Mod Pathol 36 (2023) 100118

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Computer-Assisted Diagnosis of Lymph Node Metastases in Colorectal Cancers Using Transfer Learning With an Ensemble Model

Amjad Khan[a], Nelleke Brouwer[b], Annika Blank[c], Felix Müller[a], Davide Soldini[d], Aurelia Noske[d,e], Elisabeth Gaus[d], Simone Brandt[d], Iris Nagtegaal[b], Heather Dawson[a], Jean-Philippe Thiran[f,g], Aurel Perren[a], Alessandro Lugli[a], Inti Zlobec[a,*]

[a] *Institute of Tissue Medicine and Pathology, University of Bern, Bern, Switzerland;* [b] *Department of Pathology, Radboud University Medical Centre, Netherlands;* [c] *Institute of Pathology, City Hospital Triemli, Zürich, Switzerland;* [d] *Institute of Clinical Pathology Medica, Zürich, Switzerland;* [e] *Institute of Pathology, School of Medicine, Technical University of Munich, Munich, Germany;* [f] *Department of Radiology, Lausanne University Hospital, Lausanne University and Centre d'Imagerie Biomédicale, Lausanne, Switzerland;* [g] *Swiss Federal Institute of Technology Lausanne, Signal Processing Laboratory, Lausanne, Switzerland*

ARTICLE INFO

Article history:
Received 27 October 2022
Revised 14 December 2022
Accepted 21 January 2023
Available online 2 February 2023

Keywords:
colorectal cancer
ensemble model
histopathology
lymph nodes
metastasis detection
transfer learning

ABSTRACT

Screening of lymph node metastases in colorectal cancer (CRC) can be a cumbersome task, but it is amenable to artificial intelligence (AI)-assisted diagnostic solution. Here, we propose a deep learning−based workflow for the evaluation of CRC lymph node metastases from digitized hematoxylin and eosin−stained sections. A segmentation model was trained on 100 whole-slide images (WSIs). It achieved a Matthews correlation coefficient of 0.86 (±0.154) and an acceptable Hausdorff distance of 135.59 μm (±72.14 μm), indicating a high congruence with the ground truth. For metastasis detection, 2 models (Xception and Vision Transformer) were independently trained first on a patch-based breast cancer lymph node data set and were then fine-tuned using the CRC data set. After fine-tuning, the ensemble model showed significant improvements in the F1 score (0.797-0.949; P <.00001) and the area under the receiver operating characteristic curve (0.959-0.978; P <.00001). Four independent cohorts (3 internal and 1 external) of CRC lymph nodes were used for validation in cascading segmentation and metastasis detection models. Our approach showed excellent performance, with high sensitivity (0.995, 1.0) and specificity (0.967, 1.0) in 2 validation cohorts of adenocarcinoma cases (n = 3836 slides) when comparing slide-level labels with the ground truth (pathologist reports). Similarly, an acceptable performance was achieved in a validation cohort (n = 172 slides) with mucinous and signet-ring cell histology (sensitivity, 0.872; specificity, 0.936). The patch-based classification confidence was aggregated to overlay the potential metastatic regions within each lymph node slide for visualization. We also applied our method to a consecutive case series of lymph nodes obtained over the past 6 months at our institution (n = 217 slides). The overlays of prediction within lymph node regions matched 100% when compared with a microscope evaluation by an expert pathologist. Our results provide the basis for a computer-assisted diagnostic tool for easy and efficient lymph node screening in patients with CRC.

Introduction

The Tumor-Node-Metastasis staging system, defined by the American Joint Committee on Cancer, is a prognosis-prediction

* Corresponding author.
E-mail address: inti.zlobec@unibe.ch (I. Zlobec).

https://doi.org/10.1016/j.modpat.2023.100118

factor and a determining factor in decision making for stage-based therapeutic strategies in patients with colorectal cancer (CRC). N-staging of a patient with CRC is determined by the number of metastatic lymph nodes, which includes the presence of micrometastases, ranging in size from 0.2 to 2.0 mm.[1] A minimum of 12 lymph nodes need to be assessed, but the number of lymph nodes actually resected can often exceed 50. Therefore, the histologic evaluation of lymph node metastasis in patients with CRC can be laborious and would benefit from artificial intelligence (AI)-assisted workflow,[2] as already implemented for breast and gastric cancer.[3-5] A supervised deep learning method requires annotations of tumor regions on whole-slide images (WSIs) for training, which can be intensive and time consuming for pathologists. Multiple instance learning and end-to-end training with annotation-free WSIs have shown better performance without any regional annotation requirements.[6,7] To train such pipelines and achieve reasonable performance requires thousands of WSIs, with global slide labels of "positive" and "negative," which is not practical.

In this study, we sought to develop an AI-based screening method for lymph node metastases in CRC using a more feasible method. We used the former methods based on regional annotation, but instead of performing a vast amount of regional annotations, we trained deep learning models on a publicly available data set of lymph nodes from breast cancer cases.[8,9] Using transfer learning to fine-tune the models on a few WSIs (n = 14) from CRC lymph node tissues, we achieved exceptional performance in multiple validation cohorts. Our workflow consisted of 3 parts: first, a segmentation model trained to segment the lymph node tissues and, second, a convolutional neural network (particularly, Xception), and a Vision Transformer (ViT16), first trained on a breast cancer data set and then fine-tuned using a CRC data set. On training, both models were used to create an ensemble model to detect metastases in each segmented lymph node region. Third, the cascaded model's workflow segmented the lymph node, classified the WSI as "positive" or "negative," and eventually generated overlays of the potential metastatic regions for further evaluation by experts.

Materials and Methods

Data Sets

For the 2 main tasks (lymph node segmentation and metastasis detection), 7 different data sets with hematoxylin and eosin (HE)–stained slides were used as summarized in Table 1; these included 3 data sets for model training and development and 4 independent validation cohorts. For the segmentation model development, 100 WSIs of lymph node regions from 14 patients with CRC, annotated by 2 expert pathologists (A.B., F.M.), were used. We will refer to this data set as LnSegment throughout the article.

To develop a metastasis detection model, the PatchCamelyon tiles from lymph nodes in breast cancer cases, extracted from the Camelyon16 challenge data set, were used.[5,8] PatchCamelyon data set had 327,680 tiles, of size 96 × 96 pixels, from normal and tumoral lymph node tissue. Within each tumor tile, the central 32 × 32 pixels contained at least 1 tumor pixel. Similarly, the PatchCRC data set was constructed, which consisted of the normal and tumoral regions within the lymph nodes of 5 patients with CRC, annotated by 2 expert pathologists (A.B., F.M.). Those annotated regions were used to extract 53,814 patches of the same size as PatchCamelyon. The PatchCRC was used to fine-tune the metastasis detection model. In addition, 4225 WSIs consisting of 4 independent cohorts from the internal patient repository (Institute of Tissue Medicine and Pathology, University of Bern, Switzerland) and an external center (Department of Pathology, Radboud University Medical Centre, Netherlands) were used to validate the deep learning workflow. The external validation cohort contained mostly metastatic slides from patients with stage 3 CRC, whereas the internal cohorts were scanned regardless of patient stage and included both metastatic and normal slides. Internal cohort 1 consisted of 2803 WSIs (metastatic: 609, normal: 2194) from 298 patients with CRC. The external validation cohort contained 1033 WSIs (metastatic: 1019, normal: 14) of 400 patients with CRC. Both internal cohort 1 and the external cohort contained adenocarcinoma cases of no special type. Additionally, internal cohort 2 contained 172 WSIs of mucinous adenocarcinoma and signet-ring cell carcinomas (metastatic: 94, normal: 79) from 34 patients. Similarly, a consecutive case series of 217 WSIs (metastatic: 23, normal: 194) from 16 patients with CRC, regardless of histologic subtypes, diagnosed within the last 6 months, were acquired to form internal cohort 3, which was used to validate the workflow in clinical settings. In the validation cohorts, the slides with a pathologist's report of the presence of any metastatic cancer cells were labeled as "positive" and slides without any reported metastases were labeled as "negative." All above-mentioned data sets (except PatchCamelyon) were scanned with a Pannoramic P1000 digital slide scanner (3DHistech) at 40× magnification (ie, 20× objective magnification and 2× aperture boost), with a 0.243-μm pixel resolution. The validation WSIs contained lymph node tissues; 1 slide with no lymph node tissue was discarded.

