U-net Ensembles for Accurate Risk Stratification of Precursor Lesions of Esophageal Adenocarcinoma

Bryan Cardenas Guevara* & Vivian van Oijen*

*Graduate School of Informatics, University of Amsterdam
Incidence Rate of Esophageal Adenocarcinoma

- Sharp increase in the incidence rate

Am J Gastroenterol 2014; 109; 336-343
Barrett’s Esophagus

Gastro Esophageal Reflux disease Barrett’s Esophagus

At risk to develop Adenocarcinoma
Endoscopic surveillance with biopsies for risk stratification
Barrett’s Esophagus

Gastro Esophageal Reflux disease Barrett’s Esophagus

At risk to develop Adenocarcinoma
Endoscopic surveillance with biopsies for risk stratification

Pink discoloration
Barrett’s on Biopsy

- Normal preexisting squamous epithelium
- Squamo columnar junction
- Intestinal metaplasia (goblet cells)
- Gastric type mucosa
A proportion of patients with Barrett’s Esophagus develop Adenocarcinoma

Risk of progression of Non-Dysplastic Barrett’s Epithelium to cancer is low:
  ∘ ~0.3 - 0.6 % per year

Hvid-Jensen F. NEJM 2011, de Jonge P. Gut 2014; Dutch Barrett guideline
A proportion of patients with Barrett’s Esophagus develop Adenocarcinoma.

Risk of progression of Non-Dysplastic Barrett’s Epithelium to cancer is low:
- \( \sim 0.3 - 0.6 \% \) per year

- No Dysplasia (NDBE) → **Low interval of endoscopic surveillance**
- Low Grade Dysplasia → **High interval of endoscopic surveillance -- Endoscopic ablation therapy**
- High Grade Dysplasia → **Endoscopic ablation therapy -- surgical therapy**

CANCER

Hvid-Jensen F. NEJM 2011, de Jonge P. Gut 2014; Dutch Barrett guideline
A proportion of patients with Barrett’s Esophagus develop Adenocarcinoma

Risk of progression of Non-Dysplastic Barrett’s Epithelium to cancer is low:
- ~0.3 - 0.6 % per year

- No Dysplasia (NDBE): Low interval of endoscopic surveillance
- Low Grade Dysplasia: High interval of endoscopic surveillance -- Endoscopic ablation therapy
- High Grade Dysplasia: Endoscopic ablation therapy -- surgical therapy

Hvid-Jensen F. NEJM 2011, de Jonge P. Gut 2014; Dutch Barrett guideline
A proportion of patients with Barrett’s Esophagus develop Adenocarcinoma

Risk of progression of Non-Dysplastic Barrett’s Epithelium to cancer is low:
- ~0.3 - 0.6 % per year

- **No Dysplasia (NDBE)**: Low interval of endoscopic surveillance
- **Low Grade Dysplasia**: High interval of endoscopic surveillance -- Endoscopic ablation therapy
- **High Grade Dysplasia**: Endoscopic ablation therapy -- surgical therapy

**CANCER**

*Hvid-Jensen F. NEJM 2011, de Jonge P. Gut 2014; Dutch Barrett guideline*
A proportion of patients with Barrett’s Esophagus develop Adenocarcinoma

Risk of progression of Non-Dysplastic Barrett’s Epithelium to cancer is low:
  - \( \sim 0.3 - 0.6 \% \) per year

From Barrett’s Esophagus to Adenocarcinoma

- No Dysplasia (NDBE): Low interval of endoscopic surveillance
- Low Grade Dysplasia: High interval of endoscopic surveillance -- Endoscopic ablation therapy
- High Grade Dysplasia: Endoscopic ablation therapy -- surgical therapy

Hvid-Jensen F. NEJM 2011, de Jonge P. Gut 2014; Dutch Barrett guideline
A proportion of patients with Barrett’s Esophagus develop Adenocarcinoma

