Segmentation of oesophageal cancer lymph nodes within large H&E datasets with explainable AI

Manon Beuque[1], Avishek Chatterjee[1], Henry C. Woodruff[1,2] , Ruth E. Langley[3], William Allum[4] , Matthew G. Nankivell[3], David Cunningham[5], Philippe Lambin[1,2] , Heike I. Grabsch[6,7]

1 The D-Lab, Department of Precision Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands
2 Department of Radiology and Nuclear Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands
3 MRC Clinical Trials Unit at University College London, London, UK
4 Department of Surgery, Royal Marsden Hospital, London, UK
5 Department of Medicine, The Royal Marsden NHS Trust, London and Sutton, UK
6 Department of Pathology, GROW Research Institute, Maastricht University Medical Center+, Maastricht, The Netherlands
7 Pathology & Data Analytics, Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds, UK

Background

The analysis of resected Haematoxylin-Eosin (H&E) stained lymph nodes (LNs) is currently limited to establishing if they contain metastatic tumours. Automatic analysis of the microarchitecture is currently difficult partially due to the lack of annotated LNs for machine learning. We had access to clinical trial data (OE02) to tackle this issue.

Purpose

We hypothesize that we can design a machine learning workflow that is able to:

(1) find digital H&E slides containing LNs by using shape and statistical features extracted from deep learning generated probability maps
(2) create a scoring system and an "uncertain" class comprising less than 20% of the dataset to help make the results explainable, and
(3) segment all LNs in those images.

Materials

We used **759 H&E slides** containing LNs **delineated** by an expert pathologist, and **1018 slides without LNs**, from oesophageal cancer patients recruited into the OE02 trial. The images were extracted from files in .svs format at a maximum resolution of 2048 pixels, preserving the aspect ratio of the original image and then followed the pre-processing steps described **figure 1**.

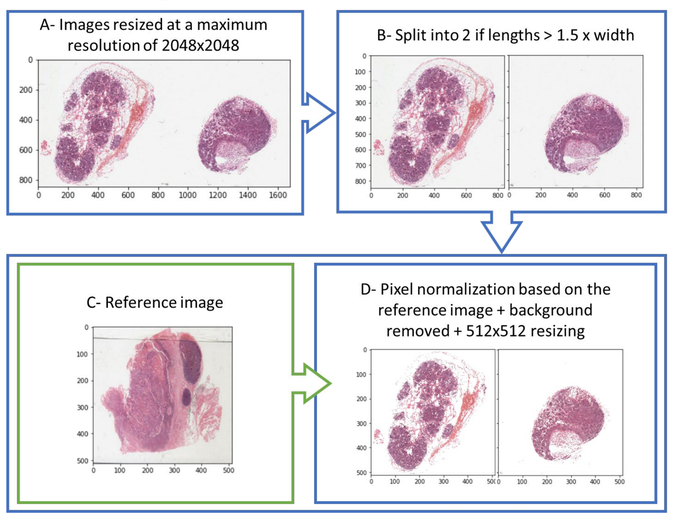

Figure 1. Pre-processing workflow A: Resizing H&E at a lower resolution. B: Image is split into two if it is rectangular. C: A reference image for color normalization is created. D: An image is chosen as reference to normalize image colour, the background is removed and the images are resized to 512x512 pixels.

Methods

The dataset was randomly divided into training (80%) and test (20%) sets. A **UNet architecture** was used to generate pixel-wise predictions for LN presence, from which handcrafted features were extracted. These features trained an optimized **xgboost model** which predicted if the region truly contained a LN. The comparative model predictions are described **figure 2**.

