Journal of Pathology Informatics 14 (2023) 100192

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Automated detection and delineation of lymph nodes in haematoxylin & eosin stained digitised slides

Manon Beuque [a], Derek R. Magee [b,c], Avishek Chatterjee [a], Henry C. Woodruff [a,d], Ruth E. Langley [e], William Allum [f], Matthew G. Nankivell [e], David Cunningham [g], Philippe Lambin [a,d], Heike I. Grabsch [h,i,*]

[a] Department of Precision Medicine, GROW School for Oncology and Reproduction, Maastricht University, Universiteitssingel 40, 6229 ER Maastricht, the Netherlands
[b] School of Computing, University of Leeds, LS2 9JT Leeds, United Kingdom
[c] HeteroGenius Limited, Leeds, United Kingdom
[d] Department of Radiology and Nuclear Medicine, GROW School for Oncology and Reproduction, Maastricht University Medical Center+, P. Debyelaan, 25 6229 HX Maastricht, The Netherlands
[e] MRC Clinical Trials Unit at University College London, 90 High Holborn, WC1V 6LJ London, United Kingdom
[f] Department of Surgery, Royal Marsden Hospital, The Royal Marsden Fulham Road, SW3 6JJ London, United Kingdom
[g] Department of Medicine, The Royal Marsden NHS Trust, The Royal Marsden Fulham Road, SW3 6JJ London, United Kingdom
[h] Department of Pathology, GROW School for Oncology and Reproduction, Maastricht University Medical Center+, P. Debyelaan, 25 6229 HX Maastricht, The Netherlands
[i] Pathology & Data Analytics, Leeds Institute of Medical Research at St. James's, University of Leeds, LS2 9JT Leeds, United Kingdom

ARTICLE INFO

Keywords:
Oesophageal cancer
Deep learning
Autodelineation
Explainability
Digital pathology
Lymph nodes

ABSTRACT

Treatment of patients with oesophageal and gastric cancer (OeGC) is guided by disease stage, patient performance status and preferences. Lymph node (LN) status is one of the strongest prognostic factors for OeGC patients. However, survival varies between patients with the same disease stage and LN status. We recently showed that LN size from patients with OeGC might also have prognostic value, thus making delineations of LNs essential for size estimation and the extraction of other imaging biomarkers.

We hypothesized that a machine learning workflow is able to: (1) find digital H&E stained slides containing LNs, (2) create a scoring system providing degrees of certainty for the results, and (3) delineate LNs in those images.

To train and validate the pipeline, we used 1695 H&E slides from the OE02 trial. The dataset was divided into training (80%) and validation (20%). The model was tested on an external dataset of 826 H&E slides from the OE05 trial. U-Net architecture was used to generate prediction maps from which predefined features were extracted. These features were subsequently used to train an XGBoost model to determine if a region truly contained a LN. With our innovative method, the balanced accuracies of the LN detection were 0.93 on the validation dataset (0.83 on the test dataset) compared to 0.81 (0.81) on the validation (test) datasets when using the standard method of thresholding U-Net predictions to arrive at a binary mask. Our method allowed for the creation of an "uncertain" category, and partly limited false-positive predictions on the external dataset. The mean Dice score was 0.73 (0.60) per-image and 0.66 (0.48) per-LN for the validation (test) datasets.

Our pipeline detects images with LNs more accurately than conventional methods, and high-throughput delineation of LNs can facilitate future LN content analyses of large datasets.

Introduction

Oesophageal and gastric cancers (OeGC) were diagnosed more than 1.5 million times worldwide in 2020 and represented 13.2% of all cancer deaths.[1] The treatment of OeGC patients depends on the disease stage, and patient performance status and preferences.[2] For Western patients diagnosed with locally advanced resectable disease, the standard of care is neoadjuvant chemo(radio)therapy followed by surgery for oesophageal cancer and perioperative chemotherapy for gastric cancer according to the ESMO guideline.[3]

The overall survival of Western OeGC patients is poor with a 3 year survival rate between 22.3% and 33.8% for gastric cancer and between 19.2% and 27.0% for oesophageal cancer.[4]

Lymph node (LN) status (presence or absence of metastasis in regional LNs) is currently the strongest prognostic factors for OeGC patients irrespective of treatment modality, grade of primary tumour regression,

Abbreviations: AUC, area under the curve; DL, deep learning; H&E, haematoxylin and eosin; LN, lymph node; OeGC, Oesophageal and gastric cancers; ROC, receiver operating characteristic.

* Corresponding author at: Department of Pathology, GROW School for Oncology and Reproduction, P. Debyelaan 25, 6229 HX Maastricht, The Netherlands.
E-mail address: h.grabsch@maastrichtuniversity.nl (H.I. Grabsch).

http://dx.doi.org/10.1016/j.jpi.2023.100192
Received 13 November 2022; Received in revised form 16 January 2023; Accepted 17 January 2023
Available online 25 January 2023

or regression in LN.[5,6] Our recent pilot study of digital haematoxylin and eosin (H&E) stained slides containing resection specimens from patients with oesophageal cancer from the OE02 trial[7] suggested that not only LN status but also the size of LNs might have prognostic value.[8] Validation of these pilot study findings is needed in at least 1 independent large study assessing thousands of LNs before pathological LN size can be considered as a useful biomarker for routine use in OeGC patient management. This and possibly other imaging biomarkers could be useful to identify patients who will benefit most from (potentially) toxic adjuvant treatment.

However, manual review of digital H&E-stained slides to identify and delineate all LNs as previously performed in the pilot study is not feasible within a reasonable time frame in large datasets. Recent phase III trials in OeGC patients typically amount to 20 000 slides and 10 000 LNs per trial, as on average 30 slides are made per resection specimen and more than 15 LNs per patient are obtained. Thus, a toolbox for the automatic identification of image files containing LNs and their automatic delineation would be very desirable for a large-scale validation of our LN size findings and as a prerequisite for further characterisation of the LN architecture by quantitative image analysis. To the best of our knowledge, there are currently no fully automated solutions available for such a task.

We hypothesized that a computational pipeline using a deep learning (DL) model combined with imaging features extracted from the generated prediction map can: (1) identify which H&E-stained digitised slides from oesophagogastric cancer resection specimens contain LNs and (2) automatically delineate the LNs with higher accuracy than current stand-alone DL solutions.

