Tissue Segmentation from Whole-Slide Images Using Lightweight Neural Networks

Steven Frank (stevenjayfrank@gmail.com)
Art Eye-D Associates LLC https://orcid.org/0000-0002-7905-7792

Article

Keywords: convolutional neural networks (CNNs), whole-slide images, tissue segmentation, lightweight neural networks

DOI: https://doi.org/10.21203/rs.3.rs-122564/v1

Abstract

Pathology slides of malignancies are segmented using lightweight convolutional neural networks (CNNs) that may be deployed on mobile devices. This is made possible by preprocessing candidate images to make CNN analysis tractable and also to exclude regions unlikely to be diagnostically relevant. In a training phase, labeled whole-slide histopathology images are first downsampled and decomposed into square tiles. Tiles corresponding to diseased regions are analyzed to determine boundary values of a visual criterion, image entropy. A lightweight CNN is then trained to distinguish tiles of diseased and non-diseased tissue, and if more than one disease type is present, to discriminate among these as well. A segmentation is generated by downsampling and tiling a candidate image, and retaining only those tiles with values of the visual criterion falling within the previously established extrema. The sifted tiles, which now exclude much of the non-diseased image content, are efficiently and accurately classified by the trained CNN. Tiles classified as diseased tissue – or in the case of multiple possible subtypes, as the dominant subtype in the tile set – are combined, either as a simple union or at a pixel level, to produce a segmentation mask or map. This approach was applied successfully to two very different datasets of large whole-slide images, one (PAIP2020) involving multiple subtypes of colorectal cancer and the other (CAMELYON16) single-type breast-cancer metastases. Scored using standard similarity metrics, the segmentations exhibited notably high recall, even when tiles were large relative to tumor features. With segmentations that can be generated locally and broadcast widely, efficiencies in utilizing expert resources can be achieved.

Introduction And Overview

Imaging modalities such as magnetic resonance imaging (MRI) produce clear, high-resolution tissue images that may be analyzed for the presence of disease or abnormality. Substantial strides have been made in automating this analysis in order to assist clinicians in making diagnostic classifications. Actually 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, 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 [1]. In particular, they have been applied to “patch-wise” techniques that analyze small regions surrounding each pixel [2,3] and “fully convolutional” approaches that make predictions for all pixels at once [4,5]. The U-Net architecture [6], developed expressly for biomedical tissue segmentation, builds on the fully convolutional architecture and is now routinely used [7]. These approaches generally process the entire image to be segmented and, as such, are subject to the size constraints affecting CNNs generally [8,9]. CNNs perform best at image sizes below 600 x 600 pixels; larger images entail complex architectures that are difficult to train, perform slowly, and require significant memory resources. Images this small can cover only a small anatomic region at a resolution sufficient to retain key detail. This may be sufficient for pedagogical purposes or if a region of interest can be localized in advance; patch-wise techniques, for example, can be applied to discrete image regions[11]. These limitations preclude CNN-based segmentation (as opposed to

classification) of large images in which subtle disease patterns may be present at unknown locations, if at all [12].

More generally, the subtleties distinguishing disease subtypes from each other and from undiseased tissue typically necessitate use of complex neural-network architectures. The original U-Net implementation had 23 convolutional layers [13], for example, and other deep-learning techniques that avoid hand-engineered image features — that is, which learn directly from labeled images — also involve architectures with 16 or more convolutional layers [14]. These approaches require substantial computational and memory capacity to process the large number of parameters (16 million in the case of U-net, for example [15]) inherent in their architectures. Although the computational capacity of mobile devices such as tablets and phones continues to grow, it is inevitably limited by battery life and the need to perform foreground tasks. Integrating complex deep-learning systems into clinical practices that continue to evolve toward telemedicine represents a substantial challenge.