Risk of progression of Non-Dysplastic Barrett’s Epithelium to cancer is low:
- \( \sim 0.3 - 0.6\% \) per year

**No Dysplasia (NDBE)**
- Low interval of endoscopic surveillance

**Low Grade Dysplasia**
- High interval of endoscopic surveillance -- Endoscopic ablation therapy

**High Grade Dysplasia**
- Endoscopic ablation therapy -- surgical therapy

Hvid-Jensen F. NEJM 2011, de Jonge P. Gut 2014; Dutch Barrett guideline
Observer variability has led to need for expert review

All cases of Barrett’s dysplasia should be confirmed by a second **EXPERT** gastro-intestinal pathologist.
Reliable and Reproducible diagnosis of dysplasia in patients with Barrett’s Esophagus
Reliable and Reproducible diagnosis of dysplasia in patients with Barrett’s Esophagus

Deep Learning Methods with promising results in
  - Brain Tumor Detection
Reliable and Reproducible diagnosis of dysplasia in patients with Barrett’s Esophagus

Deep Learning Methods with promising results in
- Brain Tumor Detection
- Diabetic Retinopathy classification
Reliable and Reproducible diagnosis of dysplasia in patients with Barrett’s Esophagus

Deep Learning Methods with promising results in
- Brain Tumor Detection
- Diabetic Retinopathy classification
- Detecting breast cancer metastasis (Camelyon16/17)
Aim of Study

Reliable and Reproducible diagnosis of dysplasia in patients with Barrett’s Esophagus

Deep Learning Methods with promising results in
- Brain Tumor Detection
- Diabetic Retinopathy classification
- Detecting breast cancer metastasis (Camelyon16/17)
- Microscopic nuclei classification
Reliable and Reproducible diagnosis of dysplasia in patients with Barrett’s Esophagus

Deep Learning Methods with promising results in
- Brain Tumor Detection
- Diabetic Retinopathy classification
- Detecting breast cancer metastasis (Camelyon16/17)
- Microscopic nuclei classification

Lack of research in the segmentation of dysplasia - carcinoma sequence in Barrett’s Esophagus
1. Annotate slides in high detail for:
   - Squamous epithelium (Sq)
   - Non dysplastic epithelium (NDBE)
   - Low grade dysplasia (LGD)
   - High Grade dysplasia (HGD)
1. Annotate slides in high detail for:
   - Squamous epithelium (Sq)
   - Non dysplastic epithelium (NDBE)
   - Low grade dysplasia (LGD)
   - High Grade dysplasia (HGD)

2. Create the masks
The Work Flow

1. Annotate slides in high detail for:
   - Squamous epithelium (Sq)
   - Non dysplastic epithelium (NDBE)
   - Low grade dysplasia (LGD)
   - High Grade dysplasia (HGD)

2. Create the masks

3. Extract Patches (512 x 512)
The Work Flow

1. Annotate slides in high detail for:
   - Squamous epithelium (Sq)
   - Non dysplastic epithelium (NDBE)
   - Low grade dysplasia (LGD)
   - High Grade dysplasia (HGD)

2. Create the masks

3. Extract Patches (512 x 512)

4. Artificial Neural Network Predictions
114 H&E stained whole slide images

- At 20x magnification:
  - 2 pixels = 1 μ

- 94 different patients

- Four Different Labs
  - From the Netherlands

- ~3 Biopsies per slide
The Data: Whole Slide Image

114 H&E stained whole slide images

- At 20x magnification:
  - 2 pixels = 1 μ
- 94 different patients
- Four Different Labs
  - From the Netherlands
- ~3 Biopsies per slide

- High Grade Dysplasia (HGD)
- Low Grade Dysplasia (LGD)
- Background (B)
114 whole slide images
- 93 used for training
- 21 for testing
114 whole slide images
- 93 used for training
- 21 for testing
The Data: Partitioning