The aim of the study was to develop, validate, and externally test a DL-based workflow to enable large-scale high throughput studies in digital H&E-stained LN tissue sections from resection specimens of oesophagogastric cancer patients.

Materials and methods

Haematoxylin & eosin-stained digitised tissue section collection

H&E-stained slides were collected retrospectively from resection specimens from OeGC patients recruited into the phase III randomised controlled trial, UK MRC OE02.[7] Those samples were collected from 42 European centres. Whole slides were scanned using an Aperio XT Scanner. A total of 1695 scanned H&E slides from 493 resection specimens (on average 3,4 images per specimen) were manually reviewed and classified as containing one or more LNs (N = 756 images) or no LN (N = 939). All LNs were manually delineated by an expert pathologist using the Aperio ImageScope software (ground-truth delineations) and delineations were saved in an Aperio ImageScope XML annotation file format.

The image dataset was randomly split per patient, with 394 patients (~80%) in the training dataset and 99 (~20%) patients in the validation dataset. For the external dataset, 826 H&E slides were extracted from 33 resected specimens (on average 25 images per specimen) from the UK MRC OE05 trial.[9] The OE05 dataset had 348 images with delineated LNs and 478 images identified without LNs. The study was approved by the South East Research Ethics Committee, London, UK, REC reference: 07/H1102/111.

Pre-processing of digital images for deep learning

Common pre-processing strategies for H&E-stained images as described by Li et al.[10] were applied to the original images in our database to harmonise the dataset and remove noise: Python 3.7 was used and all packages/libraries used in this study are listed in supplementary material Table 1. As the resolution of slides scanned at 40× magnification can be up to 200 000 × 200 000 pixels, scanned images were extracted from the Aperio ScanScope files at a maximum size of 2048 × 2048 pixels, preserving the aspect ratio of the original image. To extract the image at a maximum resolution of 2048 × 2048 pixels, different downsample levels were tested until reaching the maximum resolution, at which point the downsampled image was extracted. Finally, the extracted images were converted into jpeg image file format to facilitate the use of standard python packages for pre-processing. One scanned image from the dataset was randomly selected to be the reference image for Macenko's colour normalisation strategy, which consists of colour deconvolution later matched to the colour characteristics extracted from the reference image.[11] As the DL model required square images as input, scanned images with rectangular shapes (i.e., length > than 1.5 times the width) were split into 2 squares to avoid overstretching or compressing of the image. We also applied the Otsu thresholding method,[12] a histogram-based filter able to generate a binary mask that separates the foreground (tissue) from the background (empty space) setting the background values to 255 to maintain a white background.

Subsequently, the scanned and cropped images were resized to 512x512 pixels by downsampling with bicubic interpolation (see Fig. 1) to be suitable as input for U-Net.

We named "pre-processed images" the resulting images. We also derived binary masks from the coordinates of the delineations saved in XML files indicating the pixels belonging to LN tissue. The number of samples used during the study before and after pre-processing can be seen in Table 1.

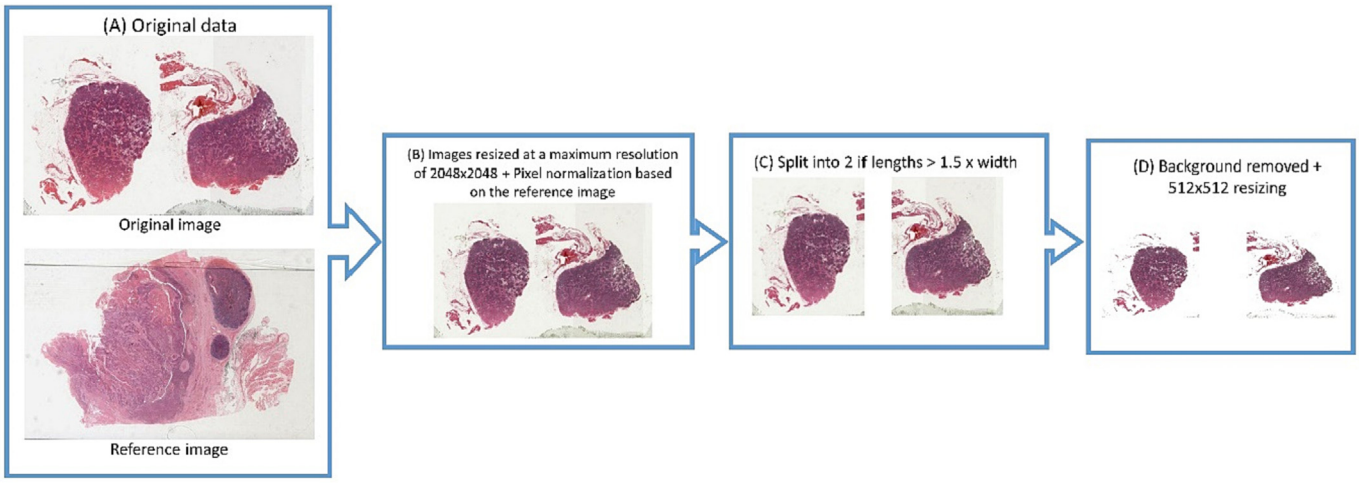

Fig. 1. Pre-processing workflow: colour normalisation, resizing, splitting, and removal of background. (A) Digitised glass slide with 2 lymph nodes; bottom: randomly chosen reference image. (B) Image after colour normalisation per pixel and downsizing to 2048 × 2048. (C) Image split into 2 sub-images in cases where the original image was rectangular. (D) Removal of background and resizing to 512 × 512 pixels.

Table 1
Description of the 3 different datasets: number of patients with scanned H&E slides, number of scanned H&E slides, and number of images after pre-processing of the scanned H&E slide.