The approach described here performs segmentation at the level of an image tile rather than a pixel, although pixel-level segmentations based on the tiles is also possible. This reduces segmentation to a tractable classification task that may be carried out using simple CNN architectures. As will be seen, even the largest medical images — whole slides of histopathology samples, which may have sizes exceeding 100,000 pixels in each dimension — can be handled.

Successful implementation depends on optimal tile sizing and selection. Larger tiles provide more anatomic information to the CNN for classification. But tile size also dictates the resolution of the segmentation and CNN complexity, as well as affecting tile selection. In particular, selection is based on a visual criterion that will locate all tiles corresponding to the diseased region while excluding as many non-qualifying tiles as possible. By presenting for CNN classification only those tiles that satisfy the visual criterion, spurious classifications are minimized and wasteful CNN processing is avoided.

As discussed in earlier work [16], image entropy reflects the degree of nonredundant information — the information diversity — in a region of pixels. For pure classification exercises, it is useful to compare the entropy of a tile with that of the larger image from which it is drawn. This approach is unsuited to segmentation analysis, however, because it does not directly discriminate between tissue types. Instead, using a training set of ground-truth slides annotated by pathology experts, tiles are derived from the diagnostic regions and their minimum and maximum entropy values noted. These values serve as boundaries or “rails” that constrain selection of tiles from a test image to be segmented: a candidate tile is retained only if its image entropy lies on or within the rails.

As shown in Fig. 1, the results of this initial sift can be striking.

As illustrated, tying the sifting criterion to the ground-truth training images instead of the image under examination, as is common in CNN-based segmentation systems, produced significantly better performance — but only if the visual criterion is well chosen. Tile-based classification and segmentation

systems commonly select tiles based on image density [17] or background percentage [18]. Even if distortions due to variations in staining and digital acquisition of slides can be overcome [12], all of the undiseased tissue regions will still be present in the resulting tiles, complicating the classification task.

The technique considered here may be applied to whole-slide images as follows. First, an annotated training set of images is obtained and downsampled. The degree of downsampling and the tile size into which the downsampled image is decomposed represent parameters specific to the tissue morphology under study. Optimized together, they permit the use of source images small enough to be processed on mobile devices and a tile size that balances segmentation accuracy with resolution. The training images are segregated into two sets, one with just the diseased regions and the other with disease regions masked. Training tiles are then created from each set. The tiles overlap sufficiently so their total, after initial sifting to remove tiles with excessive background area, is adequate for training — generally tens of thousands of tiles per class. The data redundancy resulting from overlap, it is found, matters less than tile population; for training purposes, quantity is quality.

If there is only one disease type, the classification task is binary. Particularly in studies of malignancy, however, multiple disease subtypes often exist and must be reckoned with. Even if the objective is simply to highlight disease regions of whatever subtype, combining diagnostic subtypes into a single training class is frequently a mistake; any visually significant differences among morphologies will likely skew classification and degrade accuracy. For example, one subtype may, to the CNN, appear less distinct from normal tissue than from the other subtype. The technique we describe here is suited to multiclass training with a softmax activation function and should be employed in this way when subtypes can be visually distinguished.

The minimum and maximum entropies of the diagnostic tiles, assessed over all disease types, are obtained. The nondiagnostic (i.e., normal tissue) tiles are then further sifted based on the entropy rails thus established. (Of course, in practice, sifting based on background content and entropy can take place simultaneously.) The CNN is then trained to distinguish between diagnostic and non-diagnostic tiles and, in the case of multiple disease types, among those classes as well.

Following CNN training, a new whole-slide image may be segmented by first resampling at the lower resolution 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. In a binary classification, pixel-level diagnostic probabilities may be computed by averaging (or otherwise combining) the tile-level probabilities for each pixel. This approach substantially improves the resolution obtainable by simple tile overlap without probability averaging. Softmax probabilities are more complex, however, and better segmentations may be obtained simply by taking the union of relevant overlapping tiles. In a multinomial classification, a tile is considered relevant if it corresponds to the dominant subtype, i.e., the subtype accounting for the majority of diagnostic tiles. Intriguingly, the accuracy of the segmentation is not necessarily compromised by an inaccurate subtype classification.

