Journal of Pathology Informatics 14 (2023) 100174

Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

Accurate diagnostic tissue segmentation and concurrent disease subtyping with small datasets

Steven J. Frank *

*Med*A-Eye Technologies, Framingham, MA 01702, United States*

ARTICLE INFO

Keywords:
Digital pathology
Tissue segmentation
Deep learning
Whole slide images

ABSTRACT

Purpose: To provide a flexible, end-to-end platform for visually distinguishing diseased from undiseased tissue in a medical image, in particular pathology slides, and classifying diseased regions by subtype. Highly accurate results are obtained using small training datasets and reduced-scale source images that can be easily shared.

Approach: An ensemble of lightweight convolutional neural networks (CNNs) is trained on different subsets of images derived from a relatively small number of annotated whole-slide histopathology images (WSIs). The WSIs are first reduced in scale in a manner that preserves anatomic features critical to analysis while also facilitating convenient handling and storage. The segmentation and subtyping tasks are performed sequentially on the reduced-scale images using the same basic workflow: generating and sifting tiles from the image, then classifying each tile with an ensemble of appropriately trained CNNs. For segmentation, the CNN predictions are combined using a function to favor a selected similarity metric, and a mask or map for a a candidate image is produced from tiles whose combined predictions exceed a decision boundary. For subtyping, the resulting mask is applied to the candidate image, and new tiles are derived from the unoccluded regions. These are classified by the subtyping CNNs to produce an overall subtype prediction.

Results and conclusion: This approach was applied successfully to two very different datasets of large WSIs, one (PAIP2020) involving multiple subtypes of colorectal cancer and the other (CAMELYON16) single-type breast cancer metastases. Scored using standard similarity metrics, the segmentations outperformed more complex models typifying the state of the art.

1. Introduction and overview

1.1. Problem statement and related work

Whole-slide imaging platforms allow glass biopsy slides to be scanned and digitized at high resolution.[1,2] The resulting WSIs preserve minute anatomic detail but are quite large, typically exceeding 100,000 pixels in each dimension, making them cumbersome to review and store and difficult to share.[3–5] Substantial strides have been made in automating analysis of WSIs in order to assist clinicians in making diagnostic classifications.[6–8] Identifying and labeling diagnostic regions within a medical image represents a separate, and more difficult, computational task known as segmentation. Although often pursued alongside classification tasks such as subtyping, segmentation is far more granular and therefore more challenging.

CNNs have been used to segment images, including medical images of tissue, into distinct labeled regions.[9] They have been applied to "patch-wise" techniques that analyze small regions surrounding each pixel[10,11] and "fully convolutional" approaches that make predictions for all pixels at once.[12,13] The U-Net architecture,[14] developed expressly for biomedical tissue segmentation, builds on the fully convolutional architecture and is now routinely used,[15] particularly when combined with other architectures as discussed below.

These approaches usually process the entire image to be segmented and, as such, are subject to the size constraints affecting CNNs generally.[16,17] Commonly used CNNs running on standard hardware can comfortably handle image dimensions up to 1000 × 1000 pixels; larger images may require more complex architectures that are difficult to train, perform slowly, and require significant memory resources. While conventional CNN capabilities suffice for low-resolution images such as mammograms and chest radiographs, no CNN can process more than a minuscule portion of a histology WSI at a resolution sufficient to retain key anatomic detail. Analysis of a representative portion of an image may suffice if a region of interest can be localized in advance; patch-wise techniques, for example, can be applied to discrete image regions.[18] But it is unsuited to segmentation of large images in which subtle disease patterns may be present at unknown locations, if at all.[19]

* Corresponding author.
E-mail address: steve@medaeye.com.

http://dx.doi.org/10.1016/j.jpi.2022.100174
Received 15 July 2022; Received in revised form 1 September 2022; Accepted 15 December 2022
Available online 23 December 2022

Most tissue-segmentation tools have been developed to partition organs, sub-organs, and different classes of tissue rather than to separate diseased from undiseased regions, and typically operate on an entire image.[20–23] While helpful to clinicians, such capabilities are unlikely to reduce rates of diagnostic error involving medical images. These error rates have been estimated at 3% to 5%, resulting in approximately 40 million diagnostic errors involving medical images annually worldwide.[24] Fatigue-related errors in radiology, for example, have been well documented, with rates of retrospectively detected errors estimated to be as high as 30%.[25]

1.2. Contributions

The objective of this work is to provide a flexible platform for segmentation of diseased from undiseased tissue in a medical image — particularly large medical images such as WSIs, which are among the most challenging to analyze. The approach is equally applicable to smaller (e.g., radiology or magnetic resonance) images. Our primary emphasis is on decision support: helping clinicians locate elusive disease regions that might otherwise be missed due to their subtlety or small size, or because of simple fatigue or error. How artificial intelligence (AI) compares to human performance, on the other hand, is not considered.

The techniques we describe offer several benefits relative to other AI platforms. First, high accuracy levels are obtained even with small training sets. The expense of generating training images annotated by disease experts has long plagued development of AI-based diagnostic systems, so much so that substantial efforts have been made to augment the supply using synthetic data[26,27] and semi-supervised learning techniques.[28] The ability to use small training sets, if carefully curated to avoid underrepresentation of diverse patient populations, can speed development and broaden the range of clinical decision-support tools.

Second, both the images and the CNNs that process them are small. Our approach uses only as much image resolution as is necessary to resolve the anatomy critical to segmentation and, if concurrently implemented, disease subtyping. As a result, analyzed images are generally small enough to be transmitted conveniently and displayed, both in original form and as tissue segmentations, on a mobile device. The CNNs are lightweight and may, if desired, also be executed on a mobile device, enabling system deployment as an "edge AI" application requiring no connectivity.

Third, the same basic workflow is used both for segmentation and subtyping, enabling "end-to-end" processing of a candidate image in a single pass. The workflow involves generating subimage tiles from a downsampled version of the candidate image (for segmentation) or from the predicted tumor region (for subtyping, following segmentation); generating tile-level classification probabilities; and combining these in accordance with a selection framework to produce a pixel-level segmentation mask and a subtype classification. Although image resolution and tile sizes may differ between segmentation and subtyping tasks, this does not alter the processing flow or time to completion.

Novel aspects of the proposed technique that proved critical to success with modestly sized datasets include measures for computationally identifying and excluding tiles unlikely to contain important visual information. One such measure uses image entropy not only to identify visually rich tiles, as is now conventional, but to segregate tissue types; this confines CNN analysis to predictive tiles drawn from the most relevant image regions. Another key aspect involves combining tile-level predictions from multiple models to make best use of limited training images. The models are trained on overlapping but distinct image subsets. Combining their tile-level predictions reduces random error associated with the individual CNN models,[29] and the manner of combining predictions can be selected so as to favor a particular similarity metric.

2. Materials and methods

2.1. Methodology and workflow

The approach described here is summarized in Fig. 1. WSIs are reduced in scale, and the rescaled images are preprocessed and broken down into overlapping tiles for analysis by a CNN. The degree of downsampling and the tile size into which the downsampled image is decomposed represent parameters specific to the tissue under study as well as the task being performed. Optimized together, the rescaled image and the resulting tiles ideally retain only as much anatomic detail as is necessary to facilitate classification of tissue as diseased or undiseased in order to produce a