Data type/subset	Training dataset from OE02	Validation dataset from OE02	Test dataset from OE05
Number of patients	394	99	33
Number of images	1340	355	826
Number of images after pre-processing	1516	481	1251

Deep learning model for automatic detection and delineation of lymph nodes

The DL model chosen to detect and delineate LNs was a U-Net[13] using ResNet-50 as backbone due to its proven good performance for histopathology whole slide image delineation.[14] The loss function used was a combination of Dice loss weighted at 0.3 and binary cross entropy loss weighted at 0.7, which empirically gave the best Dice score on the validation dataset. We used Adam (adaptive moment estimation), an algorithm which optimises the model with a learning rate of 10^{-4}.[15] The model was trained using 4 GPUs (NVIDIA GeForce RTX 2080Ti) until overfitting was observed based on the surveillance of the mean Dice coefficient in the training and validation datasets calculated after each epoch, i.e., when the epoch just before the mean Dice continued increasing for the training dataset but stagnated or reduced for the validation dataset.

Identification of the images containing lymph nodes

The output of the DL model per pre-processed image was a probability map showing the predicted likelihood of a particular pixel being part of a LN. In order to convert the per-pixel prediction value into a binary classifier for the scanned image, we compared the results of 2 methods based on the probability maps. The first method was the current standard method which uses a simple threshold of the prediction map to obtain a binary mask as described by Ronneberger et al,[13] termed "conventional method" in this article. The per-pixel predictions were threshold at 0.5 probability, where every prediction higher than 0.5 was considered part of a LN, creating a binary mask. To remove potential artefacts, small areas (minimum area set at 5% of the smallest LN area found in the training set) were considered as potential outliers and excluded from further analyses. Any scanned image containing a region of interest greater than that size was labelled as potentially containing LN.

The second method used *a priori* knowledge of the LN shape (usually similar to a kidney bean) to analyse the predicted LN delineation and select the most likely correctly segmented ones. This selection allowed us to obtain a prediction score not just per pixel but per LN and quantify the results of our DL model. The following features were extracted from the prediction map of each candidate LN: descriptive statistics (geometric and harmonic means, standard deviation of prediction values, entropy, skewness, and kurtosis), and shape features (pixel count, number of delineations predicted, roundness, roundness disproportion, area, perimeter, centroid, orientation, major axis length, minor axis length, diameter, extent, solidity, eccentricity, elongation, perimeter/surface ratio). The features were normalised using z-score normalisation based on the mean and standard deviation derived from the training dataset and the correlation between features were tested using the Spearman rank correlation coefficient[16] on the training dataset. To remove redundant information, if a correlation coefficient was above 0.85 between 2 features, the feature with the highest correlation coefficient across the correlation table was deselected from the remaining feature set. The normalisation and the feature selection based on the training dataset was later applied to the validation and test datasets. Finally, recursive feature elimination with 10 cross-validation (RFECV) using default parameters was performed on the features extracted from the training set, optimising the area under the curve (AUC) of the receiver operating characteristic (ROC) score for an extreme gradient boosting (XGBoost) classifier. At every iteration of the RFECV model, the least predictive feature was removed from the dataset until only 1 remained. We visualised the RFECV curve (corresponding to the AUC against the number of features) and selected the number of features corresponding to the turning point of the curve, i.e., when no further increase in the AUC score was observed.

The selected features were used as input for an XGBoost classifier which was trained to give a prediction score between 1 and 0 whether a candidate delineation contained a LN or not. To fine-tune the parameters of this classifier, a grid-search with 10-fold cross-validation was performed on the training dataset. The parameters tested were maximum depth, minimum child weight, number of estimators, gamma, and the learning rate. The set parameters were the scoring system using the ROC AUC, the objective being binary logistic, and column sample by tree at 0.8. To determine whether a particular pre-processed image contained a LN or not, each potential LN within the pre-processed image was attributed a prediction score and the highest score was chosen to classify the pre-processed image. The workflow of both automatic classification strategies per pre-processed image (with LNs/without LNs) is shown Fig. 2.

Reusing our prediction score computed per predicted delineations and in an attempt to make the model more robust, we empirically created a third "uncertain" category based on statistics from the validation dataset, for which the model could not predict with a high enough confidence whether the pre-processed image contained a LN or not. To define this new category, the lower bound corresponding to the lowest 5% of prediction scores in the distribution of pre-processed images with LNs and the upper bound corresponding to the highest 5% of scores in the distribution of pre-processed images without LNs in the validation dataset were extracted from the distribution of the confidence scores.

Analysis of the model's lymph node detection performance

We compared the performance of the 2 methods used for classification of pre-processed images on the validation, and external test datasets by comparing the normalised confusion matrices (i.e., the rows are divided by the sum of the rows which then add up to 1). Performance metrics calculated on the validation and external datasets were balanced accuracy, sensitivity, specificity, and F1-score.

We reported the feature importance via the Gini index of the trained XGBoost model,[17] the ROC curves of the candidate delineation classification prediction on the training, validation, and external test datasets with their confidence intervals at 95% calculated with 2000 bootstrapping of the results and the AUCs, along with the calibration curve based on the predictions obtained on the validation dataset. We also reported the results using the confusion matrices including the uncertain category on the validation and the external test datasets composed of the pre-processed images. To evaluate the added value of the uncertain category, we compared the false-negative and false-positive results on the external dataset with and without the uncertain category using 2 proportion z-test at a significance level of 0.05.

We used the maximum prediction score of a candidate delineation to establish the performance of the XGBoost model on the scanned images.

The scanned image predictions were used to calculate sensitivity and balanced accuracy per patient in the validation and external test datasets. We reported the mean sensitivity and the mean balanced accuracy per patient together with the violin plots of those metrics with a 2-sided Mann–Whitney-Wilcoxon test with Bonferroni correction to assess whether distributions of balanced accuracies and specificities were different between the validation and external test datasets.

Analysis of the false-negative results were performed by an expert pathologist. Observations were reported on the scanned images of the external test dataset containing LNs which remained undetected by the model to identify potential underlying causes or trends. The scanned images were considered false negatives when none of the pre-processed images belonging to those scanned images were falling in the category "uncertain" or "contain LN".