Model Development

The complete deep learning methodology followed in this study is shown in Figure 1. In the first step, the segmentation model (UNet) was trained using the LnSegment data set to segment the lymph node tissue on each WSI using 5-fold cross-validation.[10] In the next step, 2 neural network models, Xception and ViT16, were independently trained on the PatchCamelyon

Table 1
Seven different types of data sets used in this study to develop and validate the colorectal cancer (CRC) lymph node metastases-detection workflow. All data sets are from patients with CRC, except for the PatchCamelyon data set, which is from patients with breast lymph nodes

Model type	Data set	Image type	No. of patients	Lymph node type	No. of tiles/WSIs	Tile size (pixels)	Model operation
Lymph node segmentation	LnSegment	WSIs	14	Colon	100	—	Training
Metastasis detection	PatchCamelyon	Tiles	—	Breast	327, 680	96 × 96	Training
	PatchCRC	Tiles	5	Colon	53, 814	96 × 96	Fine-tuning
	Internal cohort 1	WSIs	298	Colon	2803	—	Validation
	Internal cohort 2	WSIs	34	Colon	172	—	Validation
	Internal cohort 3	WSIs	16	Colon	217	—	Validation
	External cohort	WSIs	400	Colon	1033	—	Validation

WSIs, whole-slide images.

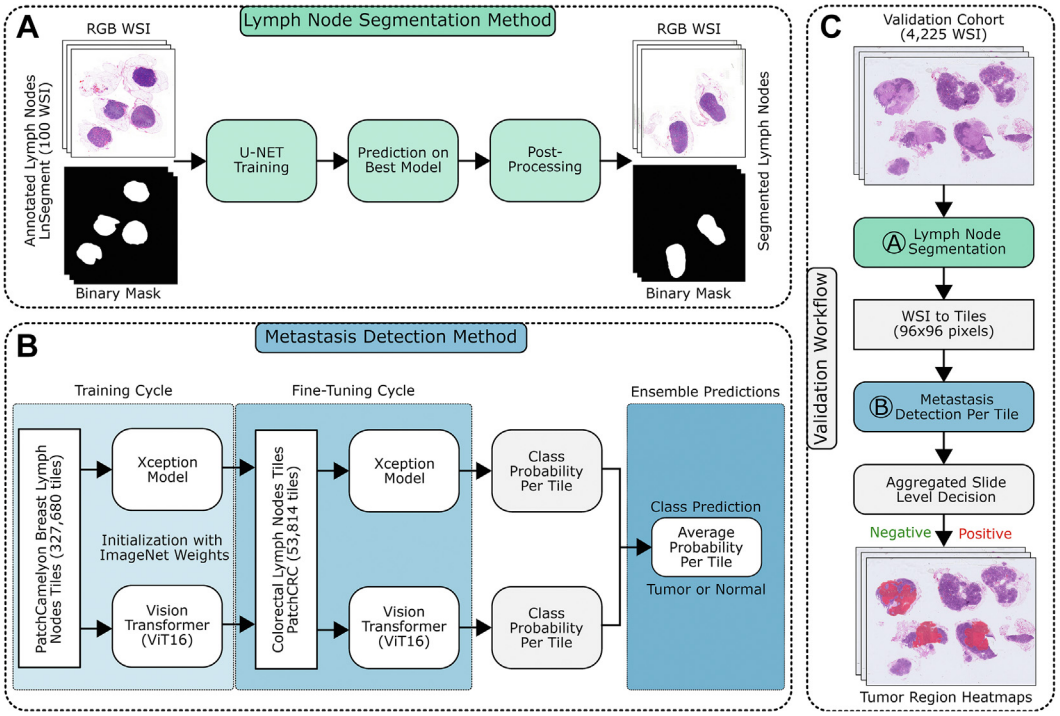

Figure 1.
The deep learning workflow for colorectal cancer (CRC) lymph node quantification consisting of 3 main tasks. In the first task (A), the segmentation model (UNet) is trained using the LnSegment data set to segment the lymph node tissue fragment on each whole-slide image (WSI) using 5-fold cross-validation. In the second task (B), 2 neural network models, Xception and Vision Transformer (ViT16), are independently trained on PatchCamelyon (a publicly available breast cancer lymph node data set) for a positive and negative tissue classification task, by initializing them on ImageNet weights for a long training cycle and are then fine-tuned on a PatchCRC data set for a short cycle. Once these models are successfully trained, they are validated on the corresponding test data set, where the average outcome of both models is used to generate the final probability per tile. In the third task (C), by cascading both segmentation (A) and metastasis detection models (B), validation is performed. Each WSI in the validation cohorts is first passed through the segmentation model (A) to segment the lymph node tissues. Each segmented tissue is then split into 96- × 96-pixel tiles and passed to the metastasis detection model (B) to obtain the per-tile probability. The most probable tiles, with a class score of 85% or higher and with the absence or presence of more than 2 such connected tiles, are used to define the slide-level label as negative or positive, respectively. Finally, the per-tile probability within each tissue fragment is used to aggregate the overlays on WSIs in QuPath (as shown in red).

data set for a positive and negative tissue classification task, by initializing them on ImageNet weights for a long training cycle, and were then fine-tuned on the PatchCRC data set for a short cycle.[11,12] Once these models were successfully trained, they were validated on the corresponding test data set, where the average outcome of both models was used to generate the final probability for each tile. In the final step, cascading both segmentation and metastasis detection models, validation was performed in large cohorts of lymph node WSIs from patients with CRC. Each WSI in the validation cohort was passed through the segmentation model to obtain the lymph node tissue. Each segmented tissue fragment was then split into tiles and fed to the metastasis detection step to obtain a per-tile probability. In the absence or presence of more than 2 connected tiles, with a class probability higher than a given threshold, the slide-level label was defined either as negative or positive, respectively. Finally, the per-tile probability within each tissue fragment was added to the aggregated overlays associated with WSIs in QuPath.[13] The following subsections will explain the model training and validation steps in detail.

