Supplemental Online Material and details to

Computer-aided Diagnosis using Deep Learning in the Evaluation of early Oesophageal Adenocarcinoma

Alanna Ebigbo[1,*], Robert Mendel[2,*], Andreas Probst[1], Johannes Manzeneder[1], Luis A. de Souza Jr.[2,3], João P. Papa[2,3], Christoph Palm[2,4] and Helmut Messmann[1]

[1] III. Medizinische Klinik, Klinikum Augsburg – Germany
[2] Regensburg Medical Image Computing (ReMIC), Ostbayerische Technische Hochschule Regensburg (OTH Regensburg) – Germany
[3] Department of Computing, São Paulo State University – Brazil
[4] Regensburg Center of Health Sciences and Technology (RCHST), OTH Regensburg – Germany

[*] Both authors contributed equally as first authors.

1. Methodological description

1.1 Patient inclusion

Patients and Images (Augsburg data)

Patients included in the study belonged to one of three categories:

1. Patients referred for standard surveillance of Barrettt's esophagus (BE)
2. Patients referred with a proven neoplasia in a random biopsy for further evaluation
3. Patients referred for endoscopic treatment of a confirmed early esophageal adenocarcinoma (EAC).

In some patients with long-segment BE, images were taken from different regions in the same patient. Images included 148 high definition (1350 $\times$ 1080 pixels) white light (WL) and narrow-band images (NBI) of 33 early EAC and 41 areas of non-neoplastic Barretts mucosa. Images were collected by three endoscopists (AE, HM, AP) experienced in the evaluation of BE. Only images of EAC with Paris 0-IIa or 0-IIb morphology were included in the study. All images were recorded using the near-focus / magnification function of a conventional video gastroscope (GIF-HQ190, Olympus Medical Systems, Tokyo, Japan). A transparent hood was attached to the tip of the scope to maintain the same distance to the mucosa during image acquisition.

Medical Image Computing and Computer Assisted-Intervention (MICCAI) data

The MICCAI data is an open access image data set provided by the Endoscopic Vision Challenge at the MICCAI conference 2015. It contains 100 high-definition (1600 $\times$ 1200 pixels) WL and pathologically validated endoscopic images from 39 patients; 17 patients with early EAC and 22 with non-neoplastic Barretts. In each image with EAC, the neoplastic area had been delineated by five international BE experts.

1.2 Patient exclusion

For the Augsburg data, exclusion criteria were:

1. Age <18 years
2. Patients under anticoagulant therapy

prohibiting esophageal biopsy or endoscopic treatment

3. ASA Grade IV

4. Images of protruding, pedunculated, depressed or excavated lesions (Paris 0-Ip/Is, 0-IIc and 0-III) were excluded in the Augsburg data set.

1.3 IRB / registration

The study was conceived as a prospective single-center trial. Ethics approval with the Reference Number 2017-11 was granted by the local institutional review board of the Klinikum Augsburg. The clinical part of the study including recruitment of patients, acquisition, and annotation of endoscopic images of the Augsburg data was done at the Department of Gastroenterology of the Klinikum Augsburg from November 2016 to November 2017. The CAD-DL evaluation of the endoscopic images was done at the Regensburg Medical Image Computing (ReMIC) lab at the Ostbayerische Technische Hochschule Regensburg (OTH Regensburg, Technical University of Applied Sciences).

1.4 Outcomes

- Comparison of the diagnostic performance of the CAD with deep learning (CAD-DL) system, trained and tested separately on two distinct image sets (Augsburg and MICCAI data).
- Comparison of the diagnostic performance of the CAD-DL system on corresponding NBI and WL images.
- Assessment of the performance of endoscopists on both data sets (MICCAI and Augsburg), compared with the diagnostic performance of the CAD-DL system.
- Computation of an automated segmentation of the tumor region based on the cancer probability provided by the deep learning system. This allows the comparison of automated and manual expert delineations.

1.5 Study approach

Endoscopic resection and histology (Augsburg data)

After endoscopic examination and imaging, EAC was resected by endoscopic submucosal dissection (ESD). For regions of normal Barretts mucosa without dysplasia, a biopsy was taken using standard biopsy forceps. For correct correlation between the image given to the CAD-DL system and the spot out of which the biopsy was taken, the examiner tried as much as possible to keep the endoscopic view in the same position between image acquisition and biopsy using the transparent hood.