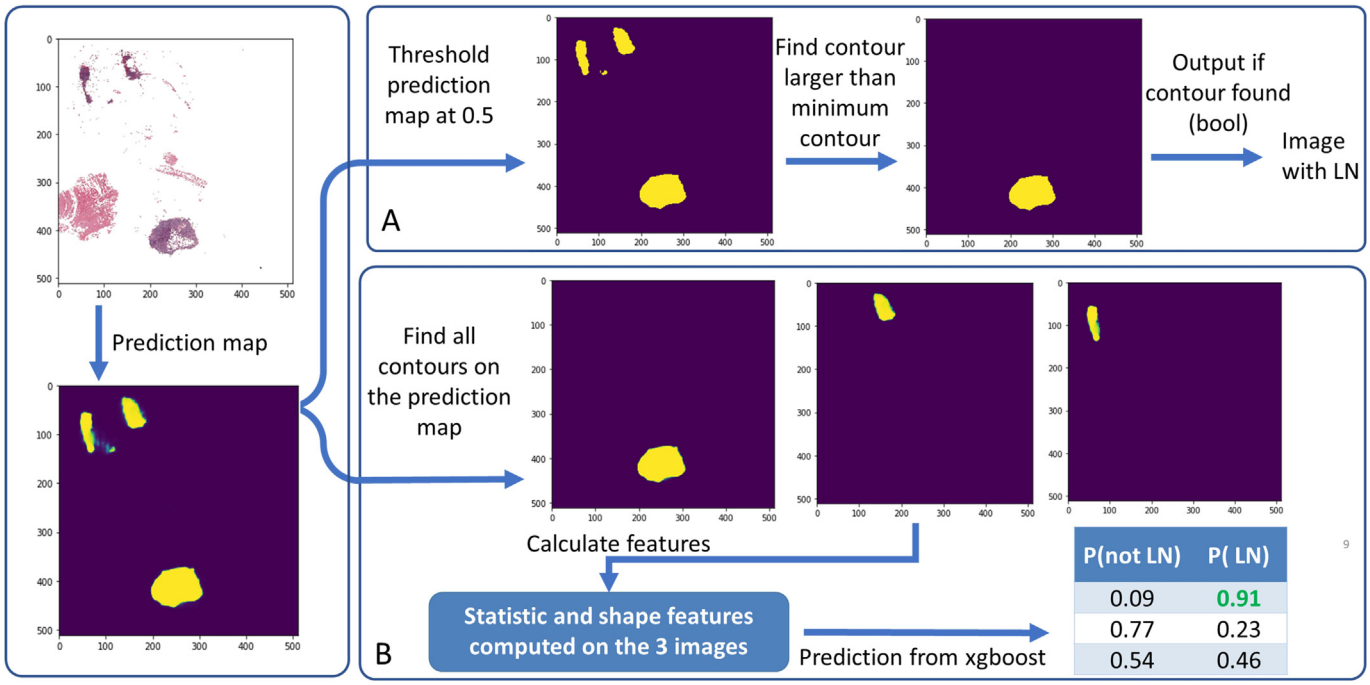

Fig. 2. The 2 strategies for predicting whether an image contains a lymph node. (A) "conventional" method, (B) our prediction score method.

Auto-delineation of the lymph nodes

Post-processing for auto-delineation extraction

If the pre-processed images had to be split in 2 during the pre-processing step, the delineations were performed on both pre-processed images and on a central image corresponding to both halves of the pre-processed images. Pixels with 2 probabilities were averaged. Finally, to evaluate the delineation results, we resized the pre-processed images to the original scanned image aspect ratio. Once the prediction was obtained from the scanned image, the delineations were automatically found as described in Suzuki and Be.[18] Areas with LN candidates were removed if the area was smaller than the minimum area computed previously. Next, the delineations were made convex to roughly resemble the natural shape of LNs.[19] From the scanned images, we filtered out the background to only delineate the tissue and increase the accuracy of the delineation, taking into consideration that potential concavity might not have been dealt with when making the delineations convex.

Analysis of lymph node delineation performance

The performance of the delineation model was reported using the average Dice coefficient [20] calculated from the original image, along with the average Dice per LNs from the validation and test dataset. Violin plots of the Dice coefficient per size category were reported. We calculated 4 size categories based on the distribution of areas of the ground-truth delineations in the original images in the training dataset, given in μm^2: (1) from the minimum area to the first quartile (Q1), (2) from Q1 to the median (M), (3) from M to the third quartile (Q3), and (4) from Q3 to the largest area. The distribution of the values within the violin plots were compared between the size categories using a 2-sided Mann–Whitney–Wilcoxon test with Bonferroni correction.

Results

After pre-processing of the datasets, the training dataset contained 1516 pre-processed images, the validation dataset 481, and the test dataset 1251 (see Table 1). Our U-Net model was trained for 28 epochs until the model began to overfit on the training dataset.

Evaluation of the lymph node detection performance

Comparison between conventional threshold method and newly developed prediction score method

Among the pre-processed images of the training dataset, 5% of the smallest LN area was equivalent to 167 pixels. The confusion matrices illustrating these results can be found in Fig. 3 panels A and B. The detection performance was reported Table 2.

The optimal number of features calculated for our newly developed prediction score system was 6, namely perimeter surface ratio, standard deviation, roundness, harmonic mean, number of contours, and centroid. The accuracy vs number of features curve calculated during the recursive feature elimination supporting the choice of number of features can be found in supplementary material Fig. 1 A. The optimum hyperparameters for XGBoost were found to be gamma = 0.8, learning rate = 0.01, maximum depth = 3, number of estimators was 1000, and minimum child weight = 5. Feature importance can be found in supplementary material Fig. 1B. The detection accuracies on the scanned images were 0.92 on the validation dataset and 0.85 on the test dataset. The confusion matrices illustrating these results can be found in Fig. 3C and D.

The AUCs of the training, validation, and test datasets were 0.98, 0.94, and 0.90, respectively. The ROC curves obtained on the training, validation, and test datasets are illustrated in supplementary material Fig. 2 and the calibration curve calculated on the validation dataset can be found in supplementary material Fig. 3.

For the interval of uncertainty we found the following values on the validation dataset: The lower boundary of the distribution scores for the pre-processed images which contained LNs at a 5% cut-off was found to be 0.48; the upper boundary at 95% obtained on the score of the pre-processed images which didn't contain LNs had a prediction score of 0.72. Table 3 displays the results found on the validation and external test datasets, respectively.