Segmentation Model Training

Precise lymph node segmentation is a critical factor in proper metastasis detection. The segmentation model training was derived from our recent study on the impact of scanner variability on lymph node segmentation and generalizing such models on WSIs from different scanners.[14] We rescanned 100 glass slides of CRC lymph nodes by using 4 different scanners from 3 different vendors. Digitized WSIs were annotated by 2 experienced pathologists for lymph node regions (ie, the LnSegment data set). We performed extensive and systematic experiments to evaluate the impact of scanner variability and generalizability of lymph node segmentation methods.[10,15-17] To cope with performance variances, we evaluated various stain normalization approaches by creating reference mosaic images from the foreground and background regions of lymph node WSIs.[18-21] Similarly, domain generalization techniques, such as stain mix-up, domain adversarial learning, and fine-tuning were evaluated to generalize the segmentation methods.[14,22,23]

The workflow of the lymph node segmentation method is shown in Figure 1A. From each WSI in the LnSegment data set, the first 20× magnification images were obtained and subsequently downscaled by a factor of 64, to allow them to fit into memory. The corresponding ground truth annotations were similarly downscaled. To train the UNet, the downscaled WSIs were rescaled further to fit the network. To evaluate the method across all samples, the LnSegment data set was divided into training ($n = 80$) and test ($n = 20$) sets by using 5-fold cross-validation. To minimize overfitting, the training samples were augmented mainly using methods that could imitate data sets from different sources, such as using contrast and stain variations as described in Table 2.[24-29] The network was trained for 200 epochs with a learning rate of 1×10^{-3} which was reduced by a factor of 0.1 on

Table 2
Different data augmentation strategies and training specifications used during the development of segmentation and metastasis detection models

Data augmentation and training specifications	Segmentation model	Metastasis detection Model	
		Xception	ViT16
Flipping horizontal and vertical	✓	✓	✓
Rotation	✗	(0°-45°)	(0°-90°)
Shifting	✗	✗	(0%-2%)
Scaling	✗	✗	(0%-2%)
Brightness change	Multiplying by a factor [0, 0.75] and adding a constant between 0 and 15[24]	✗	✗
Stain augmentation	By RGB to HED and using linear contrast between 0.5 and 0.2[24,25]	By Staintools StainAugmentor (method = Macenko, σ1 = 0.2, σ 2 = 0.5)[18,26]	By RGB to HED and using linear contrast function with a factor between 0.2 and 0.3[19,24]
Elastic transform	✗	(α = 1, σ = 3, affine = 1, interpolation = 1)[27]	✗
Contrast change	✗	Contrast limited adaptive histogram equalization (clip limit = 7.0, tile grid size = [8,8])[27]	✗
Rescaled input size (pixels)	512 × 512	150 × 150	224 × 224
Input intensity normalization	Between 0 and 1	Between −1 and 1	Between 0 and 1
Loss function	Binary cross-entropy	Binary cross-entropy	Sparse categorical cross-entropy
Optimizer	Adam[28]	Adam[28]	Adam with weight decay[29]

reaching a plateau. In the postprocessing step, the binary mask was obtained by including the pixels with a probability higher than 50%. The binary mask was refined further by using a conditional random field[30] and dilated with a filter of size 5 × 5 pixels to maximize it beyond the capsular sinus regions for the metastasis analysis in the next step.

Metastasis Detection Model Training

We anticipated that tumor and normal regions from both breast and CRC lymph nodes would have similar morphology when considering a small, tiled region. However, the data sets may also have differences owing to varied staining and scanning conditions. Therefore, to fit the model to the new data set, we used a transfer learning approach,[31-33] where we established a model on a largely annotated and publicly available data set from a different tissue type (breast cancer lymph nodes) and then retrained it with a new, small, annotated data set (CRC lymph nodes), to account for any differences. As shown in Figure 1B, first, we independently trained Xception and ViT16 neural networks on the PatchCamelyon training set for a tumor and normal tissue classification task, by initializing them with ImageNet weights. Both models were slightly modified by introducing resizing, preprocessing, and dropout layers. For the Xception model, the input size was resized by the preprocessing layer using the bilinear interpolation method, and pixel values were scaled per sample to allow the use of ImageNet weights for the initialization (Table 2). After global average pooling and before the final output layer, a 50% dropout was used to prevent potential overfitting. Initially, the Xception model was trained for 5 epochs on a batch size of 256 by freezing all of the layers containing the ImageNet weights, with a learning rate of 1×10^{-3}. The whole model was then trained by unfreezing the ImageNet weight layers for another 25 epochs, with the same batch size, but with a lower learning rate of 1×10^{-5}. Similarly, for ViT16 training, the input was resized, and pixel values were normalized with respect to the ImageNet data set, by resizing and preprocessing layers, respectively. The model was trained for 30 epochs on a batch size of 64 by using a learning rate and a weight decay factor of 2×10^{-6} and 5×10^{-7}, respectively. In ViT16, each input sample of 224 × 224 pixels was divided into 196 small patches, called visual tokens, with each token having the size of 16 × 16 pixels. To learn different features on the same training data set, different augmentation strategies were used for the 2 models. The data augmentation methods and other training specifications for both Xception and ViT16 are presented in Table 2. During the training iterations of both models, only the weights with the best validation performance were saved.

On completion of the training cycle on PatchCamelyon, the best-validated weights from both Xception and ViT16 models were used to fine-tune the training set of PatchCRC. In the fine-tuning cycle, all abovementioned parameters described for each model remained the same, except that both models were trained up to 15 epochs on the new data set (22,269, 16,770, and 14,775 tiles for training, validation, and testing, respectively). After the fine-tuning step, the outcome of each model was combined by averaging the probabilities to obtain the score per tile. Finally, the performance on the corresponding test sets before and after creating an ensemble model was reported.

Validation and Visualization

On successful training and testing of both segmentation and metastasis detection models on the corresponding training and test data sets, they were cascaded into a validation workflow (Figure 1C). In the first step, each WSI from the validation cohorts was passed through the segmentation model to obtain the lymph node tissue fragments. Each segmented region was then divided into 96 × 96 pixels tiles to pass over to the metastasis detection model to obtain the per tile probability for a tumor or normal class. In every WSI, each segmented lymph node tissue was screened to find more than 2 connected tiles, each with a probability of 85% or higher. In the presence or absence of such a single region, the slide was marked as positive or negative, respectively. These predicted slide labels were compared with ground truth labels provided by pathologists to report the final score per validation cohort. To visualize the potential metastatic areas within each lymph tissue fragment on a positive WSI, a heat map was generated for each WSI by

aggregating the per-tile probability. During the inference, the information about the coordinates of the lymph node tissue, per-tile probability, and corresponding coordinates within each tissue were exported into a standard comma-separated value (CSV) file for each WSI. The CSV file was then imported to the QuPath project, where all processed WSIs were already set up. Using a Groovy script (ie, the standard scripting language used in QuPath), the CSV file information was used to generate overlays or heat maps. These overlays were useful to assess the potential metastatic regions within each lymph node tissue visually.