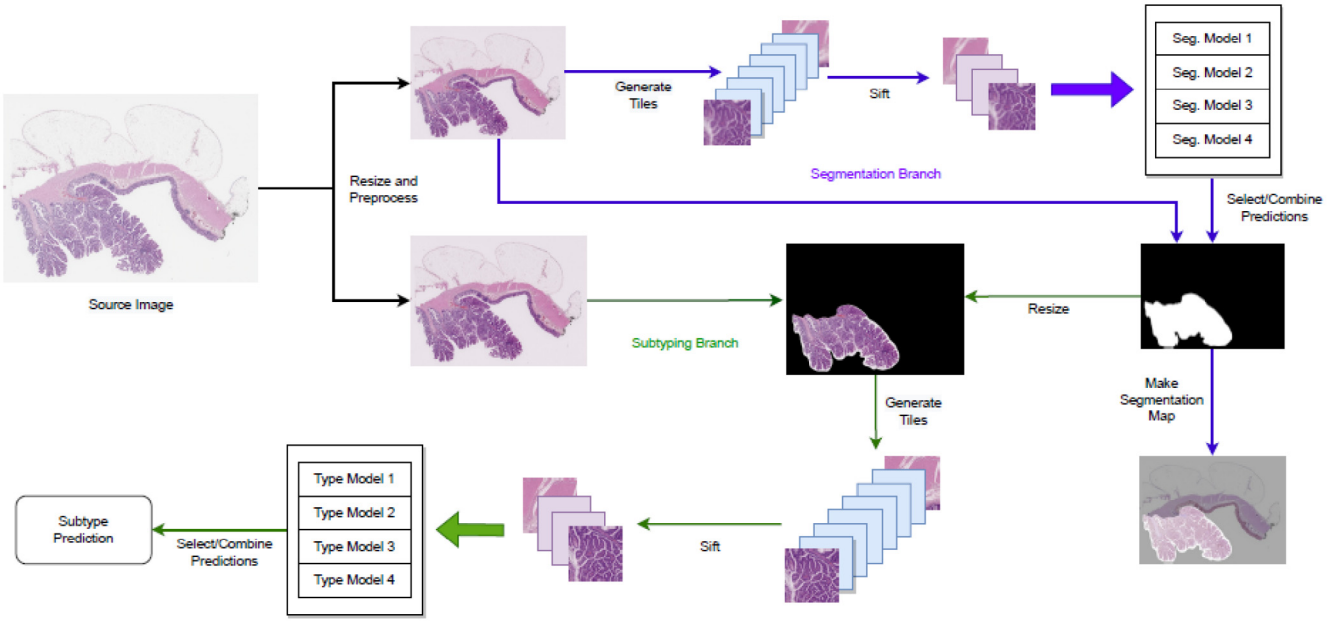


Fig. 1. A representative workflow including subtyping has two parallel branches. First, the WSI is rescaled to image sizes optimized for segmentation and subtyping. In the segmentation branch, tiles are generated from the segmentation image, sifted, and presented to CNN models that have been trained on different dataset folds to produce a segmentation mask. In the subtyping branch, the mask is resized to match the subtyping image and used to exclude regions unlikely to be diseased. Tiles are generated from the unmasked regions of the subtyping image, sifted, and presented to CNN models trained for subtyping to produce a subtype prediction.

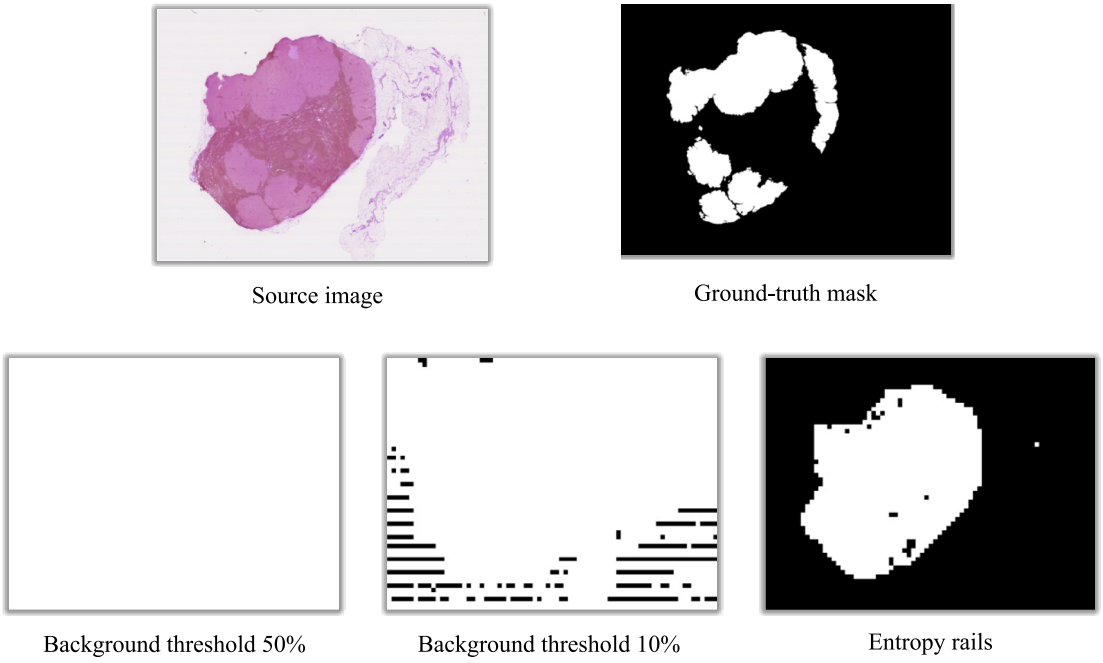


Fig. 2. Effect of sifting with entropy rails. A source histology image of metastatic lymph node tissue from the CAMELYON16 dataset is associated with a ground-truth segmentation mask, which, when applied to the image, occludes non-tumor tissue regions. Sifting with background thresholding alone is not a viable strategy. Setting the maximum per-tile background fraction at 50% excludes no tiles, so a location map of the retained tiles is completely white. Sifting with a much smaller background maximum of 10% still captures far too much non-tumor tissue for the image to be useful. The horizontal bands in the 10% image track subtle stripe artifacts in the source image, further degrading image usefulness. The union of tiles sifted with entropy rails exceeds, but roughly approximates, the tumor region, and the band artifacts are eliminated.

segmentation. If subtyping is to be performed, the same procedure is repeated on image regions within the segmentation to resolve the subtypes. As shown in Fig. 1, both the degree of image downsampling and the tile size may differ between segmentation and subtyping.

For segmentation, two sets of images are created: one with just the diseased regions masked and the other with everything except diseased regions masked. Training tiles are then created from each image set. The tiles overlap sufficiently so their total, after sifting as described below, is adequate for training — generally at least 10,000 per class per cross-validation fold, and in equal numbers. The data redundancy resulting from overlap, it is found, matters less than tile population. For subtyping, tiles generated exclusively from diseased regions may be segregated into subtype classes. While disease detection is inherently a binary classification problem, subtyping may involve discrimination among multiple disease types.

Successful implementation depends on optimal tile sizing and selection. Larger tiles provide more anatomic information to the CNN for analysis. But tile size also dictates the resolution of the segmentation and CNN complexity. Hence, the above procedure is repeated for different image scales (typically different levels associated with the source WSIs) and different tile sizes at each scale.

Before performance can be compared at different image resolutions and tile sizes, however, tiles must be sifted to eliminate those having low or irrelevant image content. It is important to be able locate, computationally, all tiles corresponding to the diseased region while excluding as many non-qualifying tiles as possible. In particular, tiles with too much background or glass-slide artifacts degrade training. A commonly used technique for background exclusion is Otsu's method.[30] Because this technique attempts to identify a single threshold, however, it is suboptimal for complex images in which background and foreground intensity overlap, as may occur for samples that undergo histological staining.[31] Even the seemingly simple problem of excluding glass-slide artifacts has engendered complex solutions involving, for example, CNNs.[32]

Here, the task of excluding spurious tiles is performed in three stages: simple background exclusion, background exclusion for stained tiles, and entropy sifting. Staging these progressively more complex techniques allows nonqualifying tiles to be identified with the least amount of processing. Simple background exclusion converts the tile to grayscale and discards it if an excessive fraction (e.g., >0.2) of the tile is pure black or pure white.[33] For slides that have been treated with hematoxylin and eosin ("H&E") stain and normalized as described below, however, this approach may fail due to the tonal shift. A second stage of background exclusion therefore thresholds tiles based on the number of identically valued pixels rather than their proximity to white or black extremes. In particular, we obtain the frequencies and populations of pixel values within a tile, and reject the tile if the total population of the highest few pixel frequencies exceeds a threshold fraction corresponding to the allowable proportion of background area. Tiles with less background area are more visually heterogeneous, with smaller pixel populations at the peak value frequencies.