114 whole slide images
  ○ 93 used for training
  ○ 21 for testing

Slicing resulted in ~5000 patches
For training and ~2000 for testing
  ○ 512 x 512 x 3
  ○ One half contains Dysplastic regions
114 whole slide images
  - 93 used for training
  - 21 for testing

Slicing resulted in ~5000 patches
For training and ~2000 for testing
  - 512 x 512 x 3
  - One half contains Dysplastic regions
The Neural Network

U-Net Architecture
- Encoding done by DenseNet-161 and ResNet-50
U-Net Architecture

- Encoding done by DenseNet-161 and ResNet-50

Ensemble of 10 Imagenet pre-trained U-nets

- 5 networks based on DenseNet-161
- 5 networks based on ResNet-50
The Neural Network

U-Net Architecture

- Encoding done by DenseNet-161 and ResNet-50

Ensemble of 10 Imagenet pre-trained U-nets

- 5 networks based on DenseNet-161
- 5 networks based on ResNet-50
Main evaluation metric:

\[ \text{Dice-Score} = F_1\text{-Score} = \frac{2|A \cap B|}{|A| + |B|} \]
Main evaluation metric:

\[ \text{Dice-Score} = \frac{2 |A \cap B|}{|A| + |B|} = \text{F}_1\text{-Score} \]

Weighted average Dice score
Main evaluation metric:

\[ \text{Dice-Score} = F_1-\text{Score} = \frac{2 |A \cap B|}{|A| + |B|} \]

Weighted average Dice score

Low Grade Dysplasia vs Background

High Grade Dysplasia vs Background
Patch Feeding the Neural Network

Extracted Patch

Data Augmentation

- Rotations
- Flips
- Random Noise
- Random Crops
Patch Feeding the Neural Network

- Extracted Patch
- Data Augmentation:
  - Rotations
  - Flips
  - Random Noise
  - Random Crops
- Neural Networks
Patch Feeding the Neural Network

Extracted Patch

Data Augmentation
- Rotations
- Flips
- Random Noise
- Random Crops

Neural Networks

Predictions
Patch Feeding the Neural Network

Extracted Patch

Data Augmentation
- Rotations
- Flips
- Random Noise
- Random Crops

Neural Networks

Post-Processing with CRF

Predictions
Patch Feeding the Neural Network

Extracted Patch

Data Augmentation
- Rotations
- Flips
- Random Noise
- Random Crops

Neural Networks

Post-Processing with CRF

Review by Pathologists

Predictions
Results: High Grade Dysplasia

Weighted Dice-Score of 0.83 on the held out set

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Pixel-wise accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
<td>-</td>
</tr>
<tr>
<td>High Grade Dysplasia</td>
<td>0.80</td>
<td>0.76</td>
<td>0.85</td>
<td>-</td>
</tr>
<tr>
<td>Weighted Score</td>
<td>0.83</td>
<td>0.8</td>
<td>0.87</td>
<td>0.94</td>
</tr>
</tbody>
</table>

WSI  

Mask
Results: High Grade Dysplasia

Weighted Dice-Score of **0.83** on the held out set

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Pixel-wise accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
<td>-</td>
</tr>
<tr>
<td>High Grade Dysplasia</td>
<td>0.80</td>
<td>0.76</td>
<td>0.85</td>
<td>-</td>
</tr>
<tr>
<td>Weighted Score</td>
<td>0.83</td>
<td>0.8</td>
<td>0.87</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Results: Low Grade Dysplasia

Weighted Dice-Score of 0.74 on the held out set

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Pixel-wise accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
<td>-</td>
</tr>
<tr>
<td>Low Grade Dysplasia</td>
<td>0.67</td>
<td>0.64</td>
<td>0.70</td>
<td>-</td>
</tr>
<tr>
<td>Weighted Score</td>
<td>0.74</td>
<td>0.72</td>
<td>0.76</td>
<td>0.86</td>
</tr>
</tbody>
</table>