Performance Evaluation

The lymph node segmentation model was quantified using 2 metrics: Matthews correlation coefficient (MCC)[34,35] and the Hausdorff distance (HD).[36] The MCC between the ground truth and segmented labels was calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} (1),$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. MCC ranges between −1 and +1, where +1 represents a perfect prediction, 0 an average random prediction, and −1 an inverse prediction, when compared with the ground truth. Boundary differences were measured using the HD. HD calculates the maximum Euclidean distance from all minimum distances between the boundaries of ground truth (A) and the boundaries of the segmentation region (B), as follows:

$$HD(A, B) = max(h(A, B), h(B, A)) (2),$$

where h(A, B) is the directed HD, based on the following equation:

$$h(A, B) = max_{a \in A}\, min_{b \in B} \parallel a - b \parallel (3),$$

where $\|a - b\|$ represents the Euclidean distance. The HD between 2 perfectly overlapping boundaries is equal to zero. Expected ranges of HD for good, acceptable, and bad scores could be <100 μm, 100-150 μm, and >150 μm, respectively.[14] Similarly, for evaluating our metastasis detection model, we used a weighted F1 score, the area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity. For statistical analysis, the Wilcoxon signed-rank test was used.[37]

Results

The results obtained in different experiments, using various test and validation data sets, as evaluated using metrics such as MCC, HD, AUROC, F1 score, sensitivity, and specificity, are presented in Table 3.

Lymph Node Segmentation

As an initial step in this study, the UNet model was trained on the LnSegment data set to segment the lymph node tissue on WSIs. We evaluated the trained UNet model using 5-fold cross-validation and achieved an MCC score of 0.86 (±0.154) and boundary losses (HD) of 135.59 μm (±72.14 μm).[14]

Metastasis Detection

For metastasis detection, the Xception and ViT16 models were independently trained on PatchCamelyon and fine-tuned on PatchCRC data sets. The results of both training and fine-tuning cycles and of the ensemble model, along with AUROC curves, are presented in Figure 2. The Xception model achieved AUROC and F1 scores of 0.968 and 0.902, respectively, when evaluated on the PatchCamelyon test set. Likewise, when the ViT16 model was assessed on the same test set, it yielded an AUROC and F1 score of 0.962 and 0.891, respectively. The Xception and ViT16 models showed statistically significantly different performance on the same test samples (P <.00001). Therefore, their ensemble model showed slightly improved AUROC and F1 scores (0.974 and 0.910, respectively) when compared with the Xception (P =.012) and ViT16 (P <.00001) models alone (Fig. 2A).

When tested on the PatchCRC data set, the Xception, ViT16, and ensemble model trained on the PatchCamelyon data set attained AUROCs of 0.949, 0.958, and 0.959, respectively, and F1 scores of 0.775, 0.799, and 0.797, respectively (Fig. 2B). The ensemble model did not show statistically significantly improved performance when compared with the ViT16 (P =.673) model on the PatchCRC test set. However, other comparisons (ie, Xception vs ViT16 and Xception vs ensemble model) were statistically significantly different (P <.00001).

The fine-tuning step, which used the PatchCRC data set for further training of the models that had previously been trained only on the PatchCamelyon data set, yielded improved performance (P <.00001) compared with the PatchCamelyon-only trained models when both were evaluated using the PatchCRC

Table 3
The results of lymph node segmentation and metastases detection (ensemble) model when evaluated on various test and validation data set using corresponding metrics, such as MCC, HD, AUROC, F1 score, sensitivity, and specificity

Model type	Data set	MCC (±SD)	HD (±SD), μm	AUROC	F1 score	SN	SP
Lymph node segmentation	LnSegment	0.860 (±0.154)	135.59 (±72.14)	—	—	—	—
Metastasis detection	PatchCamelyon	—	—	0.974	0.910	—	—
	PatchCRC (without fine-tuning)	—	—	0.959	0.797	—	—
	PatchCRC (with fine-tuning)	—	—	0.978	0.949	—	—
	Internal cohort 1	—	—	—	0.974	0.995	0.967
	Internal cohort 2	—	—	—	0.901	0.872	0.936
	Internal cohort 3	—	—	—	1.0	1.0	1.0
	External cohort	—	—	—	1.0	1.0	1.0

MCC, Matthews correlation coefficient; HD, Hausdorff distance; AUROC, area under the receiver operating characteristic curve; SN, sensitivity; SP, specificity.

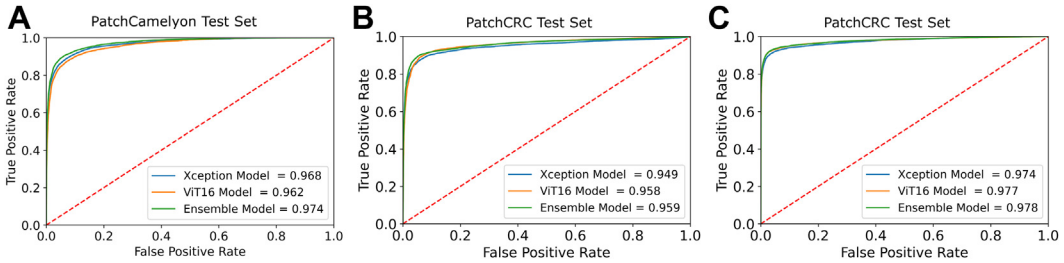

Figure 2.
The areas under the receiver operating characteristic curves (AUROC) for Xception, ViT16, and their ensemble model: (A) trained and tested on the tiles from PatchCamelyon, (B) trained on PatchCamelyon and tested on PatchCRC, and (C) trained on PatchCamelyon, fine-tuned on PatchCRC, and tested on PatchCRC. The AUROC scores show that the model trained on breast lymph nodes has improved performance on the colorectal cancer (CRC) data set when the model is fine-tuned on the CRC lymph node data set.

data set. The fine-tuned Xception and ViT16 models yielded AUROCs of 0.974 and 0.977, respectively, and F1 scores of 0.939 and 0.949, respectively. Their fine-tuned ensemble model presented the highest performance, with an AUROC of 0.978 and an F1 score of 0.949, when tested on the PatchCRC data set (Fig. 2C).

Validation on 3 Independent Cohorts

On successful evaluation of both segmentation and metastasis detection models on the respective testing data sets, they were applied to validation cohorts by following the steps of the validation workflow in Figure 1C. The final ensemble model showed excellent performance in both external and internal validation cohorts when comparing predicted positive and negative slide-level labels with the pathologists' ground truth. The ensemble model performed 100% accurately on the external cohort of patients with CRC when evaluated by using F1 score, sensitivity, and specificity metrics. Some examples of detected lymph node metastases in the external validation cohort are shown in Figure 3A, B.

However, in internal cohort 1, consisting of the same type of patients, the model performance dropped slightly to 0.974, 0.995, and 0.967 on F1, sensitivity, and specificity, respectively, compared with the external cohort results. Approximately 0.5% of positive slides were incorrectly classified as negative, owing to small, isolated tumor cell clusters (<110 μm). Figure 4 shows some examples of lymph node metastases from internal cohort 1, with Figure 4A, B showing correctly detected metastases and Figure 4C-E showing micrometastases that went undetected by the model. Moreover, approximately 3.3% of negative slides were falsely detected as positive, mainly owing to tissue folds and active germinal centers.

A subgroup of validation cases from internal cohort 2 consisted of mucinous adenocarcinoma and signet-ring cell carcinoma cases. In this group, the performance was slightly worse because of limited examples of mucinous adenocarcinoma and signet-ring cell carcinoma in the data sets used for training both the segmentation and metastasis detection models. Owing to the presence of this subgroup, the performance in internal cohort 2 yielded an F1, sensitivity, and specificity of 0.901, 0.872, and 0.936, respectively. A few examples from internal cohort 2 with overlay predictions are shown in Figure 5A, B.

