\text{ReLU}[\text{CONV}_m\{x_m\} + b_m]\}] \quad (1)$$

$\text{CONV}\{\cdot\}$  is the convolution operator,  $\text{ReLU}[\cdot]$  is the rectified linear unit function:  $f(x) = \max(0, x)$ ,  $\text{MAX}\{\cdot\}$  is the max pooling layer with a receptive field of  $(2 \times 2)$  and a stride of 2,  $m \in \{1, 2, \dots, 13\}$  represent index of the convolution layer and  $b_m$  is the learned bias of  $m^{th}$  layer. The decoder consists of an upsampling, convolution and batch normalization layer. An activation map  $y_{m+1}$  of the  $(m + 1)^{th}$  decoder layer is given as:

$$y_{m+1} = \text{NORM}[\text{CONV}_m\{\text{US}(y_m) + b_m\}] \quad (2)$$

Here,  $y_m$  is the activation map of  $m^{th}$  layer,  $\text{US}(\cdot)$  is the upsampling layer and  $\text{NORM}[\cdot]$  is the batch normalization layer. The features from the last layer of the decoder are fed to a softmax activation layer. The output of the softmax layer is a  $c$  channel image, where  $c$  represents the number of classes. A segmented image is generated by assigning each pixel to the class, which had the maximum probability.

### 3.3.2 Segmentation Workflow

The semantic segmentation of the multimodal image into four regions was performed using a patch based convolutional neural network approach (Jaremenko et al., 2015). This workflow was implemented using Python with the Deep Learning Library Keras (Chollet et al., 2015) with Theano backend (Theano Development Team, 2016).

The model construction started with a patch extraction and a patch selection. Each multimodal image (denoted by  $\mathbf{I}$ ) of size  $M \times N$  was divided into patches (denoted by  $\mathbf{P}$ ) of size  $256 \times 256$  pixels without any overlap. The number of patches per image was different as our images were different in size. Each patch  $P_i$  can be partitioned into at most four sub-regions (denoted as  $R_0, R_1, R_2, R_3$ ) such that  $\bigcup_{l=0}^3 R_l = P$  where  $\bigcup$  represents union set. To remove the background patches from the training set, a homogeneity factor ( $H = \sum_{s=1}^K \sum_{t=1}^K \frac{p_{st}}{1+|s-t|}$ , where  $p_{st}$  is the probability of relative position of a pixel pair,  $K$  is the distinct intensity level) was calculated for each patch and a threshold of 60% was optimized such that all the patches belonged to the tissue section. This led to 9.228 training patches. The patches from validation and test set were used for model evaluation. Table 1 shows an overview of the dataset and the patches.

For patch training, the SegNet model (Badrinarayanan et al., 2015) was trained end-to-end to classify the pixels of the multimodal patch into the

four regions. The input of the SegNet model (Badrinarayanan et al., 2015) was a multimodal patch and the output of the model was a segmented patch. The weights of the encoder layers were initialized using VGG16 model pre-trained on ImageNet dataset (Simonyan and Zisserman, 2014) (Russakovsky et al., 2014). We trained the model using a mini-batch of five patches and the stochastic gradient descent optimizer to minimize the cross-entropy loss function. The learning rate was set to  $10^{-4}$  and the training was terminated when the validation loss converged. The total training time was approximately 3 hours on a single NVIDIA GeForce GTX 1060 (6GB memory).

The model performance was evaluated on the test patches. The predicted patches were combined into a whole image, which was called 'segmented map'. This segmented map was post-processed using morphological operations like removing blobs and filling holes. The segmented map was visualized as a false-colour image, wherein the regions  $R_0$  (mucosa without crypt),  $R_1$  (crypt),  $R_2$  (non-mucosa),  $R_3$  (background) were indicated in green, red, blue and black, respectively. The segmented map was visually evaluated, and the quantitative evaluation of the segmented regions was performed by calculating the F1 score and recall as explained in section 4.

## 4 RESULTS

### 4.1 Qualitative Evaluation

We visually inspected the segmented map of the validation and the test images. The segmentation of regularly shaped crypts for images with architecture = 0, chronicity = 0, activity = 0, was good. On the other hand, the model performed poorly for segmenting irregularly shaped crypts observed in architecture > 0 and chronicity > 0. The segmentation of the mucosa region was good for all images. We believe that training the SegNet model (Badrinarayanan et al., 2015) with more images of histological index greater than 0 can improve the segmentation performance for images with higher histological indexes, e.g. with stronger altered crypt structures. Also a good quality image with high SNR is required for training the model.