The uncertain category, i.e., a category which would require manual rechecking of the original image by a pathologist, comprised 6% of the validation dataset. The same proportion was obtained on the external test dataset using the lower and upper bounds calculated on the validation dataset. Comparing the results obtained in the uncertain table score to the confusion matrix in Fig. 3D, the false-negative rate was similar: 23 % in

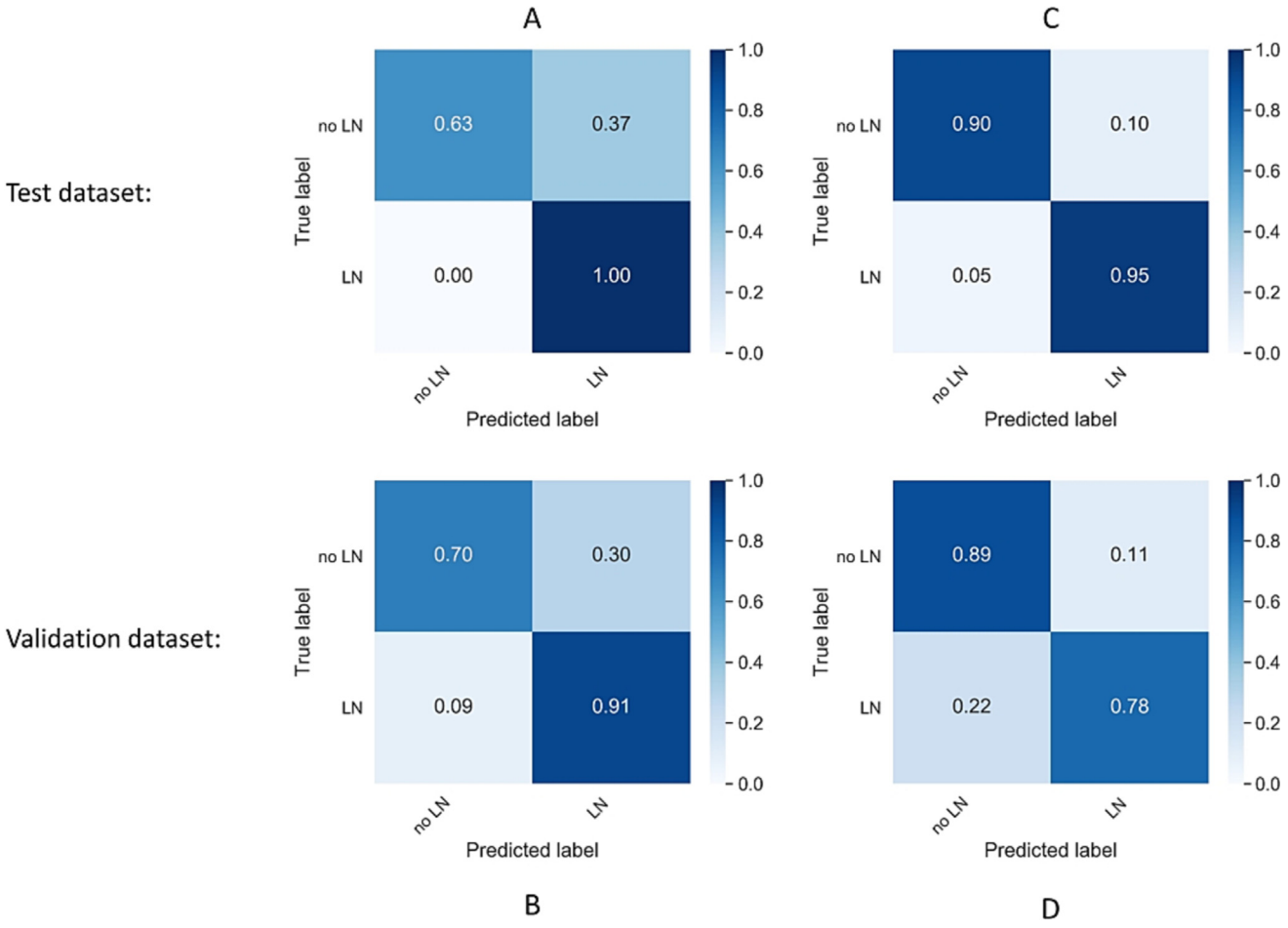

Fig. 3. Comparison of the results obtained on the original images to detect LNs in the validation and external test dataset between the "conventional" method and our prediction score method. (A) Confusion matrix "conventional" method for the validation dataset, (B) confusion matrix "conventional" method for the external test dataset, (C) confusion matrix using the prediction score method on the validation dataset, (D) confusion matrix using the prediction score method on the external test dataset.

Table 2
Balanced accuracy, specificity, sensitivity, and F1-score calculated on the validation and external test datasets for comparing the 2 classification methods. Bold indicates best performance on the external test dataset.

Method	Dataset	Balanced accuracy	Specificity	Sensitivity	F1-score
Our method	Validation	0.93 CI [0.90,0.95]	0.90 CI [0.87,0.94]	0.95 CI [0.92,0.98]	0.92 CI [0.89,0.94]
	Test	**0.83 CI [0.81,0.86]**	**0.89 CI [0.86,0.91]**	0.78 CI [0.74,0.82]	**0.77 CI [0.74,0.81]**
Conventional method	Validation	0.81 CI [0.79,0.84]	0.63 CI [0.57,0.68]	1.00 CI [1.00,1.00]	0.81 CI [0.76,0.84]
	Test	0.81 CI [0.78,0.83]	0.70 CI [0.67,0.73]	**0.91 CI [0.88,0.94]**	0.72 CI [0.69,0.75]

Table 3
Predictions on the validation and external test datasets split into 3 categories according to the level of certainty. The ground truth (image with or without LN) was obtained from manual review by a pathologist.

	Pre-processed images predicted to contain one or more LN n (%)	Uncertain category n (%)	Pre-processed images predicted to contain no LN n (%)	Total
	Dataset		Validation dataset	
Images with LN	181 (0.87)	17 (0.08)	11 (0.05)	209 (1)
Images without LN	14 (0.05)	11 (0.04)	247 (0.90)	272 (1)
Total	258 (0.53)	28 (0.06)	195 (0.41)	481 (1)
	Dataset		External test dataset	
Images with LN	280 (0.69)	32 (0.08)	92 (0.23)	404 (1)
Images without LN	61 (0.07)	33 (0.04)	753 (0.89)	847 (1)
Total	341 (0.38)	65 (0.06)	845 (0.56)	1 251 (1)

our uncertain table versus 22% in the confusion matrix (P = .73). However, we observed a significant decrease in false-positive findings: 7% in our uncertain table versus 11% in the confusion matrix (P < .05).