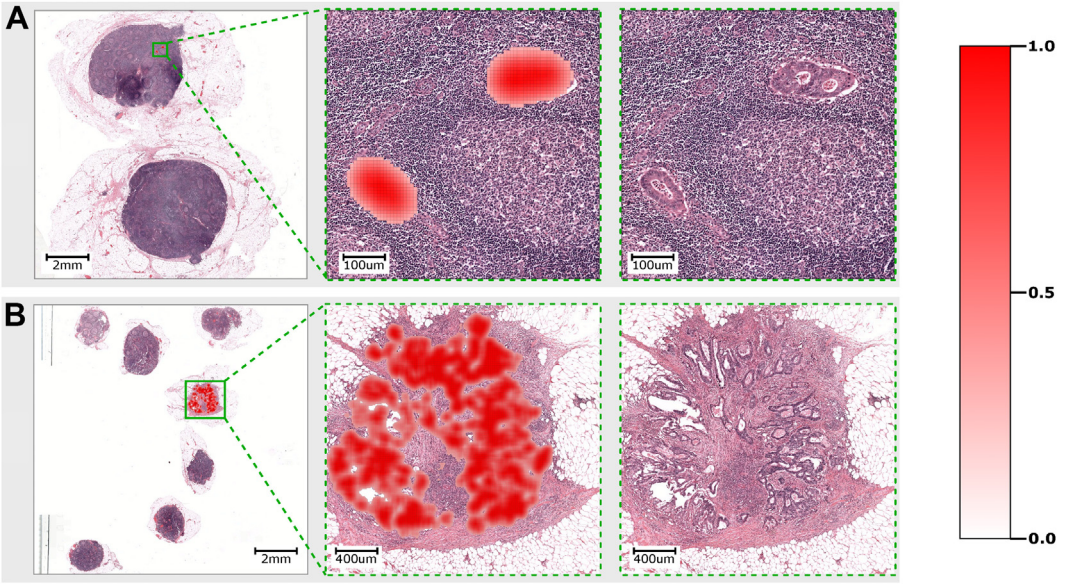

Figure 3.
Examples from the external cohort. The detected metastatic lymph node and overlays of predicted regions (with a tile probability between 0 and 1; see the color-map legend on the right), showing (A) a detected micrometastasis and (B) a tumor deposit.

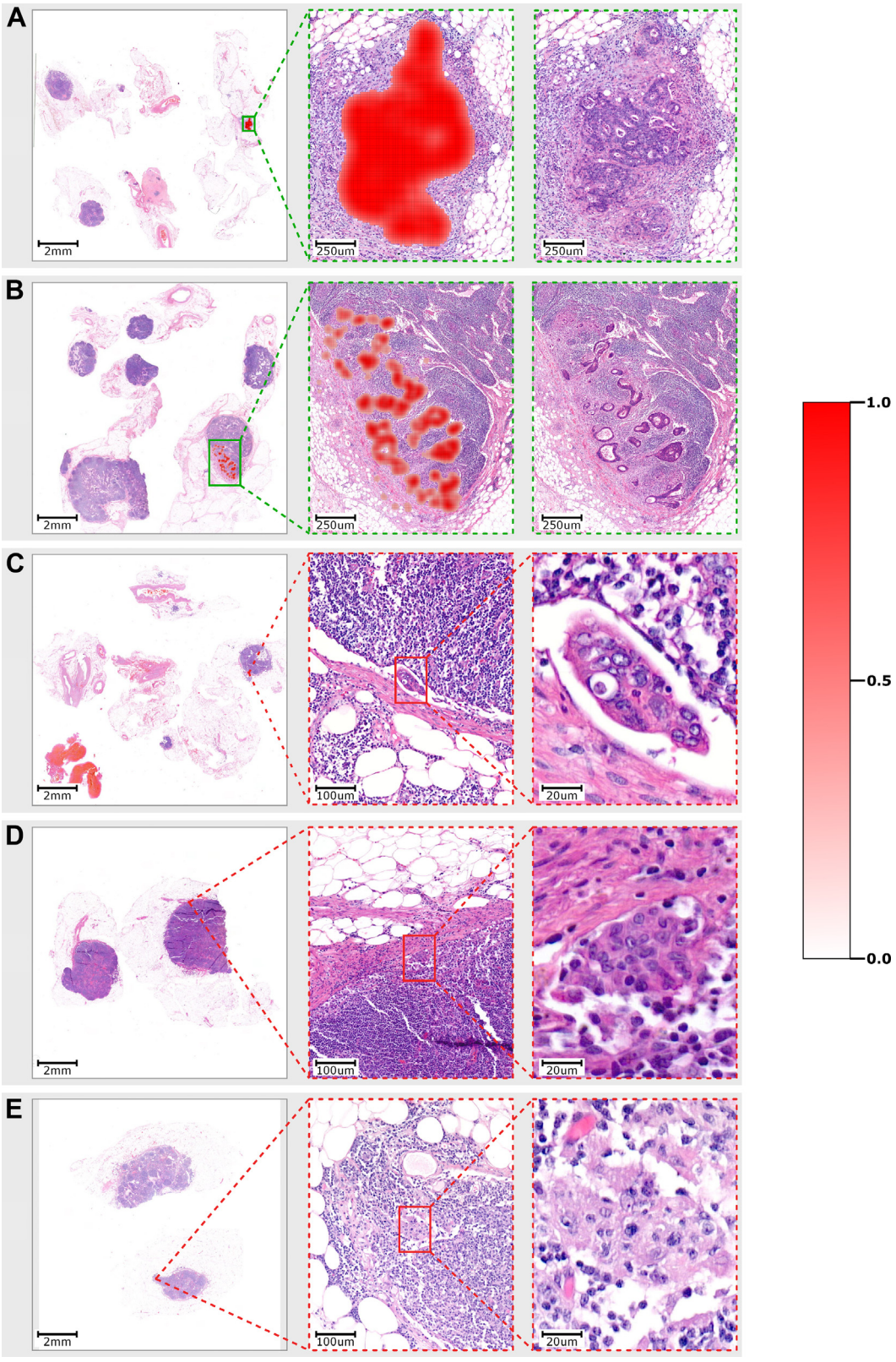

Figure 4.
Examples from internal cohort 1. The lymph node whole-slide images (WSIs) of patients with colorectal cancer (CRC) showing the heat map overlays (with a tile probability between 0 and 1; see the color-map legend on the right) of metastasis regions (A and B) detected by our ensemble model. The model missed 3 cases (C, D, and E) with a maximum metastasis of <110 μm in diameter.