### 4.2 Quantitative Evaluation

One of the evaluation metrics for classification problems is accuracy, which is misleading for unbalanced class sizes. In our case the number of background

Table 2: Comparison of machine learning and deep learning prediction for  $R_0$  (mucosa without crypt),  $R_1$  (crypt),  $R_2$  (non-mucosa) and  $R_3$  (background) based on F1 score and recall. The values correspond to mean ( $\pm$  standard deviation). The number in bold is the best score for classical machine learning and deep learning.

	F1 score				Recall			
	$R_0$ (mucosa without crypt)	$R_1$ (crypt)	$R_2$ (non-mucosa)	$R_3$ (background)	$R_0$ (mucosa without crypt)	$R_1$ (crypt)	$R_2$ (non-mucosa)	$R_3$ (background)
<b>Deep learning</b>	<b>0.55</b> ( $\pm 0.17$ )	<b>0.63</b> ( $\pm 0.13$ )	<b>0.64</b> ( $\pm 0.14$ )	0.95 ( $\pm 0.02$ )	<b>0.57</b> ( $\pm 0.22$ )	<b>0.63</b> ( $\pm 0.17$ )	<b>0.76</b> ( $\pm 0.15$ )	0.92 ( $\pm 0.03$ )
<b>Machine learning</b>	0.27 ( $\pm 0.11$ )	0.18 ( $\pm 0.18$ )	0.56 ( $\pm 0.23$ )	<b>0.96</b> ( $\pm 0.02$ )	0.45 ( $\pm 0.12$ )	0.55 ( $\pm 0.24$ )	0.44 ( $\pm 0.25$ )	<b>1</b> ( $\pm 0.00$ )

pixels is much higher than the number of pixels belonging to the crypt region, hence accuracy is an inappropriate choice for an evaluation metric.

We evaluated the model performance using F1 score and recall for each region. The number of pixels in the segmented map that intersect with its manually annotated image is considered as true positive. The higher the number of true positives, higher is the F1 score and recall. The two metrics are given by  $F_1 = 2TP/(2TP+FN+FP)$  and  $Recall = TP/(TP+FN)$ , where TP is true positive, FP is false positive and FN is false negative.

In table 2, we report the mean and standard deviation of the F1 score and recall for each region of the segmented maps after post-processing. These values did not change significantly before and after post-processing. The overall segmentation accuracy for the region  $R_0$ ,  $R_1$ ,  $R_2$  shows that SegNet model (Badrinarayanan et al., 2015) outperformed the classical machine learning approach. Specifically, an overlap of the predicted crypts with manually annotated crypts was 18% and 63% using classical machine learning and deep learning, respectively.

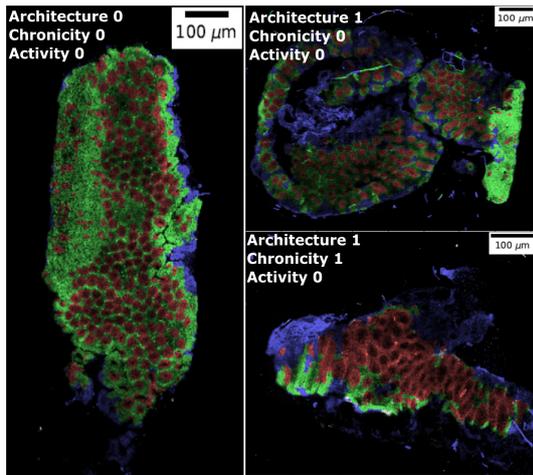


Figure 2: Segmented map superposed with the grey scale multimodal image along with the histological indexes. Regularly shaped crypts (left image) are well segmented whereas a poor segmentation of irregularly shaped crypts is observed (right bottom).

In table 3, we report the F1 scores of  $R_0$  (mucosa without crypt) and  $R_1$  (crypt) for the test and validation image along with its histological indexes for both learning approaches. The evaluation of the segmented regions is important as chronicity of the mucosa and the architecture of crypts serve as predictive marker for the IBD diagnosis. From the results, we observe that using deep learning the two regions for different histological index levels are efficiently segmented. Furthermore, the SegNet model (Badrinarayanan et al., 2015) shows remarkable performance on images with lower histological index (like test image 4). The F1 score for both regions  $R_0$  (mucosa without crypt) and  $R_1$  (crypt) were 0.75 and 0.71, respectively.