These two stages are sufficient when generating training tiles, which are drawn from annotated regions. Tiles derived from an image under examination are subjected to a further stage of sifting based on image entropy, which reflects the degree of nonredundant information — the information diversity — in a region of pixels. Whereas earlier work[33] tests tile entropy against a threshold minimum, we have found that different tissue types may exhibit image entropies falling within a characteristic band. Therefore, when training tiles are generated, the minimum and maximum entropy values of tiles drawn from the diagnostic regions are noted. These values serve as boundaries or "rails" that constrain selection of tiles from a candidate image to be segmented: a candidate tile is retained only if its image entropy lies on or within the rails. This test not only ensures that the tile contains sufficiently diverse visual information to support classification by the CNN but also excludes clearly irrelevant tissue types.

Fig. 2 illustrates the benefits of tile sifting with entropy rails. Due in part to staining, conventional sifting measures based on background fraction fail to exclude a significant fraction of tiles drawn from irrelevant image regions. Sifting with entropy rails, by contrast, excludes most of the background and obviously non-diagnostic tissue. The narrower the entropy range manifested by diseased tissue, the more successful entropy sifting will be in excluding non-diagnostic tissue from consideration. Of course,

proper CNN training ensures that only a relatively small proportion of tiles will be misclassified in any case; but in this study, sifting with entropy rails improved tile-level classification accuracy by about 10%.

The CNN architecture primarily employed in this study[1] was selected to minimize the number of convolutional layers (five are used) and consequent trainable parameter count. Three dropout layers mitigate the risk of overfitting to a small dataset of images. For both datasets studied and for both segmentation and subtyping, we trained for 75 epochs in each training/test partition using a batch size of 16, a categorical cross-entropy loss function, an Adam optimizer, a learning rate of 0.0001, and sigmoid activation. Data augmentation consisted of random horizontal and vertical flips as well as brightness variation. More significant data augmentation resulted from the degree of tile overlap. Model weights were saved following each epoch and, for each training subset, the weights producing the smallest binary cross-entropy loss were retained. The model was trained and evaluated on an Nvidia GeForce RTX 2070 GPU; due to the simplicity of the model architecture, each epoch completed in less than two minutes.

Following CNN training, a new WSI may be segmented by first downsampling to the resolution identified as optimal and decomposing the resampled image into overlapping tiles whose size matches the training tiles. Once classified, these tiles are used to generate a segmentation mask or segmented image. Pixel-level diagnostic probabilities are computed by averaging, for each pixel, the tile-level prediction probabilities of tiles containing that pixel.

As a candidate image is evaluated, the tiles are sifted and presented to each of the subset-trained CNNs. The manner in which tile-level predictions are chosen or combined favors, for segmentation, a selected similarity metric. The most common metric, intersection over union ("IoU"), quantifies the degree of overlap between the prediction P and the ground truth T:

$$\text{IoU} = \frac{P \cap T}{P \cup T}$$

Precision represents the proportion of pixels classified as positive (e.g., as tumor pixels) that are, in fact, positive while recall (or sensitivity) corresponds to the proportion of all positive pixels correctly classified as such. In terms of true positives (*TP*), false positives (*FP*), and false negatives (*FN*),

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Averaging predictions across models maximizes IoU, yielding segmentations that balance precision and recall. Selecting the maximum prediction for each tile favors recall. A high recall score ensures that most or all of the diseased region is made visible in the segmentation. Selecting the minimum prediction for each tile favors precision. Depending on characteristics of the tissue as well as the imaging modality, the difference may be pronounced or relatively minor, as shown in Fig. 3.

For subtyping, the image is resampled from the WSI at the resolution optimal for this classification task. The segmentation mask is resized if necessary and applied to the subtyping image to exclude regions predicted to be non-diagnostic; classification tiles are generated only from the diagnostic region. In this case, masks optimized for precision are utilized since, for subtyping purposes, including potentially misleading non-diagnostic tiles is more harmful to prediction accuracy than excluding some diagnostic tiles (which are unlikely to skew toward any subtype). As will be seen, the CNN models used to classify the image and those used for segmentation may have different architectures.

[1] The architecture is illustrated and described in detail in earlier work.[33] Source code is available at https://github.com/stevenjayfrank/A-Eye.

2.2. Datasets

Two very different benchmark datasets were utilized in this study. Both consist of extremely large WSIs, which would be impossible to segment in raw form using conventional deep-learning techniques. The annotated training and validation slides of the PAIP2020 challenge[2] were selected to investigate segmentation of multiple disease subtypes. We employed the CAMELYON16 dataset to investigate segmentation of a single disease type, cancer metastases in lymph nodes, based on samples having very small diagnostic regions.

2.2.1. The PAIP2020 Dataset

The PAIP2020 dataset contains annotated WSIs exhibiting different degrees of microsatellite instability (MSI), a molecular phenotype of colorectal cancer that arises from a defective DNA mismatch repair system. MSI status in colorectal cancer has prognostic and therapeutic implications. In particular, a high degree of MSI (MSI-H) is associated with a better prognosis than a low degree (MSI-L). Moreover, MSI-H appears to predict the efficacy of immune checkpoint inhibitors in solid tumors.[34] Although CNNs have been successfully used to identify MSI status in gastrointestinal cancer patients,[35,36] distinguishing between MSI-L and MSI-H subtypes is far more difficult, typically requiring DNA testing.[37,38]

The PAIP2020 training image dataset consists of 47 whole slides — 12 of which are labeled as MSI-H and the remaining 35 as MSI-L — provided in multilevel SVS format. The levels and their corresponding image parameters are shown in Table 1.

The slides contain varying amounts of non-tissue background. The dataset also includes binary segmentation masks defining the tumor regions. An unannotated, unlabeled validation set consists of 31 additional slides.

For training, the WSIs at scaling levels 2 and 3 (L2 and L3) were investigated. The provided ground-truth masks were rescaled to each level and used to create new, separate images of the tumor and non-tumor portions of each image. Four different subsets or folds of the 47 training slides were defined. Each fold included 36 training images (8 MSI-H images and 28 MSI-L images) and 11 test images (4 MSI-H images and 7 MSI-L images) to preserve, in each fold, a training/test split above 3:1. Each of the MSI-H test sets was unique, i.e., contained no images found in any other test set. Because segmenting tumor regions does not require distinguishing between MSI-H and MSI-L tumor types — that is., both types can be considered a single class — this imbalance can be tolerated as long as the training sets contain similar distributions of both tumor types.

The typically small image area occupied by the tumor represents a further source of class imbalance, i.e., between tumor and non-tumor tissue. To address this discrepancy, which is far more significant in the CAMELYON16 dataset, we obtained similar numbers of tumor and non-tumor tiles by overlapping them to different degrees. Overlap levels ranged from 80% to 96% depending on the image size and the number of images in each training class. While this approach does not address the underlying imbalance between tumor and non-tumor data, training based on equivalent numbers of tiles did not impair the ability to obtain useful predictions.

The L2 and L3 images were stain-normalized using Reinhard normalization[39,40] and decomposed into tiles ranging in size from 200 × 200 to 600 × 600 pixels. This range of tile sizes accords with recent work identifying a similar range as producing peak CNN performance for anatomic subject matter including tumor tissue.[41] Training tiles were sifted using the two-stage background exclusion procedure described above, with tiles having more than 20% background excluded. The background of tumor images — MSI-H and MSI-L images were

[2] De-identified pathology images and annotations used in this research were prepared and provided by the Seoul National University Hospital by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316). The PAIP 2020 datasets are provided by the Seoul National University Hospital, South Korea. See https://paip2020.grand-challenge.org.