The histology of the resection specimen or the biopsy served as the reference standard for the characterization of images. Based on the results of histology, the endoscopic images were divided into two categories:

- Early EAC (pT1)
- Non-neoplastic Barretts mucosa (BE)

Histological diagnosis was validated by a second pathologist.

Tumor segmentation / Delineation of cancer margins (Augsburg data)

After image acquisition, a region of interest was delineated using the open source image editing program GIMP. The delineation was performed by one of the three experienced endoscopists and routinely re-evaluated by another one of the three. The manual delineations show the tumor margins and suspicious regions in EAC and BE images, respectively. They were used for training of the deep learning classification system and additionally served as the reference standard for the tumor segmentation task subsequently given to the CAD-DL system.

Image evaluation by endoscopists

To establish a basis of comparison for the results of the CAD-DL system, 13 endoscopists blinded to the true diagnosis were asked to characterize images of both data sets. During image evaluation, the endoscopists provided a level of certainty by stating whether they felt confident or not-confident about their diagnosis. For purposes of uniformity between both data sets, only WL images were evaluated in this part of the study. For the MICCAI data, endoscopists were offered the first WL image of each region (39 images), and regarding the Augsburg data, endoscopists evaluated 74 WL images.

1.6 Device and technique

Computer-Aided Diagnosis / Deep Learning system

The proposed CAD-DL system is a special case of a deep convolutional neural network (CNN) with a residual net (ResNet) architecture [1]. For this study, our ResNet consisted of 100 layers, where the parameters are learned during a training phase and the diagnosis is estimated during a test phase. Both, training and testing were done completely independent for the two data sets.

To ensure a strict separation of training and testing data as well as to increase training and test corpus size at the same time, a leaving-one-patient-out cross-validation (LOPO-CV) approach was conducted. LOPO-CV means, that all information of one patient (left-out patient) were taken out of the training set. Then, the deep learning system was trained on the images of the remaining patients. Hereafter, the image of the left-out patient was used for classification. Thus, looping over all patients and taking P patients into account, each patient served $P-1$ times in a training set and once in a test set.

For training, small patches were generated from the endoscopic color images and augmented to simulate similar instances of the same class. Finally, the parameters were adjusted using the training data. For classification, firstly the class probabilities for each patch of the test image were estimated. Then, the class decision for the full image was compiled from the patch class probabilities. See Figure 1 (main text) for an overview of the system.

A definition of a particular number of images sufficient for training a CAD-DL system is virtually impossible. Although the patch-based approach in combination with LOPO-CV increased the corpus size for training considerably, a minimum number of $60-80$ images were assumed to be required. For our experiments, we used 7219 patches and 5359 patches for Augsburg and MICCAI data respectively, without taking augmentation into account.

1. **Training patch generation:** The full images were sampled randomly by extracting patches with a size of 224×224 pixels. The class of each patch was determined by the delineations of the human experts. For the MICCAI data with five delineations, the intersection area was chosen as the ground truth. The MICCAI data contained two kinds of areas: delineations of dysplastic regions; all other regions were defined as non-dysplastic. The Augsburg data showed three different kinds of areas: tumor margins, suspicious BE regions and background. Therefore, the MICCAI data were treated as a two-class problem, the Augsburg data as a three-class problem.

2. **Augmentation:** To enhance the generalization ability of the network, the patches were modified systematically. Rotation, translation, mirroring along the horizontal and vertical axis as well as contrast, brightness hue and saturation jittering were applied to each patch in a randomized fashion.

3. **Training the model:** Training a deep learning model means minimizing a given loss function, which expresses how close the prediction is to the ground truth. For our experiments, we employed the margin loss function [2]. For a reasonably proper initialization of the ResNet parameters, the transfer learning approach was chosen [1, 3].

4. **Patch classification:** For testing, the full image was sampled equidistantly into overlapping patches of size 224 × 224 pixel with an offset of 50 pixels. Then, each sample was propagated feed-forward through the ResNet. This resulted in a probability for each class for each patch.

5. **Image classification and segmentation:** For classification, a threshold t regarding these patch-specific class probabilities was considered. Only if the probability of class cancer exceeded t for at least one patch, the image was classified as cancer. For segmentation, t was applied to all patches resulting in binary images.

Hardware and Software

The deep learning experiments were applied on a Graphics processing unit (GPU) cluster with 8 Nvidia 1080Ti. The software was based the Pytorch framework, which is publicly available and open source.