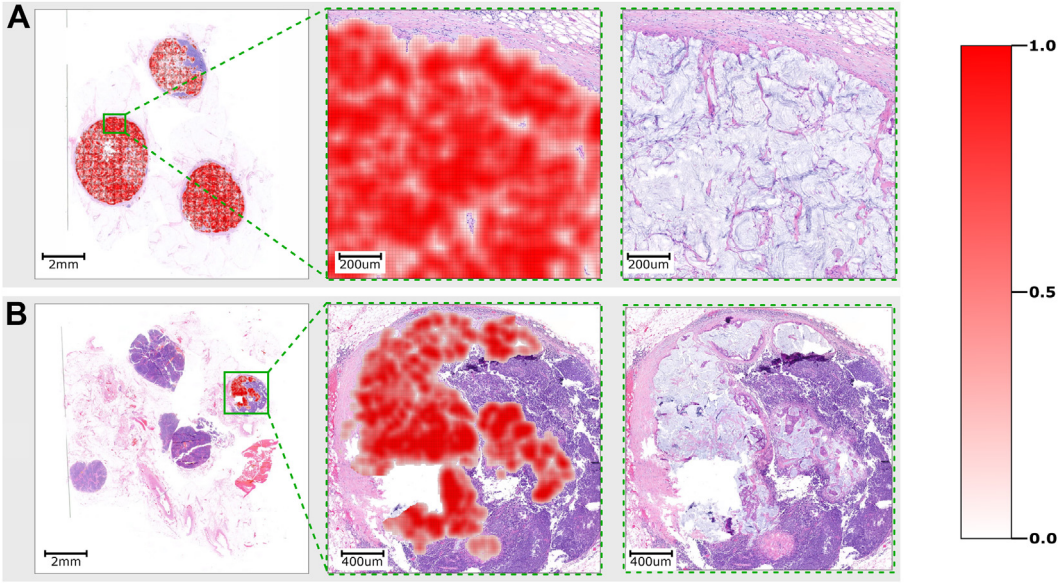

Figure 5.
Examples from internal cohort 2. The detected metastatic lymph node and overlays of predicted regions (with a tile probability between 0 and 1; see the color-map legend on the right) with detected mucinous adenocarcinoma regions (A and B).

Validation on Consecutive Case Series

To determine the performance of our approach in a simulated real-life setting, we scanned consecutive cases diagnosed between January 2022 and June 2022, comprising an additional independent internal cohort, internal cohort 3, without any selection for histologic subtypes. We applied our approach to all 217 CRC lymph node WSIs and generated overlays of the metastatic regions in QuPath. To validate our findings, the same glass slides were scored by a pathologist using a microscope. In parallel, another expert verified the model's predictions with the pathologist's outcome and reported any discrepancies. In this experiment, the lymph node status assigned by the model was a 100% match to the outcome of the pathologist. A few examples of metastatic detection from internal cohort 3 are presented in Figure 6A, B. Our approach was able to detect additional useful information for pathologists on a few WSIs, such as those where tumor deposits and metastatic fragments from the primary colon appeared on the slides along with lymph nodes (Fig. 6C). Similarly, in rare cases, such as sinus histiocytosis (ie, a benign disorder of lymph nodes), despite resembling tumor cell morphology, was correctly detected as normal tissue (Fig. 6D). Furthermore, a lymph node with cutting artifacts was correctly detected by the model as negative (Fig. 6E).

Discussion

In this study, we used segmentation, transfer learning, and ensemble methods to generate a workflow for a lymph node metastasis detection algorithm with high accuracy across multiple validation sets from 2 different institutions.

Deep learning models for histopathology images are typically trained on small tiles extracted from larger WSIs or annotated regions owing to computational memory constraints. Such small tile–based approaches have successfully been performed for various tasks at the cost of extensive regional annotations.[5,6,38-41] Previous studies have shown that such methods for lymph node metastasis detection in breast cancer, using 399 regionally annotated slides for model development, performed better (AUROC = 0.885) than pathologists (AUROC = 0.808) with a time constraint.[5] Similarly, the application of such a method to clinical experiments has shown significant improvement in sensitivity, from 83% to 91% (P =.02), when used as an assisting tool.[3] The multiple instance learning method, which requires slide-level labels to train the model, has previously achieved an AUROC of 0.966 when trained on 6,500 slides from axillary lymph nodes.[6] End-to-end WSI-based methods have shown improved performance over previous methods (with AUROCs of 0.959 and 0.941 for adenocarcinoma and squamous cell carcinoma) when trained on lung cancer on 5,045 slides.[42] A similar method has succeeded in obtaining even higher scores (AUROC of 0.999) when trained and tested on 1963 and 1000 CRC lymph node slides, respectively.[7]

To reduce the workload of collecting regional annotations, we developed a computer-assisted diagnostic workflow based on segmentation, slide-level classification, and visualization of potential metastatic lymph node regions for patients with CRC by leveraging a public data set of breast cancer lymph nodes. Our study had the following advantages: by using transfer learning and ensemble models, we developed a workflow for accurate detection of metastatic CRC lymph nodes with less effort in collecting regional annotations and without a large stack of globally labeled WSI data sets. Our workflow has shown excellent performance with very high sensitivity (0.995, 0.872, 1.0, and 1.0) and specificity (0.967, 0.936, 1.0, and 1.0) on 3 internal and 1 external validation cohorts (n = 4225 slides) when comparing slide-level labels with the pathologist ground truth. Moreover, regarding performance, our approach was in line with other similar approaches used for lymph node metastasis detection in different contexts, requiring larger regional or global annotated data sets for development.[3,5-7,42-44] Additionally, as our workflow is validated on large independent cohorts with higher sensitivity and specificity scores, our approach could significantly reduce the workload of pathologists for N-staging of patients with CRC.