Accuracy and sensitivity distributions

The balanced accuracy and sensitivity distribution for the detection of LN per scanned image reported per patient in the validation and external test datasets is illustrated in supplementary material Fig. 4 . The mean sensitivities of the LN detection were 1 and 0.72 for the validation and test dataset, respectively. The mean balanced-accuracies for the LN detection were 0.58 for both the validation and test datasets.

False-negative analysis on the test dataset

30 (9%) scans out of 348 were classified as "not containing LNs" although they contained a LN. Fig. 4 summarises the description of those images.

We observed that some of the LNs 'undetected' by the algorithm were: (a) very small collections of lymphocytes which did not have a capsule or (b) did not display the usual LN microarchitecture with loss of lymphocytes and massive increase of macrophages occupying large part of the node, while others appeared 'empty', i.e., devoid of immune cells.

Evaluation of the model's lymph nodes delineation performance

The delineation performance was computed on the validation and external test datasets, comparing the ground truth delineated by a pathologist with the fully automatic delineation in original images containing LNs. The mean Dice score per original image was 0.73 and the mean Dice score per LNs was 0.66 for the validation dataset and 0.60 per original image and 0.48 per LNs for the external test dataset. The parameters used to create different intervals computed on the distribution of delineation areas in the train dataset were: Q1 = 278 061.5, M = 737 090.6, and Q3 = 1 707 021.1 (areas in μm^2). The violin plots of the Dice scores per interval for the validation and external test datasets are displayed in Fig. 5. Examples of different quality of auto-delineations are displayed in supplementary material Fig. 5.

Accurate delineation of small LNs (<Q1) seem to be significantly lower than the detection of the LNs at another size range, for both the validation and test datasets. However, the results are significantly different for the first and last categories (Fig. 5).

Discussion

In the current study, we developed a novel machine learning based pipeline to: (1) find and (2) delineate LNs in large collections of digitised H&E-stained slides from oesophagogastrectomy specimens and tested the performance on 1 independent dataset, while attempting to increase the explainability of the models. For finding the digital images containing LNs, we compared the performance of a conventional U-Net with thresholding method with our newly developed prediction score approach and observed a lower number of false-positives in both the validation and test datasets using our method. Furthermore, our approach had a higher accuracy in predicting whether a pre-processed image contains LN or not in the external test dataset (0.77 conventional method vs 0.81 our approach).

Another study to delineate LNs in H&E-stained images for gastric cancer patients using a U-Net architecture and thresholding reports a Dice score of 0.986 on the validation dataset.[21] The main difference to our study is that the training dataset of the U-Net model consisted exclusively of H&E-stained images containing LNs in every slide, meaning that the network would only have to exclude the background and small artefacts to allow LN delineation. Moreover, metrics per LN such as sensitivity or Dice score were not reported, leading to the performance of this model on small LNs to remain unknown.

When inspecting the feature importance within the XGBoost model ranked by the Gini coefficient, we observe that roundness and perimeter-to-surface ratio where among the 3 most important features. This correlates well with semantic knowledge that LNs often have an oval shape,[19] leading to irregular shapes being filtered out by our model. The delineation results obtained on the test dataset were adequate (mean Dice of 0.60 per original image, 0.48 for per LNs). When looking at the results divided by LN size in Fig. 5, the Dice scores for smaller LNs were significantly lower than for larger LNs. This is partially due to the penalisation of small structures by the Dice score, but might also be partially due to mislabelled data, where artefacts were wrongly attributed a label during pre-processing. Further analysis of the small LNs is being conducted. Although we attempted to limit the change of appearance of our data by splitting our images in 2 when the length/height difference was greater than 1.5, resizing the images into square images compresses the data and possibly negatively impacted our feature extractions and thus our results. Another pre-processing method such as tiling the extracted images instead of resizing them could alter the effect of the resizing.

Individual LNs can vary substantially in their microarchitecture[22] which can impact on the successful training of a DL model to identify LNs.[23] Delineations of the lymph nodes could be impacted by tumour invasion, as this

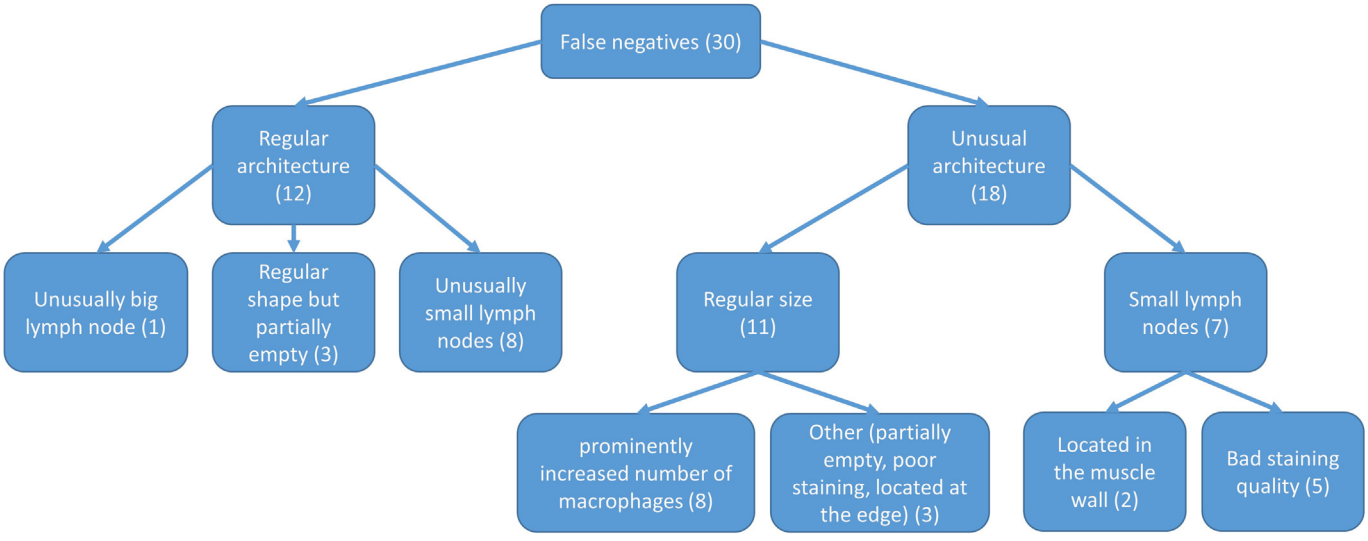

Fig. 4. Analysis of the LNs architecture in the scanned images belonging to the external test dataset wrongly categorised without LNs.