1.7 Data analysis

The performance of the CAD-DL system was evaluated separately for the classification and the segmentation task. The classification results were shown in terms of sensitivity (SE), specificity (SP) and $F1$-measure as the harmonic mean of SE and SP:

$$SE = \frac{\text{\# true positive images}}{\text{\# positive images}} \tag{1}$$

$$SP = \frac{\text{\# true negative images}}{\text{\# negative images}} \tag{2}$$

$$F1 = \frac{2 \cdot SE \cdot SP}{SE + SP} \tag{3}$$

The segmentation results were evaluated using the Dice coefficient D [4]. D takes the overlap of an automated segmentation A and the binary mask resulting from the intersection of expert delineations E into account. Then, D is defined as:

$$D = \frac{2|E \cap A|}{|E| + |A|} \tag{4}$$

In this study, the images were cropped by at least 87 pixels (patch size / 2 - offset / 2) before D was computed, because only patches which were completely inside the image were processed.

Statistical analysis

An exact 2-sided McNemar test was performed to check the sensitivity and specificity results of endoscopists against the CAD-DL system as well as the CAD-DL results on WL against those on NBI. For p values below .05, the results were assumed to be statistically significantly different. The testing was performed with the help of the software *R*.

2. Details of results

2.1 Patient characteristics

For the Augsburg data, a total of 74 regions from 62 patients (2 females, 60 males) with a median age of 66 years underwent endoscopic imaging using near-focus WL as well as NBI. The MICCAI data set is described above.

2.2 Technical details

Performance of CAD-DL in images of early EAC

After training, our CAD-DL system was able to diagnose EAC in WL images for the best $F1$ value of 92% with a sensitivity of 97% and a specificity of 88%, when a threshold of $t = 0.90$ was used. Changing the

Augsburg data, WL

t	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
SE	1.00	1.00	1.00	0.97	0.97	**0.97**	0.85	0.74
SP	0.34	0.49	0.54	0.68	0.76	**0.88**	0.93	0.98
$F1$	0.72	0.76	0.78	0.83	0.86	**0.92**	0.88	0.79

Augsburg data, NBI

t	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
SE	1.00	1.00	1.00	1.00	0.97	0.97	**0.94**	0.74
SP	0.35	0.35	0.36	0.43	0.53	0.60	**0.80**	0.95
$F1$	0.72	0.72	0.73	0.75	0.77	0.80	**0.86**	0.82

MICCAI data

t	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
SE	0.92	0.92	**0.92**	0.90	0.81	0.79	0.69	0.50
SP	0.98	0.98	**1.00**	1.00	1.00	1.00	1.00	1.00
$F1$	0.95	0.95	**0.96**	0.95	0.90	0.88	0.81	0.67

Tab. 1: *Diagnostic performance of CAD-DL on the Augsburg and MICCAI images for different thresholds t.*

threshold resulted in varied values for sensitivity and specificity accordingly. For NBI images, the best $F1$ value with 86% was achieved for $t = 0.95$ resulting in a sensitivity and specificity of 94% and 80%, respectively, again with varying results depending on the threshold value used (Table 1). Sensitivity and specificity of CAD-related WL and NBI results were statistically equal.

In the MICCAI images, CAD-DL achieved a sensitivity and specificity of 92% and 100%, respectively, for the best $F1$ value of 96% at a lower threshold of 0.75 (Table 1). Therefore, our previous results [3] were enhanced further. The performance measures remained constant in the range of $t = 0.65$ to $t = 0.80$.

	Augsburg data		**MICCAI data**	
	μ	σ	μ	σ
SE	0.76	0.11	0.99	0.03
SP	0.80	0.12	0.78	0.10
$F1$	0.77	0.05	0.87	0.07

Tab. 2: *Diagnostic performance of standard endoscopists on WL images with mean μ and standard deviation σ.*

Performance of endoscopists in both data sets

The performance of the 13 endoscopists is shown in Table 2. The mean sensitivity and specificity value were 76% and 80%, respectively, for the Augsburg data. For the

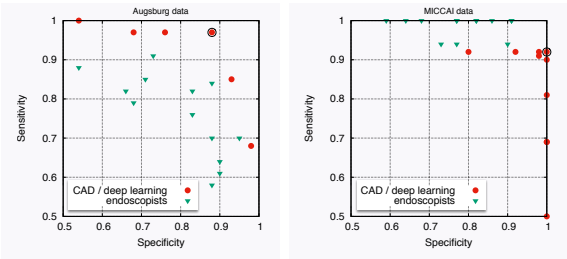

Fig. 1: *ROC curves – Augsburg WL (left) and MICCAI (right) – illustrating the diagnostic performance of endoscopists as compared with the CAD-DL system. Marked with a black circle is the overall best result of CAD-DL for the F1 value (cf. Table 1). These values were used for the McNemar significance tests.*