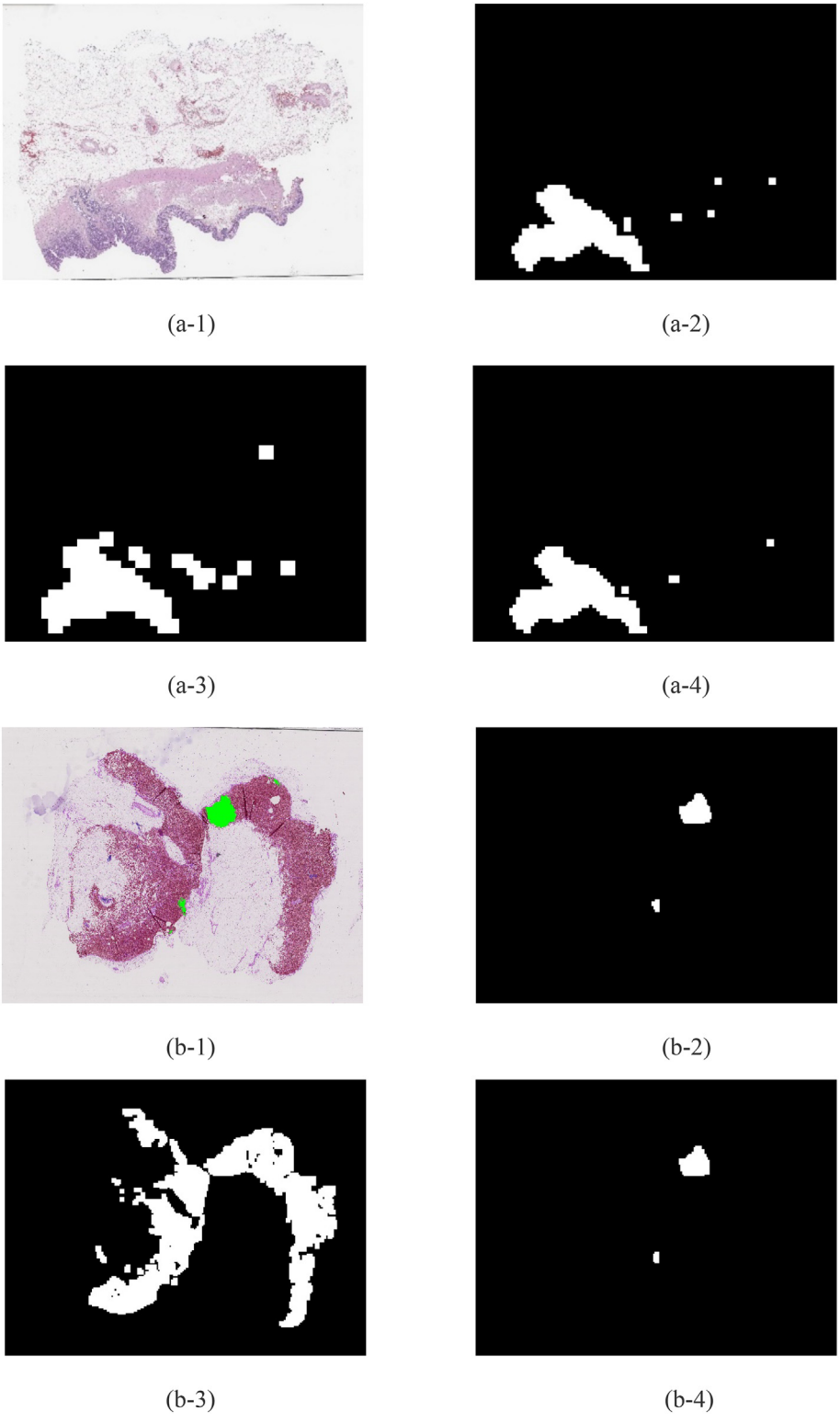


Fig. 3. Similarity-metric optimization. (a-1) Validation image from PAIP2020 dataset. (a-2) Segmentation mask for image shown in (a-1), optimized for IoU. (a-3) Segmentation mask for image shown in (a-1), optimized for recall. (a-4) Segmentation mask for image shown in (a-1), optimized for precision. (b-1) Validation image from CAMELYON16 dataset with tumor region marked in green (ground-truth annotations are not available for PAIP2020 validation images). (b-2) Segmentation mask for image shown in (b-1), optimized for IoU. (b-3) Segmentation mask for image shown in (b-1), optimized for recall. (b-4) Segmentation mask for image shown in (b-1), optimized for precision. For the PAIP2020 image, optimizing for different similarity metrics has little overall effect on the resulting masks. The opposite is true for the CAMELYON16 image: optimizing for precision produces positive mask regions confined to the primary tumor sites, while optimizing for recall results in a diagnostically unhelpful mask that excludes very little tissue.

Table 1
Scaling factors for PAIP2020 SVS files.

Level	Average Dimensions (pixels)	Downsample Factor	Magnification
0	116,214 × 88,094	1	40×
1	29,053 × 22,023	4	20×
2	7263 × 5505	16	10×
3	3498 × 2662	32	5×

preprocessed identically — consisted of a solid black background surrounding the tumor regions, while non-tumor images included black regions corresponding to the tumor locations with the remainder of the slide unmodified. After sifting, subsets of tiles corresponding to the image folds were drawn from this superset to create segmentation training sets. For each training set at each tile size, the tumor and non-tumor class populations each exceeded 12,500 tiles.

For each tile size, the maximum and minimum image entropies of qualifying tumor tiles (with no distinction drawn between MSI-L or MSI-H tiles) were noted. Segmentation test tiles were prepared by decomposing each test image into overlapping tiles. This time, the tiles were sifted based not only on background content but image entropy as well. In particular, tiles with above-threshold background regions or whose image entropies were not on or between the entropy rails were excluded.

To determine the optimal image rescaling and tile size for segmentation, training and testing were carried out with tiles of varying size corresponding to a single fold of the L2 and L3 images. In each case, model weights were saved after each of the 75 training epochs and the weights producing the smallest binary cross-entropy loss for each were compared across tile sizes; the smallest loss corresponded to the optimal tile size.

For subtyping, tile sets were assembled from the tumor portions of the labeled images, with greater tile overlap for the MSI-H images to achieve class-level parity in tile populations. Although this creates greater data redundancy in the MSI-H tiles, once again the class imbalance did not preclude accurate predictions and produced better results, particularly with adjustment to the decision boundary, than alternatives such as efforts to calibrate the models.[42] The above procedures using a single image fold were repeated to identify the optimal image rescaling and tile size for subtype classification.

Once obtained, these parameters were used to generate tiles for the remaining image folds. The best-performing segmentation and subtyping model weights were identified and saved for each image fold. These models were then tested against tiles derived from the validation images using the same procedures — image rescaling, tile generation, and exclusion based on background content and entropy — described above. Predictions were generated for each tile by all four best-performing fold models. Comparative segmentation masks were created by varying the prediction actually assigned to each tile: the average of all four fold-level model predictions, the minimum prediction or the maximum prediction. The assigned tile-level prediction was stored against each image pixel location spanned by the tile; the more tiles that intercepted a pixel, the larger the number of predictions that were associated with that pixel. A pixel-level probability map corresponding to the image under study was created by averaging, for each pixel, the assigned prediction probabilities. The resulting binary segmentation masks are white where the averaged probabilities equal or exceed the decision boundary and are otherwise black.

For subtyping, the segmentation masks having the highest precision scores were applied to the validation images. The first stage of background exclusion, which rejects tiles having too much black or white, constrains tile retention to tiles drawn from tumor regions of the images. These were further sifted as described above and presented to all four best-performing fold models. The prediction probabilities from the four fold models were averaged for each tile, and the resulting tile-level predictions averaged to produce an overall image-level classification. Pixel-level probabilities, in other words, were not generated for this task.

Table 2
Scaling factors for CAMELYON16 TIFF files.

Level	Average Dimensions (pixels)	Downsample Factor	Magnification
0	106,754 × 190,790	1	40×
1	53,377 × 95,395	2	30×
2	26,688 × 47,697	4	20×
3	13,344 × 23,848	8	15×
4	6672 × 11,924	16	10×

2.2.2. The CAMELYON16 dataset

Metastatic involvement of lymph nodes corresponds to a poorer prognosis for survival of breast cancer and, like microsatellite instability in colorectal cancer, is difficult and time-consuming to diagnose from visual examination of histopathology images. In the CAMELYON16 dataset, a sample is either normal, i.e., the lymph node contains no metastatic tissue, or malignant. The CAMELYON16 dataset is especially challenging in that diagnostic regions may be quite small — in some cases just a few pixels out of billions, and may not be organized into a discrete mass. Whereas the tumor regions in the PAIP2020 dataset tend to be large and contiguous, the CAMELYON16 lymph-node lesions often present as a dusting of tiny features. The latter morphology tested the lower limits of useful tile sizes.