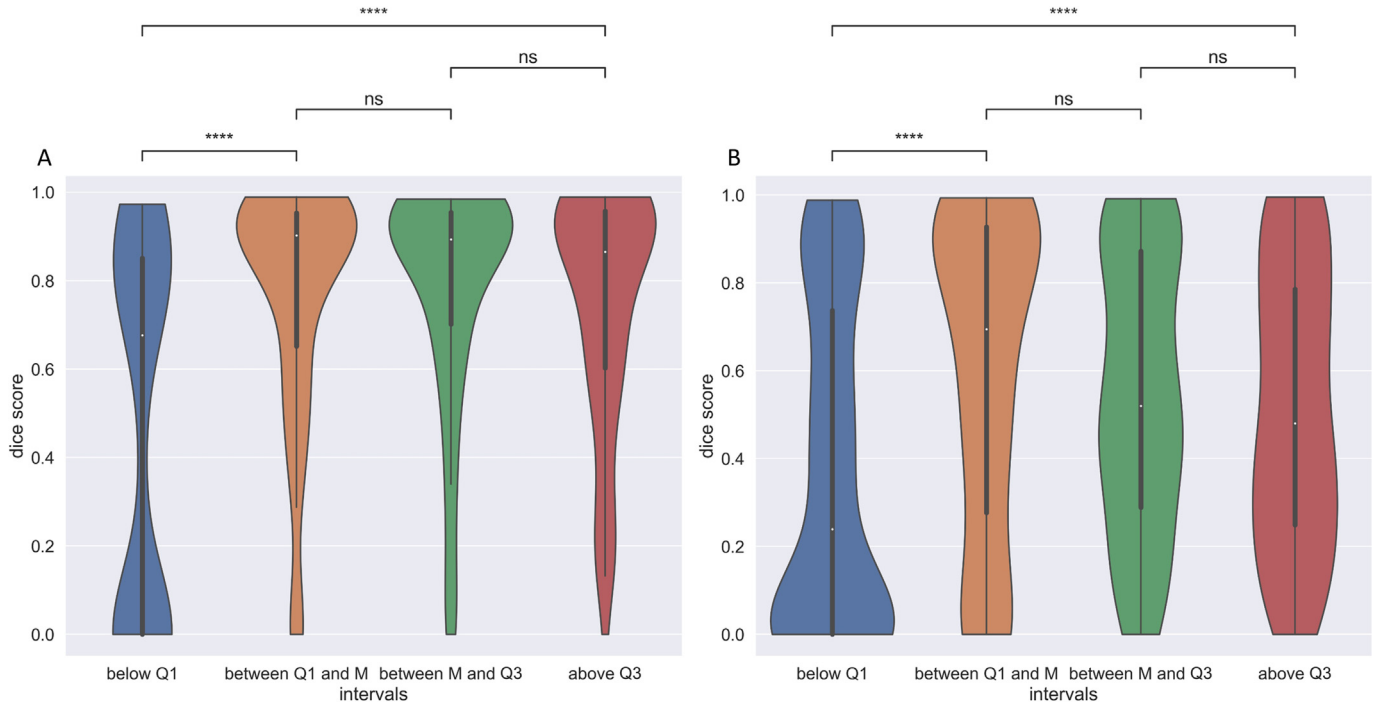

Fig. 5. Violin plot of the dice score per LNs split into 4 intervals defined based on the quartiles found on the ground-truth delineations distribution of the training dataset in μm^2 (A) on the validation dataset and (B) the external test dataset. Legend of the annotations (P-values): ns = non-significant i.e. $5 \times 10^{-2} <$ P-value $<= 1$; **** = P-value$<= 10^{-4}$

could change their structure and appearance. A follow-up study could evaluate the results on the positive and negative lymph nodes and show if a significant difference exists between the results in the 2 categories. Furthermore, H&E-stained tissue sections can vary in colour even if they originate from the same laboratory. Our analysis pipeline therefore included normalisation of the data, making the datasets less dependent from differences in staining. It is also possible that our normalisation method was not sufficient to prevent a domain shift in the external test dataset. Other method such as Fourier-based data augmentation as described in Wang et al.[24] could be adopted in a follow-up study to overcome this issue. We have chosen to train a U-Net model as this has been shown to be one of the most often-used models for automatic delineation in histopathology.[14] Further, our attempt to make the results of a U-Net model trained on histopathology data more explainable (certainty score, uncertain class creation, and most important features extracted from the prediction map) could solve 2 major roadblocks to clinical implementation: DL models lack explainability (the "black-box problem") and are incapable of assessing whether a new dataset is useable or should be rechecked by a pathologist (the "generalisability problem"). Our "uncertain" class could help solve this issue although our current results don't generalise well on the external dataset, with almost a quarter of pre-processed images being classified as not containing LNs while containing LNs (23%).

The different results observed between the validation and test datasets could be due to the fact that the validation dataset is from the same source as the training dataset while the test dataset is from another cohort.

Looking towards clinical application, our model for the detection and delineation of LNs could be integrated into software used for reviewing H&E-stained slides in the diagnostic setting and tested prospectively on H&E data from UGI patients. To obtain better delineations and detection results, we suggest implementing a continual learning process which would retrain the model with corrected delineations and detections predictions on the new dataset such as in Perkonigg et al.[25] A follow-up project will introduce analysis of handcrafted features extracted from the H&E-stained images (histomics analysis) to complete the work performed here and predict tumour infiltration within LNs.