WSI

Mask
Results: Low Grade Dysplasia

Weighted Dice-Score of 0.74 on the held out set

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Pixel-wise accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
<td>-</td>
</tr>
<tr>
<td>Low Grade Dysplasia</td>
<td>0.67</td>
<td>0.64</td>
<td>0.70</td>
<td>-</td>
</tr>
<tr>
<td>Weighted Score</td>
<td>0.74</td>
<td>0.72</td>
<td>0.76</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Results: Low Grade Dysplasia

Weighted Dice-Score of 0.74 on the held out set

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Pixel-wise accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
<td>-</td>
</tr>
<tr>
<td>Low Grade Dysplasia</td>
<td>0.67</td>
<td>0.64</td>
<td>0.70</td>
<td>-</td>
</tr>
<tr>
<td>Weighted Score</td>
<td>0.74</td>
<td>0.72</td>
<td>0.76</td>
<td>0.86</td>
</tr>
</tbody>
</table>

WSI

Mask

Prediction

WSI

Weighted Dice-Score: 0.94
Variability in the Predictions

WSI

Mask for High Grade Dysplasia
Variability in the Predictions

WSI

Mask for High Grade Dysplasia

Model Prediction for HGD
Variability in the Predictions

WSI

Model Prediction for HGD

Mask for High Grade Dysplasia

Mask for Low Grade Dysplasia
Variability in the Predictions

WSI

Mask for HGD and LGD
Variability in the Predictions

WSI

Mask for HGD and LGD

Model Prediction for LGD
Good results with the models trained on HGD

Both models make a lot of false positives

The models have more uncertainty regarding LGD
Good results with the models trained on HGD

Both models make a lot of false positives

The models have more uncertainty regarding LGD

Visually Useful Predictions
Annotations could be coarse
Annotations could be coarse
Improvements

Annotations could be coarse

Mask

Prediction
Annotations could be coarse

Patch

Mask
Annotations could be coarse

Patch

Mask
Annotations could be coarse

Patch

Mask

Prediction
Improvements

Annotations could be coarse

Patch

Mask

Prediction

More data!

Multiple annotators to isolate regions of interest: Inject Expert Knowledge
References

- Shidan Wang et al, Pathology Image Analysis Using Segmentation Deep Learning Algorithms, The American Journal of Pathology, V. 189, 9, 2019

- G Litjens et al: A survey on deep learning in medical image analysis, Arxiv 2017

- Olga Russakovsky et al., ImageNet Large Scale Visual Recognition Challenge, CoRR, 2014

- Gao Huang et al., Densely Connected Convolutional Networks, CoRR, 2016

- Hao Dong et al, automatic Brain Tumor Detection and Segmentation Using U-net Based Fully Convolutional Networks, CoRR 2018

- Kaiming He et al., Deep Residual Learning for Image Recognition, CoRR, 2015


- Yun Liu et al, Artificial intelligence-Based Breast cancer Nodal metastasis Detection, APLM, 2018

References

- Md. Z. Alom et al., Microscopic Nuclei Classification, Segmentation and Detection with improved Deep Convolutional Neural Network Approaches, 2018

- Qikui Zhu et al., Deeply-Supervised CNN for Prostate Segmentation, CoRR, 2018

- Andreas Maier et al., A gentle introduction to deep learning in medical image processing, zemedi, V. 29, I. 2 86-101, 2019

- Shidan Wang et al., Pathology Image Analysis Using Segmentation Deep Learning Algorithms, AJPath, V.189, I.9 1686-1698, 2019

- Luis A. de Souza Jr. et al., A survey on Barrett’s Esophagus analysis using machine learning, CBM, 96, 203-213, 2018

- Olaf Ronneberger et al., U-Net Convolutional Networks for Biomedical Image Segmentation, CoRR, 2015

- Tsung Yi Lin et al., Focal Loss for Dense Object Detection, CoRR, 2017
Acknowledgments