However, the trained SegNet model (Badrinarayanan et al., 2015) shows under-segmentation for some cases shown in table 4. The first column is a multimodal image patch, the second column is the manual annotation, the third and fourth column is the prediction by classical machine learning and deep learning (without post-processing), respectively. The crypt segmentation using the handcrafted features show worse performance, whereas the SegNet (Badrinarayanan et al., 2015) model can efficiently segment regularly shaped and distorted crypts. However, the SegNet model (Badrinarayanan et al., 2015) leads to under segmentation of closely located crypts shown in the fourth and fifth row. Also, a noisy patch shown in the last row can degrade the quality of the segmentation.

Table 3: F1 score of the region  $R_0$  (mucosa without crypt) and  $R_1$  (crypt) for validation and test images along with its histological indexes. The best performance is obtained for the image with architecture = 0, activity = 0, chronicity = 0.

Image	Architecture	Chronicity	Activity	$F_1(R_0)$	$F_1(R_1)$
Test image 1	1	1	0	0.69	0.60
Test image 2	1	1	0	0.67	0.69
Test image 3	1	0	0	0.49	0.61
Test image 4	0	0	0	0.76	0.71
Test image 5	1	0	0	0.26	0.32
Val image 6	2	2	1	0.37	0.68
Val image 7	1	0	0	0.74	0.74
Val image 8	1	1	0	0.46	0.65
Val image 9	1	0	0	0.56	0.70

Table 4: The first three rows show regularly sized and deformed crypts with a precise segmentation (without post-processing) using our trained SegNet model. The last three rows illustrate typical cases of under-segmentation, possibly due to closely located crypts or a noisy patch. The region  $R_0$  (mucosa without crypt) is shown in green,  $R_1$  (crypt) is shown in red,  $R_2$  (non-mucosa) is shown in blue and  $R_3$  (background) is shown in black.

	Multimodal image	Manual annotation	Machine learning	Deep learning
Regularly shaped crypts				
Elongated crypt				
Closely located crypts				
Noisy patch				

## 5 DISCUSSION

In this paper, we presented a semantic segmentation of non-linear multimodal images to automatize the predictive modelling of histological indexes for characterizing inflammatory bowel disease stages. We used a SegNet (Badrinarayanan et al., 2015) model for the segmentation of multimodal images into mucosa and crypt regions. Moreover, we compared the SegNet (Badrinarayanan et al., 2015) based semantic segmentation of multimodal images with a classical machine learning approach.

For the classical machine learning approach, texture features and linear classifier (PCA-LDA) was chosen due to simplicity. In order to make a fair comparison between the two approaches, same set of training images were used and the window size in the machine learning approach was set comparable to the receptive field of the SegNet model. Optimizing the window size for the machine learning approach did not affect the performance significantly, rather smaller window size increased the computation time. It was observed that the classical machine learning approach along with the hand-crafted features lack the ability to segment the tissue regions, due to a disturbing biological variance resulting from different grades of IBD. As these hand-crafted features are calculated using the intensity at pixel-level, it failed to retain the intrinsic shape information of the crypts. While manually calculated texture features were incapable of segmenting the crypt and mucosa regions, deep neural network like SegNet (Badrinarayanan et al., 2015) achieved reasonable to good result.

Our SegNet model was trained using categorical cross entropy loss function which considers every pixel as an independent sample and asserts equal learning for all pixels. This is a drawback for images with unbalanced classes. Therefore, we believe that weighted pixel wise cross entropy and dice loss function can segment the multimodal images effectively. The weighted pixel wise cross entropy loss in the U-net (Ronneberger et al., 2015) assisted the segmentation of closely located cells in biomedical images. Similarly for closely located crypts more advanced loss functions (Hashemi et al., 2018) can be implemented.

Deep learning approach can generalize the diversity in the underlying data and learn domain-specific representations, although it manifests certain drawbacks. Firstly, it is difficult to understand the contribution of the CARS, TPEF and SHG signal intensity for the segmentation of the mucosa and the crypts. Secondly, a deep learning approach requires large amount of good quality data which is difficult to

obtain particularly in a new technique like non-linear multimodal imaging. Thus, a data augmentation was needed.

For data augmentation, the multimodal images were randomly rotated to consider arbitrary orientations of the multimodal images. This helped to construct a rotation-invariant model. The patches outside the image grid were zero-padded and were filtered by the patch selection process before training the model. However, another possibility could be to mask these zero-padded regions in the loss calculation during model training. Other augmentation techniques like zooming, shearing and resizing of the images affected the spatial resolution and the crypt architecture in the multimodal image. Therefore these techniques were not applied.