The CAMELYON2016 dataset consists of WSIs provided in multilevel TIFF format. The dataset includes several hundred slides, 111 of which have annotations prepared by expert pathologists and which define metastatic regions. Scaling and magnification factors appear in Table 2.

The tumor morphologies necessitated a much larger rescaled image size. In order to be classified properly, the tile must contain enough diagnostic information to permit the CNN to distinguish reliably among classes. At the same time, for the contours of a tile-based segmentation to exhibit reasonable fidelity to the represented tumor region, the tile size must be smaller (ideally, considerably smaller) than that region. We tested the L1, L2, and L3 images, as well as a set rescaled to a maximum dimension of 15,000 pixels (i.e., intermediate between L3 and L4).

Of the 111 annotated tumor-containing WSIs, 91 were selected for training and the remaining 20 served as the validation set. Tiles were prepared at different sizes (ranging from 100 × 100 to 250 × 250 pixels) for the tumor and non-tumor portions of each image as described above.

3. Results

3.1. PAIP2020 slides

The best segmentation performance for the PAIP2020 validation set was obtained with L2 images at a tile size of 150 × 150 pixels.

As shown in Table 3, relatively high scores were obtained for all similarity metrics. Segmentation masks employed to generate tiles for subtyping benefitted from the highest possible precision, since non-tumor tiles are potentially confounding while completeness is unnecessary; only enough tiles to support classification are needed. For clinical purposes, a case may be made for IoU or recall. For safety, i.e., to ensure that nothing significant is missed, recall is the critical metric; the IoU score that accompanies maximum recall is sufficiently high that false-positive regions are unlikely to pose a significant distraction. On the other hand, the recall level of 0.92 that accompanies maximum precision is still quite high; any missed regions will very likely be noticed by the clinician and are probably cumulative.

Table 3
Segmentation scores for L2 validation images with 150 × 150 pixel tiles.

	PAIP2020 Validation Slides – Segmentation		
	Mean IoU Score	Mean Precision	Mean Recall
Averaged predictions	**0.77**	0.84	0.92
Min. predictions	0.76	**0.90**	0.86
Max. predictions	0.71	0.74	**0.96**

Table 4
Classification accuracies for L2 validation images with 200 × 200 pixel tiles

	PAIP2020 Validation Slides – Classification		
	Five-layer CNN	EfficientNetB0	ResNet50
Accuracy	0.81	0.94	0.81

The L2 images also produced the best subtype classification performance, with identical results using tiles of 200 × 200 or 400 × 400 pixels. We repeated the testing procedure with two commonly used but more complex architectures, EfficientNet and ResNet50. As shown in Table 4, EfficientNetB5, with an input image size of 456 × 456, delivered the best performance. An advantage to the proposed framework is the ease with which different CNN architectures with varying input image sizes can be interchanged. The classification accuracy achieved with EfficientNetB5 in this study is clinically promising.

3.2. CAMELYON16 slides

As earlier noted, the highly variable morphology of the metastatic lesions in the CAMELYON16 dataset makes well-targeted detection, with minimal misclassification of normal tissue, difficult. Although tile size represents the atomic unit of resolution, both for detection and representation, the minimum achievable tile size is limited by the anatomy and visually manifested disease features. The rescaled image size, on the other hand, dictates its ease of handling (e.g., to store in cache memory), transmission and computational processing — e.g., the ability to analyze results or even perform the entire processing sequence on a mobile device. The IoU, precision, and recall similarity metrics establish the minimum usable image resolution and tile size, and therefore the smallest tumor feature that will appear in the segmentation. Halving the degree of downscaling doubles the required tile size, since ultimately the same critical area of tissue anatomy must be analyzed by the CNN. For the CAMELYON16 dataset, the greatest usable degree of rescaling limited the largest image dimension to 15,000 pixels, i.e, between L3 and L4 dimensions, with a tile size of 200 × 200 pixels. Averaging predictions from the lowest-loss fold models maximized the IoU score, producing the results shown in Table 5.

While the five-layer CNN outperformed ResNet50, both models failed to identify any tumor regions in three of the images. The lesions in at least one and arguably two of these images, both discussed in greater detail below, have dimensions putting them below the micrometastasis threshold. The third image, however, contained a micrometastasis that a clinically useful digital pathology system should not have missed. A better approach is needed, and one is proposed in sec. 3.4.

3.3. Comparison with architectures based on U-Net

For tissue segmentation, the current state of the art is typified by CNN architectures based on U-Net, often modified by using another architecture such as EfficientNet as the encoder stage (also known as the "backbone").[44–46] We compared the performance of our five-layer CNN to this architecture for both the PAIP2020 and CAMELYON16 datasets. U-Net systems are trained with segmentation datasets consisting of input images and ground-truth segmentation masks. The CNN learns to map the pixels of an image to the training masks, which may have several levels, each corresponding to a different segmentation class. Here the segmentation task is binary; the CNN maps pixels to true or false values denoting whether the tissue is or is not tumor. The U-Net architecture with an EfficientNet B0 backbone that we employed has 271 layers and 10 million trainable parameters.

Table 5
Segmentation scores for CAMELYON16 validation images resized to a maximum dimension of 15,000 pixels with 200 × 200 pixel tiles, and comparison to ResNet50, which performed best with 224 × 224 pixel tiles

	CAMELYON16 Validation Slides – Segmentation		
	Mean IoU Score	Mean Precision	Mean Recall
Five-layer model (200×200 pixel tiles)	0.21	0.22	0.69
ResNet50 (224×224 pixel tiles)	0.12	0.14	0.61

Table 6
Performance comparison between our five-layer model and hybrid U-Net architecture on the same tiles derived, in the case of PAIP2020, from a single training fold; and for CAMELYON16, from the entire training set. A single fold was used in the PAIP2020 comparison because ground-truth masks are not available for the validation images.

	PAIP2020 Slides (256 × 256 tiles)			CAMELYON16 Slides (224 × 224 tiles)		
	Mean IoU	Mean Precision	Mean Recall	Mean IoU	Mean Precision	Mean Recall
5-layer model	0.789	0.895	0.877	0.25	0.29	0.64
U-Net/EfficientNet	0.471	0.504	0.937	0.071	0.074	0.696

We prepared segmentation sets from tiles derived from the downscaled WSI images described above. We found that the U-Net/EfficientNet hybrid architecture performs best on images that have been thresholded using Otsu's method[30] before tiling. Creating segmentation masks from the resulting tiles is complicated by the fact that, with the U-Net architecture, the prediction is itself a mask tile that usually contains noise. We obtained optimal performance by treating a prediction as positive over the entire tile if at least 70% of pixels were classified as tumor. For the PAIP2020 dataset, we trained the hybrid model using 256 × 256 pixel tiles derived from images in the second training fold and selected the best performer over 20 epochs, though overfitting became evident after fewer than 10 epochs. We trained our five-layer model on the same tiles.

For the CAMELYON16 dataset, we utilized 224 × 224 pixel tiles and trained over the entire training set rather than in folds, then tested against the validation images. While not advisable for models intended to generalize beyond the training and validation sets, this single-set strategy is well-suited to a comparative study. For the hybrid model, overfitting began immediately – by the second epoch even with a learning rate of 1×10^{-5}.

As shown in Table 6, the five-layer model outperformed the hybrid model over all metrics except recall, and even here the difference was quite small — particularly when compared to the large discrepancies in IoU and precision, the "price" of the hybrid model's high recall scores. Although the hybrid model produced regions of positive prediction for all test images, these were often the wrong regions. Visual inspection of the final prediction masks reveals that the hybrid model tended to classify most of the tissue regions of a slide as tumor — i.e., it learned to identify tissue rather than to discriminate between tumor and non-tumor tissue regions. A far greater proportion of the tissue in the PAIP2020 dataset is cancerous as compared with the CAMELYON16 dataset, so in the former case even a CNN trained merely to distinguish bulk tissue from the glass slide will deliver reasonable precision scores. The hybrid model may be better suited to identifying sharply defined tissue structures with clear contrast than subtle disease markers.