Augsburg data, WL

t	0.50	0.55	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
D_μ	**0.72**	0.71	0.67	0.65	0.62	0.60	0.55	0.48	0.41	0.18
D_σ	**0.18**	0.19	0.21	0.23	0.24	0.24	0.25	0.26	0.22	0.16

Augsburg data, NBI

t	0.50	0.55	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
D_μ	**0.72**	0.71	0.67	0.65	0.62	0.59	0.57	0.51	0.43	0.35
D_σ	**0.20**	0.20	0.23	0.24	0.25	0.26	0.26	0.28	0.30	0.25

MICCAI data

t	0.50	0.55	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
D_μ	**0.56**	0.56	0.51	0.48	0.46	0.44	0.45	0.42	0.40	0.38
D_σ	**0.18**	0.19	0.21	0.23	0.25	0.26	0.24	0.25	0.26	0.28

Tab. 3: *Dice coefficient as measure of overlap in tumor segmentation between CAD-DL and experienced endoscopists in terms of mean D_μ and variance D_σ for different thresholds t.*

MICCAI data sensitivity and specificity of 99% and 78%, respectively, were achieved. The ROC curves (Figure 1, Supplemental Material) illustrate the variability of sensitivity and specificity achieved by the endoscopists in comparison with the CAD-DL system which had more stable values across both data sets.

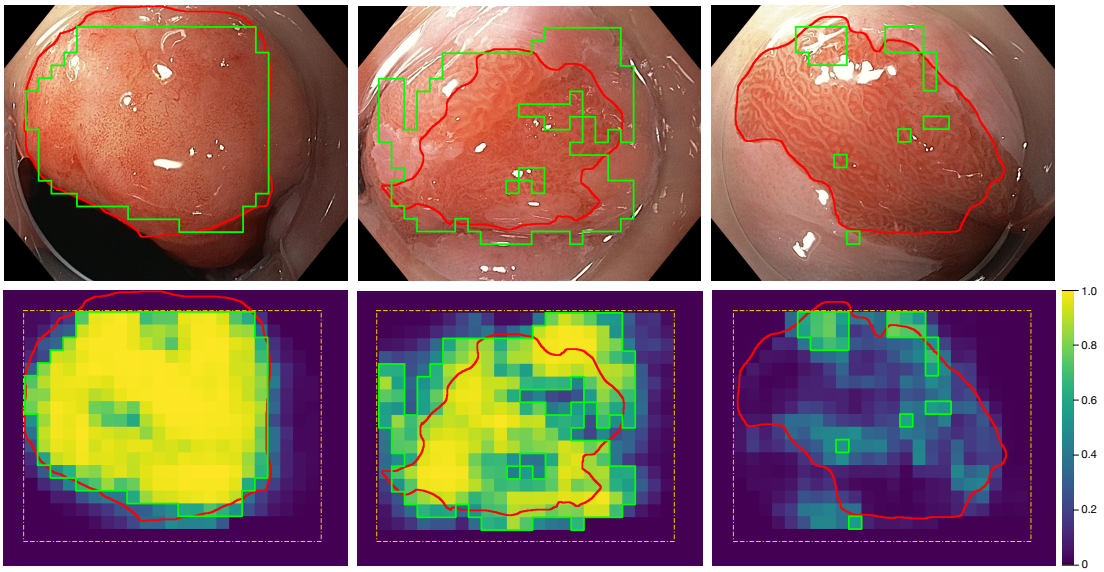

Fig. 2: *Examples of the tumor segmentation of Augsburg WL data are shown by a green contour overlaid on the original image (top) and the pseudo-colored patch-based probability maps (bottom). For comparison, the manual segmentation of an experienced endoscopist is drawn in red. The images represent the best result (left column, $D = 0.96$), one example of a mean result (middle column, $D = 0.72$), as well as the worst result (right column, $D = 0.15$). Note, that the CAD-DL segmentation is restricted to the area indicated by the orange dashed line.*