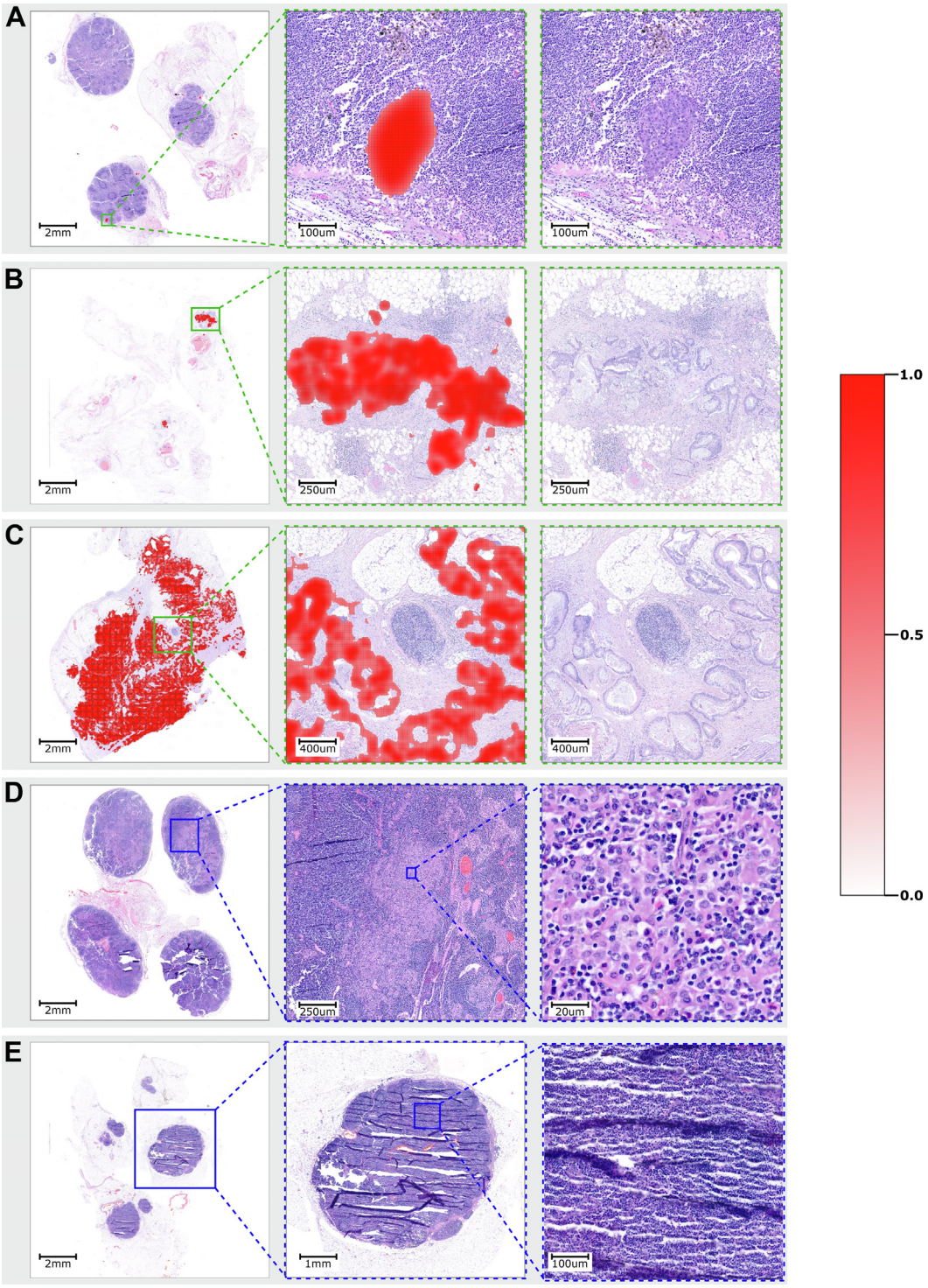

Figure 6.
Examples from internal cohort 3. The lymph node whole-slide image of samples from patients with colorectal cancer showing the heat map overlays (with a tile probability between 0 and 1; see the color-map legend on the right) of metastases regions. (A) A lymph node metastasis of around 400 μm in diameter is detected accurately. (B) An example of a tumor deposit that is detected with positive results. (C) A case with a negatively detected lymph node surrounded by colon cancer tissue. A sinus histiocytosis (ie, a benign disorder of lymph nodes) having a very distinct morphology similar to tumor cells in hematoxylin and eosin staining, and a lymph node with cutting artifacts (D and E, respectively). In both cases, the model correctly assigned the lymph node tissue as negative.

Despite an excellent performance, our study had some limitations. The training and validation data sets mostly contained adenocarcinoma cases (no special type) and only a few cases of other histologic subtypes. Nonetheless, our workflow showed high accuracy in cases with mucinous and signet-ring cell histology. In addition, to simulate the real-life diagnostic slide quality setting, no slide was excluded up front. There were 2 different types of artifacts considered: scanning artifacts (which were

eliminated at the source after the original scanning step and were considered to be of insufficient quality for diagnostics) and histologic artifacts (eg, folds, stain, and ink), which were still included. Therefore, the internal validation cohorts were selected without any quality control (for histologic artifacts, such as folds, stain, and ink) on the scanned WSIs, which resulted in a higher false-positive rate caused by tissue folds (~70% of all false-positive cases) and other artifacts (~30% of all false-positive cases). For clinical workflow implementation, some form of quality control could be automated to avoid false positives. Tools, such as HistoQC,[45] may be implemented to request rescans and ensure that this pipeline is robust to such problems. Similarly, in the HE-stained slides, cells from the reactive germinal centers may be confused with tumor cells and could contribute to higher false-positive rates. Furthermore, the external validation cohort lacked negative cases and was highly imbalanced. It was also observed that this cohort contained mostly larger metastasis regions, thus making it easier to outperform. However, we complemented this cohort with a consecutive series of unselected cases and achieved perfect agreement between the pathologist report and AI output. Extensive validation of diverse CRC lymph node data sets from different centers would be essential to assess the generalizability of the approach.

In addition, to address the question of isolated tumor cells, we performed immunohistochemistry staining on a series of apparently node-negative cases (n = 272 slides) and found 15 slides with isolated tumor cells. Moreover, these could not be detected on the corresponding HE-stained slides by our approach. Finally, our approach was limited to a classification task, rather than a segmentation task. This further restricts accurate measurement of the diameter of metastatic regions, which is useful for evaluating the detected lymph node metastases. Therefore, in the future, it would be interesting to obtain an accurate measurement of the detected regions and compare this with immunohistochemistry reference standards. Moreover, although our models are intended for HE-stained slides, the overlaid probability maps could help in deciding whether further immunohistochemistry staining is required. An interesting premise for future work is the automatic "labeling" of cases for subsequent immunohistochemistry based on these maps.

In conclusion, our lymph node metastasis detection approach is reproducible, sensitive, and specific; provides a visualization of the predicted region; and saves computational resources. This provides an excellent basis for future implementation in routine histopathology workflows.

Acknowledgments

The authors thank Ana Frei Leni, Elias Baumann, Dr Philipp Zens, Mauro Gwerder, Linda Studer, Christian Abbet, Dr Andreas Fischer, and Dr Behzad Bozorgtabar for valuable feedback during this work; Lize Dekkers, Dr Irene Centeno Ramos, Samuel Kuhn, Loredana-Ionela Daminescu, Carmen Cardozo, Dr José A. Galvan, Therese Waldburger, Stefan Reinhard, and Julius Babendererde for staining, scanning, and technical assistance; and the UBELIX (http://www.id.unibe.ch/hpc) team, the HPC cluster at the University of Bern, for providing efficient services during this study.

Author Contributions

I.Z. and A.K. conceptualized the study. A.K. and I.Z. conceived the methodology. A.K. designed the software. A.K., A.L., and I.Z validated the software. A.K., I.Z., N.B, and A.L. performed the formal analysis. A.K., N.B., A.L., A.B., H.D., F.M., and I.Z. performed the investigation. F.M., A.L., A.B, H.D., D.S., A.N., E.G., S.B., I.Z., N.B., I.N., and J.P.T. were responsible for resources. A.K., F.M., I.Z., N.B., I.N., D.S., A.N., E.G., S.B., and A.B. curated the data. A.K. wrote the original draft. I.Z., N.B., A.B., A.L., A.P., and I.N. reviewed and edited the manuscript. A.K., A.L, A.P, and I.Z. visualized the study. I.Z. and J.P.T. supervised the study. I.Z. administrated the project. I.Z. acquired the funding. All authors read and approved the final version of the paper.

Data Availability

The data used and/or analyzed during this study are available from the corresponding author on reasonable request.

Funding

This study was supported by the Swiss Cancer Research Foundation (KFS-4427-02-2018).

Declaration of Competing Interest

None of the authors have any conflicts of interest to report.

Ethics Approval and Consent to Participate

The study was approved by the ethics committee of the Canton of Bern under project number b2021-00033 and was performed according to the Human Research Act HFG 2014.