3.4. Rethinking visualization

In clinical practice, pathologists characterize tumor cell clusters 2 mm or larger as "macrometastases," smaller clusters between 0.2 mm and 2 mm as "micrometastases," and still smaller clusters (or single cells) as "isolated tumor cells." The largest available metastasis determines the slide-level diagnosis.[47]

This being the case, conventional similarity metrics, while important, may have less relevance to clinical practice than a "hit rate" — that is, whether the segmentation visibly identifies the diagnostically essential

tissue regions, even if not all diseased regions are detected or marked with perfect fidelity. In Figs. 3(b-2) and (b-4), for example, the segmentation masks do not include the two smallest tumor regions. But the largest organized feature in the small lesion at the top right of Fig. 3(b-1) has average dimensions of about 105 × 160 μm, smaller than a micrometastasis; and the lesion at the bottom left is actually an archipelago of tiny features, the largest of which is a mere 50 × 50 μm. Their omission from the segmentation should not impair its diagnostic value since the two largest features, the smaller one a micrometastasis, are marked.

In Fig. 4, our system's segmentation captures the densest regions of the tumor feature cluster rather than representing its full extent. But for visualization purposes, to make efficient use of review time, it may suffice even if the contour match is imperfect. On the other hand, missing tumor regions entirely (at least those meeting the micrometastasis threshold) will not. An unreliable AI tool will not be used. Nor will it be used if achieving high reliability comes at the cost of too much normal tissue misclassified as tumor (i.e., trading off precision to achieve high recall). A compromise strategy is to color-code regions of high and moderate interest based on probability levels, rather than segmenting in a purely binary fashion (as in the segmentation map of Fig. 1). This strategy draws initial attention to the regions where tumors are most likely present. If lesions exist, the largest will probably be found in the high-interest regions, in which case the primary review objective is attained. If none is found, the reviewing pathologist can move on to the regions of moderate interest.

The question is where to draw the probability lines so that all detected tumors are marked but lie primarily in the red regions. For the high-probability threshold, we used the mean probability of all tiles classified as tumor using the best fold models and the prediction-selection criterion chosen to maximize the similarity parameter of interest (e.g., recall). All pixels whose probabilities (averaged over all tiles containing the pixel) equal or exceed the high-probability threshold are colored red. The low-probability threshold is based on the variance of probabilities assigned to the tiles classified as tumor. Typically, the low threshold is one standard deviation below the high threshold, though it may be lower if a single standard deviation does not bring the low threshold below the decision boundary. All pixels whose averaged probabilities fall between the low and high thresholds are colored yellow.

Selecting the maximum predictions for 200 × 200 tiles in order to bias the classification toward finding elusive tumor regions — favoring reliability over noise minimization, in other words — results in high and low probability thresholds of 0.63 and 0.41, respectively, for the CAMELYON16 dataset. (Balancing reliability and noise by averaging predictions unsurprisingly produces a mean probability at the decision boundary of 0.5.) The results of these operations are summarized in Table 7 and illustrated in Fig. 5.

In 90% (18/20) of the validation images, at least some portion of the largest tumor region is colored — usually in red but, in a single case, in yellow only. If the hit rate represents successful tumor identification, the precision score reveals the amount of normal tissue misclassified as tumor and, consequently, wasted review time. Because of the small tumor features, however, even a low-precision probability map does not include very much noise. The price of accuracy is not terribly high even in the worst cases since most of the slide is (correctly) uncolored. Particularly if the segmentation is a translucent overlay on the original image or the clinician can easily toggle between them, the objective of drawing attention to the region of diagnostic interest without much distracting error is met.

Of the two slides with tumors that our system did not detect, one involved features smaller than a micrometastasis, with an average diameter of about 0.1 mm. The other presented long, narrow lesions with lengths up to 0.4 mm but averaging only about 0.04 mm in width. Such morphologies can elude tile-based detection by snaking through tiles without occupying enough area to trigger positive classification. Based on this small validation set, our five-layer CNN appears capable of detecting micrometastases so long as the definitional criteria are satisfied along both dimensions.

4. Discussion

While headlines frequently tout AI models whose performance matches or exceeds that of human experts, few such systems have achieved routine use in clinical workflows.[48–53] In 2018, Tizhoosh & Pantanowitz[53] listed ten key challenges to commercial and clinical acceptance of AI-driven digital pathology systems, chief among them the lack of labeled data. Clinical decision-support systems must demonstrate their utility to practitioners, and hence marketplace viability, as a prerequisite to investment backing; but annotated training sets depend on expensive expert time and effort.

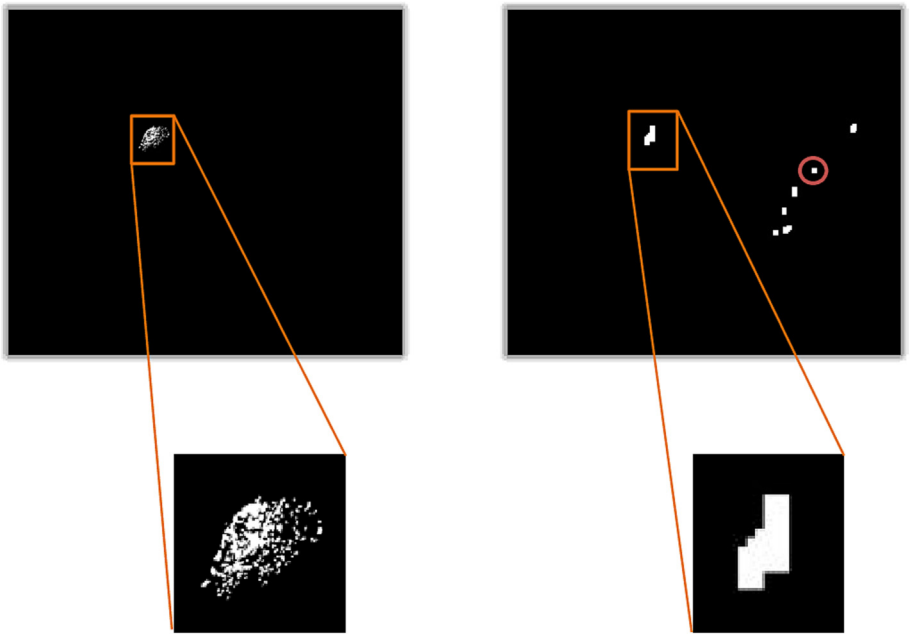
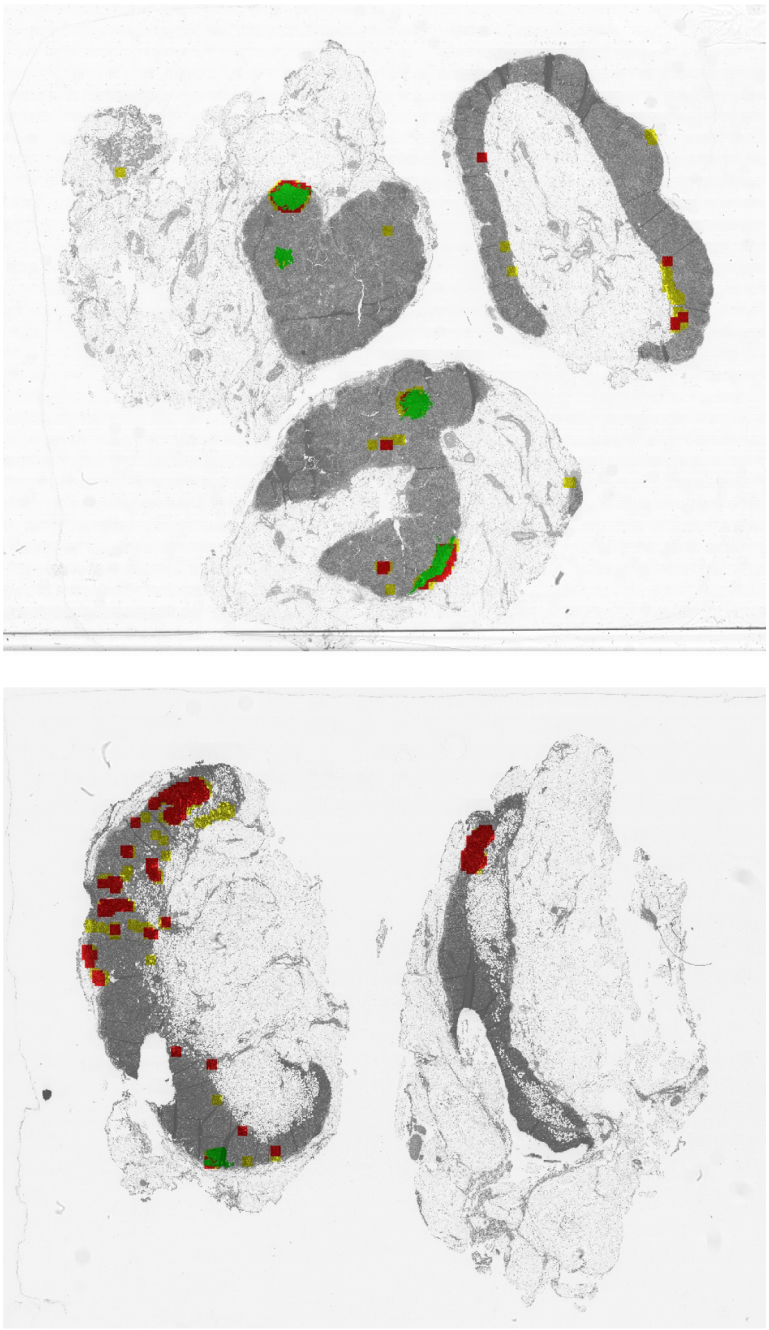