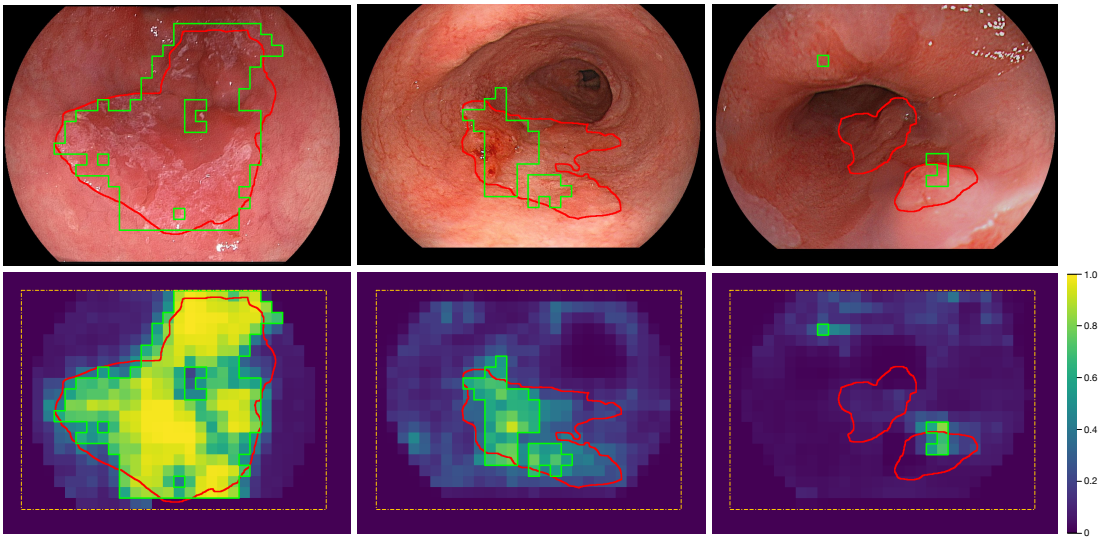

Fig. 3: *Examples of the tumor segmentation of MICCAI data. The images represent the best result (left column, $D = 0.87$), one example of a mean result (middle column, $D = 0.56$), as well as the worst result (right column, $D = 0.11$). The red contour is the interception of the manual delineations of five experienced endoscopists.*

Furthermore, the difference in difficulty diagnosing both data sets was shown by the proportion of confident statements given by the endoscopists. Taking only the cancer images into account, for the Augsburg data on average 69% of the statements were labeled as confident compared to 79% for the MICCAI data. 81% and 100% of these confident statements were correct for the Augsburg data and MICCAI data, respectively.

Performance comparison of endoscopists and CAD-DL

The McNemar test revealed statistically significant outperformance of the CAD-DL system for eleven of the 13 endoscopists for the Augsburg data either for sensitivity or specificity or for both. Only two endoscopists performed equal to the CAD-DL system. In seven cases the CAD-DL system showed a better sensitivity but equal specificity, in two cases equal sensitivity and better specificity and in one case both better sensitivity and specificity.

Since the performance of the endoscopists was much higher for the MICCAI data, a statistically significant difference was harder to achieve. Nevertheless, the CAD-DL system still outperformed four endoscopists in terms of sensitivity even for the MICCAI data. All other results were statistically equal.

Performance of CAD-DL in tumor segmentation and delineation

The measure of overlap (Dice coefficient D) between the segmentation of CAD-DL and that of experienced endoscopists can range between 0 (no overlap) and 1 (complete overlap). D was computed only for images correctly classified by CAD-DL as cancerous, and the results for the Augsburg and the MICCAI data are shown in Table 3. At a threshold of $t = 0.5$, a mean value of $D = 0.72$ was computed for the Augsburg data, equally for WL and NBI images (Figure 2, Supplemental Material). In the MICCAI data, D was 0.56 on average (Figure 3, Supplemental Material).

References

[1] K. He et al. Deep Residual Learning for Image Recognition. In: *Proc IEEE Conf Comp Vis Pat Recog*, 770–778, 2016.

[2] J. Gu et al. Recent advances in convolutional neural networks. *Pat Recog* 77, 354–377, 2018.

[3] R. Mendel et al. Barretts Esophagus Analysis Using Convolutional Neural Networks. In: *Bildverarbeitung für die Medizin*, 80–85, 2017.

[4] W. Jones and G. Furnas. Pictures of Relevance: A Geometric Analysis of Similarity Measures. *J Am Soc Inf Sc* 38(6), 420–442, 1987.