Computational Resources and Software

The experiments in this study were conducted on the UBELIX High-Performance Computing (HPC) cluster at the University of Bern. We used an Nvidia Geforce RTX3090 graphics processing unit, with 24 GB of GDDR6X memory. Training and inference of the models, reading of the whole-slide images, and statistical analysis and coding were performed in TensorFlow 2.4.1, OpenSlide 3.4.1, and Python 3.8.6, respectively.[46-48] The visualization of overlays was performed in QuPath 0.3.0.[13]

References

1. Amin MB, Edge SB, Greene FL, et al, American Joint Committee on Cancer (AJCC). *AJCC Cancer Staging Manual*. AJCC; 2017:211–212.
2. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27(5):775–784. https://doi.org/10.1038/s41591-021-01343-4
3. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018;42(12):1636–1646. https://doi.org/10.1097/PAS.0000000000001151
4. Huang S-C, Chen C-C, Lan J, et al. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. *Nat Commun*. 2022;13(1):1–14. https://doi.org/10.1038/s41467-022-30746-1
5. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–2210. https://doi.org/10.1001/jama.2017.14585
6. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309. https://doi.org/10.1038/s41591-019-0508-1

7. Chuang W-Y, Chen C-C, Yu W-H, et al. Identification of nodal micrometastasis in colorectal cancer using deep learning on annotation-free whole-slide images. *Mod Pathol*. 2021;34(10):1901–1911. https://doi.org/10.1038/s41379-021-00838-2
8. Veeling BS, Linmans J, Winkens J, Cohen T, Welling M. Rotation equivariant CNNs for digital pathology. In: Bertino E, Gao W, Steffan B, et al., eds. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); vol 11071 LNCS*. Springer; 2018:210–218. https://doi.org/10.1007/978-3-030-00934-2_24
9. Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience*. 2018;7(6):giy065. https://doi.org/10.1093/gigascience/giy065
10. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Bertino E, Gao W, Steffan B, et al., eds. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); vol 9351*. Springer; 2015:234–241. https://doi.org/10.1007/978-3-319-24574-4_28
11. Chollet F. Xception: deep learning with depthwise separable convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2017:1800–1807. https://doi.org/10.1109/CVPR.2017.195
12. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR 2021)*. ICLR; 2021. Accessed June 29, 2022. http://arxiv.org/abs/2010.11929
13. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1), 16878. https://doi.org/10.1038/s41598-017-17204-5
14. Khan A, Janowczyk A, Müller F, et al. Impact of scanner variability on lymph node segmentation in computational pathology. *J Pathol Inform*. 2022;13, 100127. https://doi.org/10.1016/j.jpi.2022.100127
15. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9(1):62–66. https://doi.org/10.1109/TSMC.1979.4310076
16. Marquez-Neila P, Baumela L, Alvarez L. A morphological approach to curvature-based evolution of curves and surfaces. *IEEE Trans Pattern Anal Mach Intell*. 2014;36(1):2–17. https://doi.org/10.1109/TPAMI.2013.106
17. Chan TF, Vese LA. Active contours without edges. *IEEE Trans Image Process*. 2001;10(2):266–277. https://doi.org/10.1109/83.902291
18. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE; 2009:1107–1110. https://doi.org/10.1109/ISBI.2009.5193250
19. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. *Anal Quant Cytopathol Histopathol*. 2001;23(4):291–299.
20. Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans Med Imaging*. 2016;35(8):1962–1971. https://doi.org/10.1109/TMI.2016.2529665
21. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Comput Graph Appl*. 2001;21(4):34–41. https://doi.org/10.1109/38.946629
22. Scannell CM, Chiribiri A, Veta M. Domain-adversarial learning for multi-centre, multi-vendor, and multi-disease cardiac MR image segmentation. In: Bertino E, Gao W, Steffan B, et al., eds. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 12592 LNCS. Springer; 2020:228–237. https://doi.org/10.1007/978-3-030-68107-4_23/TABLES/1
23. Chang J-R, Wu M-S, Yu W-H, et al. Stain mix-up: unsupervised domain generalization for histopathology images. In: Bertino E, Gao W, Steffan B, et al., eds. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); vol 12903 LNCS*. Springer; 2021:117–126. https://doi.org/10.1007/978-3-030-87199-4_11
24. Jung A.B. imgaug. Accessed February 24, 2022. https://github.com/aleju/imgaug
25. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal*. 2019;58, 101544. https://doi.org/10.1016/j.media.2019.101544
26. Byfield P. StainTools Library. Accessed February 24, 2022. https://github.com/Peter554/StainTools
27. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. *Information*. 2020;11(2):125. https://doi.org/10.3390/info11020125
28. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings*. ICLR; 2014. Accessed July 24, 2020. http://arxiv.org/abs/1412.6980
29. Loshchilov I, Hutter F. Decoupled weight decay regularization. 7th International Conference on Learning Representations, ICLR 2019. ICLR; 2019. https://doi.org/10.48550/arXiv.1711.05101
30. Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*. NeurIPS; 2012. https://doi.org/10.48550/arxiv.1210.5644
31. Mormont R, Geurts P, Maree R. Comparison of deep transfer learning strategies for digital pathology. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE; 2018:2343. https://doi.org/10.1109/CVPRW.2018.00303
32. Sharma Y, Ehsan L, Syed S, Brown DE. HistoTransfer: understanding transfer learning for histopathology. June 2021. Accessed October 24, 2022. http://arxiv.org/abs/2106.07068
33. Tsuneki M, Abe M, Kanavati F. A deep learning model for prostate adenocarcinoma classification in needle biopsy whole-slide images using transfer learning. *Diagnostics*. 2022;12(3):768. https://doi.org/10.3390/diagnostics12030768
34. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6. https://doi.org/10.1186/S12864-019-6413-7
35. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta - Protein Struct*. 1975;405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9
36. Taha AA, Hanbury A. An efficient algorithm for calculating the exact Hausdorff distance. *IEEE Trans Pattern Anal Mach Intell*. 2015;37(11):2153–2163. https://doi.org/10.1109/TPAMI.2015.2408351
37. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull*. 1945;1(6):80. https://doi.org/10.2307/3001968
38. Zhang W, Zhu C, Liu J, Wang Y, Jin M. Cancer metastasis detection through multiple spatial context network. In: *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*. ACM; 2019:221–225. https://doi.org/10.1145/3373509.3373567
39. Davri A, Birbas E, Kanavos T, et al. Deep learning on histopathological images for colorectal cancer diagnosis: a systematic review. *Diagnostics*. 2022;12(4):837. https://doi.org/10.3390/diagnostics12040837
40. Liu Y, Li F, Yu H, Zhang Z, Li H, Han C. A novel screening framework for lymph node metastasis in colorectal cancer based on deep learning approaches. In: *2022 7th International Conference on Multimedia and Image Processing; vol 7*. ACM; 2022:28–34. https://doi.org/10.1145/3517077.3517082
41. Liu Y, Kohlberger T, Norouzi M, et al. Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. *Arch Pathol Lab Med*. 2019;143(7):859–868. https://doi.org/10.5858/ARPA.2018-0147-OA
42. Chen CL, Chen CC, Yu WH, et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat Commun*. 2021;12(1):1193. https://doi.org/10.1038/s41467-021-21467-y
43. 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(1):1637. https://doi.org/10.1038/s41467-021-21674-7
44. Yu G, Sun K, Xu C, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. *Nat Commun*. 2021;12(1):6311. https://doi.org/10.1038/s41467-021-26643-8
45. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin Cancer Inform*. 2019;3(3):1–7. https://doi.org/10.1200/cci.18.00157
46. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: a vendor-neutral software foundation for digital pathology. *J Pathol Inform*. 2013;4(1):27. https://doi.org/10.4103/2153-3539.119005
47. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. March 2016. Accessed September 27, 2022. http://arxiv.org/abs/1603.04467
48. Van Rossum G, Drake FL. *Python 3 Reference Manual*. CreateSpace; 2009.