Fig. 4. Tumor features in representative CAMELYON16 image. Left, ground-truth segmentation mask with enlargement of tumor region, which consists of a cluster of features each only a few pixels in extent. Right, segmentation mask for the same image generated by averaging predictions of the best-performing fold models (i.e., maximizing IoU score). Only the denser regions of the tumor were recognized and segmented properly. The size of a tile — an example of which is circled — limits the ability both to detect and represent tumor features.

Table 7
Precision scores and hit rates for red regions (consisting of pixels with averaged probabilities of at least 0.63) and yellow regions (consisting of pixels with averaged probabilities between 0.41 and 0.63). The hit rate corresponds to the percentage of tumor-containing slides in which any portion of at least the largest tumor was identified.

CAMELYON16 Slides (200 × 200 tiles)					
	High, low thresholds	Mean precision (red)	Mean precision (red + yellow)	Hit rate (red)	Hit rate (red + yellow)
Best fold models, selection criterion = max	0.63, 0.41	0.24	0.18	80%	90%



(a)

(b)

Fig. 5. Representative high- and low-precision color-coded probability maps for CAMELYON16 slides. Slide (a) has a precision score of 0.35. The tumor regions, shown in green, are clearly marked by (or even fully engulfed in) red and/or yellow regions, and there is little spuriously colored non-tumor area. Slide (b) has a poor precision score of 0.03. The tumor is properly marked in red and yellow, and the erroneously colored regions, while more widespread than in (a), do not occupy a significant proportion of the image.

The work described here is intended to address this "chicken and egg" problem by facilitating development of clinically useful systems using small numbers of training images. Once deployed, ongoing use can itself provide a source of further training data to improve performance and broaden the diversity of the underlying patient pool.

Other enduring challenges include the "dimensionality obstacle" of large images; the futility of attempting to replace rather than assist human physicians; and the practical needs of those physicians: ease of use, financial return, and trust. The present work focuses squarely on these factors as well. Providing ease of use while retaining trust means minimizing false markings without compromising the reliability of tumor identification in a segmentation, or at least striking a clinically acceptable balance between these competing priorities. Ease of use is also supported by downscaled images that are conveniently handled and shared. If relatively small biopsy images can be communicated instantly to disease experts who can analyze them on mobile devices, expert resources can be assembled *ad hoc*, and a virtual team need not pore over the same image display shoulder-to-shoulder as is common today.[48] Moreover, the mere ability to examine pathology images on a mobile device, with regions of interest labeled, allows for intuitive gestural image manipulation (e.g., stretching and squeezing) unavailable on desktop workstations. Today, computational pathology applications are not widely used on mobile devices,[54] although some image-classification applications, such as skin lesion detection, have been proposed.[55] This is unsurprising in view of the dimensionality problem alone.

Ultimately, it is possible to envision complete end-to-end analysis of a scaled-down WSI on a mobile device. Tiling is a mechanically simple array operation that requires only rudimentary computational resources, and sifting is staged to reserve more complex operations for tiles that have cleared simpler screening criteria. Many mobile devices now include neural processing units that can execute CNNs with low computational overhead and low power consumption, albeit with some tradeoff against accuracy.[56] Indeed, highly efficient architectures such as MobileNet currently enable mobile devices to classify images,[57] and the five-layer model described here is considerably simpler than MobileNet. (MobileNet v2, for example, uses 2,422,081 parameters at a tile size of 200 × 200; the five-layer model uses 439,793 parameters at the same tile size.) Whether such capabilities actually prove useful in a world of ubiquitous connectivity, where lightweight mobile apps derive hefty computational assists from remote servers, is debatable at present. Far less debatable is the ongoing need to address the challenges that continue to limit adoption of promising AI-driven pathology tools for diagnostic support.

Declaration of Competing Interest

There are no financial or personal relationships with other people or organizations that could inappropriately influence or bias this work.