In conclusion, we created a pipeline using deep learning for initial detection and handcrafted features to reinforce the predictions of a semantic delineation model which outperformed the conventional approach. Thanks to our scoring model, we could create an uncertain category for which the model is not confident to classify the image into with or without LNs that pathologists would have to review. Although good performance was obtained on the validation dataset, medium performance was obtained on the test dataset for both the classification and delineation tasks, which might be due to high heterogeneity in the external test dataset which might not have been there in the training dataset. The first part of our workflow could be used in a routine diagnostic setting for H&E-stained images of esophageal tissue after further prospective validation, and the second part could be useful for further work on measuring LN areas and characterising the structure of LNs, potentially useful for personal treatment planning for patients with UGI cancer.

Ethics Approval and Consent to Participate

The study was approved by the South East Research Ethics Committee, London, United Kingdom, REC reference: 07/H1102/111.

Funding

This work was made possible through the support of Marie Skłodowska-Curie grant (PREDICT - ITN - No. 766276). Authors furthermore acknowledge financial support from ERC-2020-PoC: 957565-AUTO.DISTINCT, CHAIMELEON no. 952172, EuCanImage no. 952103, the Dutch Cancer Society (KWF Kankerbestrijding) project number 12085/2018–2 and IMI-OPTIMA n° 101034347. The material collection was funded by Cancer Research UK (grant number C26441/A8944 to PI H.I.G.). H.I.G. is supported in part by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Conflict of interest

H.W. has minority shares in the company Radiomics SA. D.C. declares grants from Medimmune/AstraZeneca, Clovis, Eli Lilly, 4SC, Bayer, Celgene, Leap, and Roche, and Scientific Board Membership for OVIBIO. P.L. has minority shares in the companies Radiomics SA, Convert pharmaceuticals, Comunicare, and LivingMed Biotech, and he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248 and PCT/NL2014/050728), licensed to Radiomics SA; one issued patent on mtDNA (PCT/EP2014/059089), licensed to ptTheragnostic/DNAmito; one non-issued patent on LSRT (PCT/ P126537PC00), licensed to Varian; three non-patented inventions (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and two non-issued, non-licensed patents on Deep Learning-Radiomics (N2024482, N2024889). He confirms that none of the above entities or funding sources were involved in the preparation of this paper. H.I.G. reports personal fees from Merck Sharp & Dohme, outside the submitted work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2023.100192.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–249. https://doi.org/10.3322/caac.21660.
2. Lordick F, Mariette C, Haustermans K, Obermannová R, Arnold D, Committee EG. Oesophageal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2016;27:v50–v57. https://doi.org/10.1093/annonc/mdw329.
3. Smyth EC, Verheij M, Allum W, Cunningham D, Cervantes A, Arnold D. Gastric cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 2016;27:v38–v49. https://doi.org/10.1093/annonc/mdw350.
4. Arnold M, Morgan E, Bardot A, et al. International variation in oesophageal and gastric cancer survival 2012–2014: differences by histological subtype and stage at diagnosis (an ICBP SURVMARK-2 population-based study). Gut 2021. https://doi.org/10.1136/gutjnl-2021-325266.gutjnl-2021-325266.
5. Smyth EC, Fassan M, Cunningham D, et al. Effect of pathologic tumor response and nodal status on survival in the medical research council adjuvant gastric infusional chemotherapy trial. J Clin Oncol 2016;34:2721–2727. https://doi.org/10.1200/jco.2015.65.7692.
6. Davarzani N, Hutchins GGA, West NP, et al. Prognostic value of pathological lymph node status and primary tumour regression grading following neoadjuvant chemotherapy - results from the MRC OE02 oesophageal cancer trial. Histopathology 2018;72:1180–1188. https://doi.org/10.1111/his.13491.
7. Medical Research Council Oesophageal Cancer Working G. Surgical resection with or without preoperative chemotherapy in oesophageal cancer: a randomised controlled trial. Lancet 2002;359:1727–1733. https://doi.org/10.1016/S0140-6736(02)08651-8.
8. Kloft M, Ruisch JE, Raghuram G, et al. Prognostic significance of negative lymph node long axis in esophageal cancer: results from the randomized controlled UK MRC OE02 trial. Ann Surg 2021. https://doi.org/10.1097/sla.0000000000005214.
9. Alderson D, Cunningham D, Nankivell M, et al. Neoadjuvant cisplatin and fluorouracil versus epirubicin, cisplatin, and capecitabine followed by resection in patients with oesophageal adenocarcinoma (UK MRC OE05): an open-label, randomised phase 3 trial. Lancet Oncol 2017;18:1249–1260. https://doi.org/10.1016/S1470-2045(17)30447-3.
10. Li X, Li C, Rahaman MM, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. Artif Intel Rev 2022. https://doi.org/10.1007/s10462-021-10121-0.
11. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Xiaojun G et al. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 1107-1110.
12. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 1979;9:62–66. https://doi.org/10.1109/TSMC.1979.4310076.
13. Ronneberger O, Fischer P, Brox T. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing. 2015:234–241.
14. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal 2021;67, 101813. https://doi.org/10.1016/j.media.2020.101813.
15. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint 2014. arXiv:1412.6980.
16. Spearman C. The proof and measurement of association between two things. By C. Spearman, 1904. Am J Psychol 1987;100:441–471.
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Routledge. 2017.
18. Suzuki S, Be K. Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing, 30. 1985. p. 32–46. https://doi.org/10.1016/0734-189X(85)90016-7.
19. Ganeshalingam S, Koh D-M. Nodal staging. Cancer Imaging 2009;9:104–111. https://doi.org/10.1102/1470-7330.2009.0017.
20. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.
21. Wang X, Chen Y, Gao Y, et al. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. Nat Commun 2021;12:1637. https://doi.org/10.1038/s41467-021-21674-7.
22. Elmore SA. Histopathology of the lymph nodes. Toxicol Pathol 2006;34:425–454. https://doi.org/10.1080/01926230600964722.
23. Wu Y, Cheng M, Huang S, et al. Recent advances of deep learning for computational histopathology: principles and applications. Cancers 2022;14:1199.
24. Wang X, Zhang J, Yang S, et al. A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. Med Image Anal 2023;84, 102703. https://doi.org/10.1016/j.media.2022.102703.
25. Perkonigg M, Hofmanninger J, Herold CJ, et al. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nat Commun 2021;12:5678. https://doi.org/10.1038/s41467-021-25858-z.