Amsterdam Machine Learning Lab, University of Amsterdam
Bryan Cardenas Guevara
Vivian van Oijen
Daniel Worrall

Dept. of Biomedical Engineering and Physics
Amsterdam University Medical Center
Daniel de Bruin
Ilaria Jansen
Marit Lucas
Henk Marquering

Dept. of Pathology,
Amsterdam University Medical Center
Kiki de Laat
Dilara Savci-Heijink
Myrtle van der Wel
Onno de Boer
Sybren Meijer
Variability in the Predictions

WSI

Mask for HGD and LGD

Prediction
Variability in the Predictions

WSI

Probabilities for LGD

Mask for HGD and LGD

Thresholded Prediction for LGD
Results: High Grade Dysplasia

Weighted Dice-Score of 0.83 on the held out set

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Pixel-wise accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
<td>-</td>
</tr>
<tr>
<td>High Grade Dysplasia</td>
<td>0.80</td>
<td>0.76</td>
<td>0.85</td>
<td>-</td>
</tr>
<tr>
<td>Weighted Score</td>
<td>0.83</td>
<td>0.8</td>
<td>0.87</td>
<td>0.94</td>
</tr>
</tbody>
</table>

WSI
Weighted Dice-score: 0.754
Variability in the Predictions

WSI

Model Prediction

Mask for High Grade Dysplasia

Mask for Low Grade Dysplasia
Metasplasia - Dysplasia to Adenocarcinoma Sequence

no dysplasia  Low Grade Dysplasia  High Grade Dysplasia  Cancer

Hvid-Jensen F. NEJM 2011, de Jonge P. Gut 2014; Dutch Barrett guideline
Reliable and Reproducible diagnosis of dysplasia in patients with Barrett’s Esophagus

Deep Learning Methods with promising results in

- Brain Tumor Detection
  
  Hao Dong et al, automatic Brain Tumor Detection and Segmentation Using U-net Based Fully Convolutional Networks, CoRR 2018

- Diabetic Retinopathy classification
  

- Detecting breast cancer metastasis (Camelyon16/17)
  
  Yun Liu et al, Artificial intelligence-Based Brast cancer Nodal metastasis Detection, APLM, 2018
  

- Microscopic nuclei classification
  
  Md. Z. Alom et al., Microscopic Nuclei Classification, Segmentation and Detection with improved Deep Convolutional Neural Network Approaches, 2018

Lack of research in the segmentation of dysplasia - carcinoma sequence in Barrett’s Esophagus
The Neural Network

U-Net Architecture

- Encoding done by DenseNet-161 and ResNet-50
  
  Gao Huang et al., Densely Connected Convolutional Networks, CoRR, 2016
  
  Kaiming He et al., Deep Residual Learning for Image Recognition, CoRR, 2015

Ensemble of 10 Imagenet pre-trained U-nets

- 5 networks based on DenseNet-161
- 5 networks based on ResNet-50

Cross Entropy Loss and Focal Loss

Qikui Zhu et al., Deeply-Supervised CNN for Prostate Segmentation, CoRR, 2018

Olaf Ronneberger et al., U-Net Convolutional Networks for Biomedical Image Segmentation, CoRR, 2015

Annealed Learning rate between 2e-5 and 8e-5

Tsung Yi Lin et al., Focal Loss for Dense Object Detection, CoRR, 2017
Results: Full Dysplasia

Weighted Dice-Score of 0.77 on the held out set

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>Pixel-wise accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>0.95</td>
<td>0.99</td>
<td>0.91</td>
<td>-</td>
</tr>
<tr>
<td>Dysplasia</td>
<td>0.69</td>
<td>0.56</td>
<td>0.90</td>
<td>-</td>
</tr>
<tr>
<td>Weighted Score</td>
<td>0.77</td>
<td>0.69</td>
<td>0.90</td>
<td>0.91</td>
</tr>
</tbody>
</table>