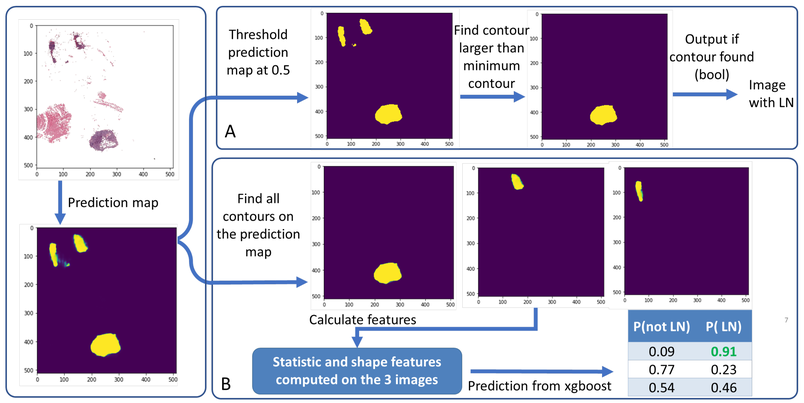

Figure 2. The two workflows for predicting whether an image contains a lymph node. A: state-of-the-art thresholding method, B: our novel prediction score method.

Results

(1) The accuracy of automatically detecting images with LNs was **0.95 with our scoring method**, compared to an accuracy of **0.89 obtained using the current state-of-the-art method** (**figure 3**).

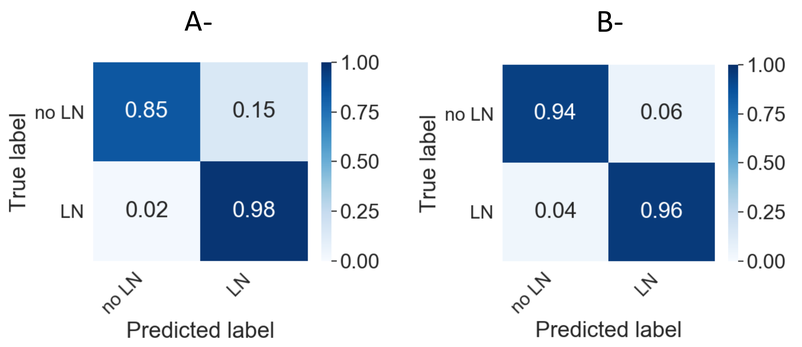

Figure 3. Results of the classification methods on the test dataset A: Confusion matrix based on the state-of-the-art method B: Confusion matrix based on the scores

(2) The **"uncertain" class comprised 13% of the entire dataset** while 2% of the images were misclassified (see **table 1**).

	Predicted images with LN (%)	Uncertain category (%)	Predicted images without LN (%)	Total per row (%)
Images with LN	140 (74)	46 (24)	4 (2)	190 (100)
Images without LN	6 (2)	15 (6)	241 (92)	262 (100)
Total	146 (32)	**61 (13)**	245 (54)	452 (100)

Table 1. Predictions split into three categories according to the level of certainty evaluated on the test dataset. Numbers in parentheses denote the fractional amount (and thus add up to 1 across the three categories).

(3) For LN segmentation, our model achieved an overall **Dice score** of **0.78** (image-level) and **0.71** (LN-level). Example of delineations are shown **figure 4**.

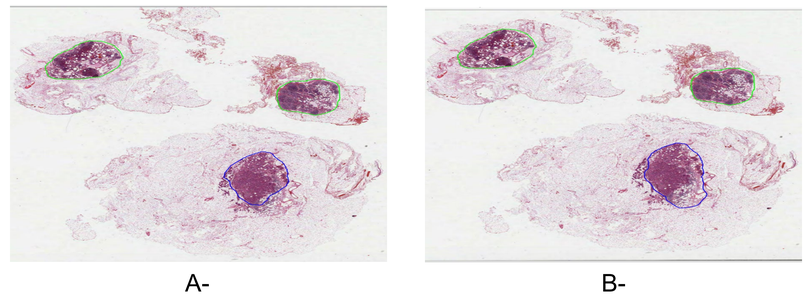

Figure 4. Automated contours: A- automatic contour correctly done in green, to correct in blue B- blue contour changed by the pathologist

Conclusion

The 1st part of our workflow can be used in **a routine diagnostic** setting; the 2nd part **limits misclassification** and **facilitates revision**; the 3rd part would **allow large-scale investigations** of LNs, identifying new **clinically relevant biomarkers** that might lead to personalized treatment. External validation on other LN H&E datasets from various cancer types is necessary before clinical implementation.

Email address: m.beuque@maastrichtuniversity.nl