References

1. Pantanowitz L, Farahani N, Parwani A. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol Lab Med Int 2015. https://doi.org/10.2147/plmi.s59826.
2. Rączkowski Ł, Możejko M, Zambonelli J, Szczurek E. ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. Sci Rep 2019. https://doi.org/10.1038/s41598-019-50587-1.
3. Lopez Barron DE, Rao P, Rao D, Tawfik O, Zachariah A. *Large-scale storage of whole slide images and fast retrieval of tiles using DRAM, in*. 2020. https://doi.org/10.1117/12.2564694.
4. Helin H, Tolonen T, Ylinen O, Tolonen P, Näpänkangas J, Isola J. Optimized JPEG 2000 compression for efficient storage of histopathological whole-Slide images. J Pathol Inform 2018;9. https://doi.org/10.4103/jpi.jpi_69_17.
5. Lujan GM, Savage J, Shana'ah A, et al. Digital pathology initiatives and experience of a large academic institution during the coronavirus disease 2019 (COVID-19) pandemic. Arch Pathol Lab Med 2021;145. https://doi.org/10.5858/arpa.2020-0715-SA.
6. Aeffner F, Zarella MD, Buchbinder N, et al. Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association. J Pathol Inform 2019;10. https://doi.org/10.4103/jpi.jpi_82_18.
7. Khened M, Kori A, Rajkumar H, Krishnamurthi G, Srinivasan B. A generalized deep learning framework for whole-slide image segmentation and analysis. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-90444-8.
8. Dimitriou N, Arandjelović O, Caie PD. Deep Learning for Whole Slide Image Analysis: An Overview. Front Med 2019;6. https://doi.org/10.3389/fmed.2019.00264.
9. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. J Digit Imaging 2017. https://doi.org/10.1007/s10278-017-9983-4.
10. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with Deep Neural Networks. Med Image Anal 2017. https://doi.org/10.1016/j.media.2016.05.004.
11. Zhang W, Li R, Deng H, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. Neuroimage 2015. https://doi.org/10.1016/j.neuroimage.2014.12.061.
12. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit; 2015. https://doi.org/10.1109/CVPR.2015.7298965.
13. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. Proc IEEE Int Conf Comput Vis 2015. https://doi.org/10.1109/ICCV.2015.178.
14. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science., in: Med. Image Comput. Comput. Interv. – MICCAI 2015MICCAI 2015. Lect. Notes Comput. Sci. 2015.
15. Liu L, Cheng J, Quan Q, Wu FX, Wang YP, Wang J. A survey on U-shaped networks in medical image segmentations. Neurocomputing 2020. https://doi.org/10.1016/j.neucom.2020.05.070.
16. Dong H, Yang G, Liu F, Mo Y, Guo Y. Automatic brain tumor detection and segmentation using U-net based fully convolutional networks. Commun Comput Inf Sci 2017. https://doi.org/10.1007/978-3-319-60964-5_44.
17. Iglovikov V, Rakhlin A, Kalinin A, Shvets A. Pediatric bone age assessment using deep convolutional neural networks. BioRxiv 2017. https://doi.org/10.1101/234120.
18. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016. https://doi.org/10.1016/j.neucom.2016.01.034.
19. Kleczek P, Jaworek-Korjakowska J, Gorgon M. A novel method for tissue segmentation in high-resolution H&E-stained histopathological whole-slide images. Comput Med Imaging Graph 2020. https://doi.org/10.1016/j.compmedimag.2019.101686.
20. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. Phys Medica 2021;85. https://doi.org/10.1016/j.ejmp.2021.05.003.
21. Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustain 2021;13. https://doi.org/10.3390/su13031224.
22. Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev 2021;54. https://doi.org/10.1007/s10462-020-09854-1.
23. Schick F. Tissue segmentation: a crucial tool for quantitative MRI and visualization of anatomical structures. Magn Reson Mater Physics, Biol Med 2016;29:89–93. https://doi.org/10.1007/s10334-016-0549-0.
24. Itri JN, Tappouni RR, Mceachern RO, Pesch AJ, Patel SH. PATIENT-CENTERED CARE fundamentals of diagnostic error in imaging. RadioGraphics 2018;38:1845–1865. https://doi.org/10.1148/rg.2018180021.
25. Stec N, Arje D, Moody AR, Krupinski EA, Tyrrell PN. A systematic review of fatigue in radiology: Is it a problem? Am J Roentgenol 2018;210. https://doi.org/10.2214/AJR.17.18613.
26. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. *Synthetic data in machine learning for medicine and healthcare*. 2022. https://doi.org/10.1038/s41551-021-00751-8.
27. Thambawitaid V, Amouei S, Hicks SA, et al. *SinGAN-Seg: Synthetic training data generation for medical image segmentation*. 2022. https://doi.org/10.1371/journal.pone.0267976.
28. Gadgil S, Endo M, Wen E, Ng AY, Rajpurkar P. CheXseg: combining expert annotations with DNN-generated saliency maps for X-ray segmentation. https://arxiv.org/abs/2102.10484v2 2021.accessed June 4, 2022.
29. Dietterich TG. Ensemble methods in machine learning. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics); 2000. https://doi.org/10.1007/3-540-45014-9_1.
30. Otsu N. Threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 1979;SMC-9. https://doi.org/10.1109/tsmc.1979.4310076.
31. Bora D Jyoti, Gupta AK. (PDF) a new efficient color image segmentation approach based on combination of histogram equalization with watershed algorithm. Int J Comput Sci Eng 2016:156–167. https://www.researchgate.net/publication/304733928_A_New_Efficient_Color_Image_Segmentation_Approach_Based_on_Combination_of_Histogram_Equalization_with_Watershed_Algorithm. (accessed June 9, 2022).
32. Bándi P, Balkenhol M, Van Ginneken B, Van Der Laak J, Litjens G. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. PeerJ 2019;2019. https://doi.org/10.7717/peerj.8242.
33. Frank SJ. Resource-frugal classification and analysis of pathology slides using image entropy. Biomed Signal Process Control 2021;66, 102388. https://doi.org/10.1016/j.bspc.2020.102388.
34. Sahin IH, Akce M, Alese O, et al. Immune checkpoint inhibitors for the treatment of MSI-H/MMR-D colorectal cancer and a perspective on resistance mechanisms. Br J Cancer 2019;121. https://doi.org/10.1038/s41416-019-0599-y.
35. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25. https://doi.org/10.1038/s41591-019-0462-y.
36. Artificial B, Hyun Park J, Young Kim E, et al. *Citation: park, J artificial intelligence for predicting microsatellite instability based on tumor histomorphology: a systematic review*. 2022. https://doi.org/10.3390/ijms23052462.

37. Evrard C, Tachon G, Randrian V, Karayan-Tapon L, Tougeron D. Microsatellite instability: Diagnosis, heterogeneity, discordance, and clinical impact in colorectal cancer. Cancers (Basel) 2019;11. https://doi.org/10.3390/cancers11101567.
38. Singh MP, Rai S, Pandey A, Singh NK, Srivastava S. Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine. Genes Dis 2021;8. https://doi.org/10.1016/j.gendis.2019.10.013.
39. Lakshmanan B, Anand S, Jenitha T. Stain removal through color normalization of haematoxylin and eosin images: A review. J Phys Conf Ser 2019. https://doi.org/10.1088/1742-6596/1362/1/012108.
40. Reinhard E, Ashikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Comput Graph Appl 2001;21. https://doi.org/10.1109/38.946629.
41. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. Radiol Artif Intell 2020;2. https://doi.org/10.1148/ryai.2019190015.
42. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc 2022. https://doi.org/10.1093/JAMIA/OCAC093.
43. Mathews MR, Anzar SM, Kalesh Krishnan R, Panthakkan A. EfficientNet for retinal blood vessel segmentation. 2020 3rd Int. Conf. Signal Process. Inf. Secur. ICSPIS; 2020. https://doi.org/10.1109/ICSPIS51252.2020.9340135.
44. Poudel S, Lee SW. Deep multi-scale attentional features for medical image segmentation. Appl Soft Comput 2021;109. https://doi.org/10.1016/j.asoc.2021.107445.
45. Hong LTT, Thanh NC, Long TQ. CRF-EfficientUNet: an improved UNet framework for polyp segmentation in colonoscopy images with combined asymmetric loss function and CRF-RNN Layer. IEEE Access 2021;9. https://doi.org/10.1109/ACCESS.2021.3129480.
46. 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 - J Am Med Assoc 2017;318:2199–2210. https://doi.org/10.1001/jama.2017.14585.
47. Homeyer A, Lotz J, Schwen L, et al. Artificial intelligence in pathology: From prototype to product. J Pathol Inform 2021;12. https://doi.org/10.4103/jpi.jpi_84_20.
48. Rakha EA, Toss M, Shiino S, et al. Current and future applications of artificial intelligence in pathology: A clinical perspective. J Clin Pathol 2021;74. https://doi.org/10.1136/jclinpath-2020-206908.
49. Acs B, Hartman J. Next generation pathology: artificial intelligence enhances histopathology practice. J Pathol 2020;250. https://doi.org/10.1002/path.5343.
50. Colling R, Pitman H, Oien K, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. J Pathol 2019;249. https://doi.org/10.1002/path.5310.
51. Matheny M, Israni ST, Ahmed M, Whicher D. *Artificial intelligence in health care: the hope, the hype, the promise, the peril.* 2019.
52. Reza Tizhoosh H, Pantanowitz L. Artificial intelligence and digital pathology: Challenges and opportunities. J Pathol Inform 2018;9. https://doi.org/10.4103/jpi.jpi_53_18.
53. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Investig 2021;101. https://doi.org/10.1038/s41374-020-00514-0.
54. Zouridakis G, Wadhawan T, Situ N, et al. Melanoma and other skin lesion detection using smart handheld devices. Methods Mol Biol 2015;1256. https://doi.org/10.1007/978-1-4939-2172-0_30.
55. Tan T, Cao G. *Efficient execution of deep neural networks on mobile devices with NPU*. 2021. https://doi.org/10.1145/3412382.3458272.
56. Morikawa C, Kobayashi M, Satoh M, et al. Image and video processing on mobile devices: a survey. Vis Comput 2021. https://doi.org/10.1007/s00371-021-02200-8.