In addition to data augmentation, a patch-based DCNN was used to increase the training data and also retain the crypt architecture. The patch size  $256 \times 256$  was optimized such that maximum tissue structure is retained. Smaller patch size failed to retain information between the crypts and generated more data making the training computationally expensive. The patch based DCNN worked efficiently, but due the combining of the patches to an image a “blocky effect” was generated. “Blocky effect” can also be generated due some other factors like the use of ‘same’ convolutions instead of ‘valid’ convolutions and odd number of feature maps before the pooling layer during training process.

To tackle this “blocky effect” simple post-processing methods were applied, which include morphological operations like remove blobs and region filling to eliminate false positive regions. These post-processing methods improved the segmentation results qualitatively. However, quantitative evaluation of these methods did not show significant changes in the F1 score and recall. Therefore, more complicated post-processing procedures like conditional random field (CRF) (Sutton and McCallum, 2012) are needed which can remove the false positives and improve results quantitatively. Nevertheless, these procedures increase the model complexity. Our post-processing methods led to an under segmentation of the crypts in some patches. This can be misleading in assessing the histological index as fusion of two regularly shaped crypts can be identified as one deformed crypt, leading to false prediction of IBD stage. Therefore, care must be taken while choosing appropriate post-processing procedures.

On the whole, a robust model can be constructed for segmenting the multimodal images with a large number of good quality images. Further, advanced loss function and post-processing procedures as men-

tioned above will need to be explored in future studies.

## 6 CONCLUSIONS

In summary, we achieved a quantitative evaluation of a semantic segmentation task of non-linear multimodal images to complement IBD diagnosis. An automatic segmentation of the crypt and mucosa region can reduce the manual diagnostic effort and can be used to predict histological indexes in real-time based on non-linear multimodal images. One limitation of our work was a small database with only a few exemplars of high histological index levels. Nevertheless, in future, a large dataset of annotated multimodal images to evaluate the model will be generated and this will improve the model presented here. In conclusion, non-linear multimodal imaging can assist the ‘gold-standard’ techniques and can be utilized under clinical conditions. Furthermore, incorporating a model for automatic segmentation of multimodal images into the multimodal microscope can provide a real-time histological index prediction and accelerate the start of a clinical therapy.

## ACKNOWLEDGEMENTS

Financial support of the EU, the ‘Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft’, the ‘Thüringer Aufbaubank’, the Federal Ministry of Education and Research, Germany (BMBF), the German Science Foundation (BO 4700/1-1, PO 563/30-1, STA 295/11-1), and Leibniz association via the ScienceCampus ‘InfectoOptics’ for the project ‘BLOODi’ are greatly acknowledged.

## REFERENCES

- Awan, R., Sirinukunwattana, K., Epstein, D., Jefferyes, S., Qidwai, U., Aftab, Z., Mujeeb, I., Snead, D., and Rajpoot, N. (2017). Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific Reports*, 7(1):16852.
- Babaie, M., Kalra, S., Sriram, A., Mitcheltree, C., Zhu, S., Khatami, A., Rahnamayan, S., and Tizhoosh, H. R. (2017). Classification and retrieval of digital pathology scans: A new dataset. *CoRR*, abs/1705.07522.
- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561.
- Chen, H., Qi, X., Yu, L., and Heng, P. (2016). DCAN: deep contour-aware networks for accurate gland segmentation. *CoRR*, abs/1604.02677.
- Chernavskaia, O., Heuke, S., Vieth, M., Friedrich, O., Schürmann, S., Atreya, R., Stallmach, A., Neurath, M. F., Waldner, M., Petersen, I., Schmitt, M., Bocklitz, T., and Popp, J. (2016). Beyond endoscopic assessment in inflammatory bowel disease: real-time histology of disease activity by non-linear multimodal imaging. *Scientific Reports*, 6:29239.
- Chollet, F. et al. (2015). Keras.
- Cicchi, R. and Pavone, F. S. (2014). Multimodal non-linear microscopy: A powerful label-free method for supporting standard diagnostics on biological tissues. *Journal of Innovative Optical Health Sciences*, 7(5):1330008.
- Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., and Tomaszewski, J. (2007). Automated grading of prostate cancer using architectural and textural image features. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1284–1287.
- Farjam, R., Soltanian-Zadeh, H., Jafari-Khouzani, K., and Zoroofi, R. A. (2007). An image analysis approach for automatic malignancy determination of prostate pathological images. *Clinical Cytometry*, 72B:227–240.
- Gunduz-Demir, C., Kandemir, M., Tosun, A. B., and Sokmensuer, C. (2010). Automatic segmentation of colon glands using object-graphs. *Medical Image Analysis*, 14:1–12.
- Guo, S., Pfeifenbring, S., Meyer, T., Ernst, G., Eggeling, F., Maio, V., Massi, D., Cicchi, R., Pavone, F. S., Popp, J., and Bocklitz, T. (2018). Multimodal image analysis in tissue diagnostics for skin melanoma. *Journal of Chemometrics*, 32:e2963.
- Hashemi, S. R., Salehi, S. S. M., Erdogmus, D., Warfield, S. K., and Gholipour, A. (2018). Asymmetric similarity loss function to balance precision and recall in highly unbalanced deep medical image segmentation. *CoRR*, abs/1803.11078.
- Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29.
- Jaremenko, C., Maier, A., Steidl, S., Hornegger, J., Oetter, N., Knipfer, C., Stelzle, F., and Neumann, H. (2015). Classification of confocal laser endomicroscopic images of the oral cavity to distinguish pathological from healthy tissue.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed ;today;].
- Kainz, P., Pfeiffer, M., and Urschler, M. (2017). Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. *PeerJ*, 5:e3874.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a). Imagenet classification with deep convolutional neu-

- ral networks. In *Advances in Neural Information Processing Systems*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12*, pages 1097–1105.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition.
- Legesse, F., Chernavskaia, O., Heuke, S., Bocklitz, T., Meyer, T., Popp, J., and Heintzmann, R. (2015). Seamless stitching of tile scan microscope images. *Journal of Microscopy*, 258(3):223–232.
- Li, C., Xu, C., Gui, C., and Fox, M. D. (2005). Level set evolution without re-initialization: a new variational formulation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 430–436 vol. 1.
- Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038.
- Naik, S., Doyle, S., Feldman, M., Tomaszewski, J., and Madabhushi, A. (2008). Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology.
- Norouzi, A., Rahim, M. S. M., Altameem, A., Saba, T., Rad, A. E., Rehman, A., and Uddin, M. (2014). Medical image segmentation methods, algorithms, and applications. *IETE Technical Review*, 31(3):199–213.
- Pathak, A. R., Pandey, M., and Rautaray, S. (2018). Application of deep learning for object detection. *Procedia Computer Science*, 132:1706 – 1717. International Conference on Computational Intelligence and Data Science.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Peng, Y., Jiang, Y., Eisengart, L., Healy, M., Straus, F., and Yang, X. (2011). Computer-aided identification of prostatic adenocarcinoma: Segmentation of glandular structures. *Journal of Pathology Informatics*, 2(1):33.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.
- Roth, H., Farag, A., Lu, L., Turkbey, E. B., and Summers, R. M. (2015). Deep convolutional networks for pancreas segmentation in CT imaging. *CoRR*, abs/1504.03967.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2014). Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575.
- Schürmann, S., Foersch, S., Atreya, R., Neumann, H., Friedrich, O., Neurath, M. F., and Waldner, M. J. (2013). Label-free imaging of inflammatory bowel disease using multiphoton microscopy. *Gastroenterology*, 145(3):514 – 516.
- Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.
- Sirinukunwattana, K., Snead, D. R. J., and Rajpoot, N. M. (2015). A stochastic polygons model for glandular structures in colon histology images. *IEEE Transactions on Medical Imaging*, 34(11):2366–2378.
- Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. *CoRR*, abs/1409.4842.
- Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.
- Travis E. O. (2006). A guide to numpy. USA: Trelgol Publishing.
- Vogler, N., Heuke, S., Bocklitz, T. W., Schmitt, M., and Popp, J. (2015). Multimodal imaging spectroscopy of tissue. *Annual Review of Analytical Chemistry*, 8:359–387.
- Waldner, M. J., Rath, T., Schürmann, S., Bojarski, C., and Atreya, R. (2017). Imaging of mucosal inflammation: Current technological developments, clinical implications, and future perspectives. *Frontiers in Immunology*, 8:1256.
- Wu, H.-S., Xu, R., Harpaz, N., Burstein, D., and Gil, J. (2005). Segmentation of intestinal gland images with iterative region growing. *Journal of Microscopy*, 220:190–204.