Two very different benchmark datasets were utilized in this study. Both consist of extremely large whole-slide images, which would be difficult to segment using conventional deep-learning techniques. The annotated training and validation slides of the PAIP2020 challenge[1] were selected to investigate segmentation of multiple disease subtypes. This dataset contains annotated whole slides 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). Histology samples are typically classified as MSI-H or MSI-L for diagnostic purposes.

We employed the CAMELYON16 [19] dataset to investigate segmentation of a single disease type — cancer metastases in lymph nodes. 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 whole-slide images, a sample is either malignant or normal, i.e., the lymph node contains no metastatic tissue. Although the binary classification task may seem simpler, the CAMELYON16 dataset is especially challenging in that diagnostic regions may be quite small — in some cases just a few pixels out of billions. Whereas the tumor regions in the PAIP2020 dataset tend to be large and contiguous, the CAMELYON16 lymph-node lesions often presented as a dusting of tiny features. The latter morphology tested the lower limits of useful tile sizes.

Where tile sizes were small enough relative to diagnostic features to produce meaningful segmentations, strong performance was observed. For the PAIP2020 dataset, segmentations achieved mean Jaccard similarity scores modestly exceeding 0.70 and recall (or sensitivity) scores well in excess of 0.90 (and as high as 0.96). Recall represents the proportion of the diagnostic region that actually appears in the segmentation map; high recall scores ensure that clinicians using the map as an aid to their practice will observe the full extent of the diseased area. Similar results were obtained for CAMELYON16 images with sufficiently large diagnostic features, although average scores were lower due to inclusion of all results that did not fail entirely.

[1] 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.

Materials, Methods, And Results

2.1 PAIP2020 Slides

The PAIP2020 training image dataset consisted of 47 whole slides — 12 of which were labeled as MSI-H and the remaining 35 as MSI-L — provided in multilevel SVS format. The slides had an average uncompressed size of 116,214 x 88,095 pixels and contained varying amounts of non-tissue

background. The dataset also included binary segmentation masks defining the tumor regions. An unannotated, unlabeled validation set consisted of 31 additional slides.

Four rescaled sets of the whole-slide images were prepared such that, in each set, the longer dimension of the rescaled images did not exceed 3500, 4500, 6000, or 8000 pixels. Each of these image sizes seemed potentially adequate to preserve the diagnostically significant anatomy without being so large as to limit the utility of tiles ranging in size from 200 x 200 to 600 x 600 pixels. The tile-preparation, training and testing procedures discussed below were carried out for each image set, and the 6000-pixel set emerged as the best performer. The images in this set had an average size of 8.5 Megapixels (MP) and a maximum size of 13.5 MP. (The maximum possible image size would be 36 MP, or 6000 x 6000 pixels.)

The provided ground-truth masks were first used to create new, separate images of the tumor and non-tumor portions of each image. To cope with the unbalanced training dataset, three different subsets of the 47 training slides were defined. Each subset included 29 training images (8 MSI-H images and 21 MSI-L images) and 18 test images (4 MSI-H images and 14 MSI-L images) to preserve, in each subset, a training/test split above 60/40. Each of the MSI-H test sets was unique, i.e., contained no images found in any other test set.

Within each training subset, the data was further unbalanced by the typically smaller image area occupied by the tumor. Consequently, to obtain similar numbers of tumor and non-tumor tiles, we overlapped them to different degrees. These ranged from 80% to 96% overlap depending on the image size and the number of images in each training class (MSI-H, MSI-L, and non-tumor).

Training tiles were sifted based on background fraction, with tiles having majority-background regions excluded. The background of tumor images — MSI-H and MSI-L images were 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. We identified background regions by creating, for each tile, an 8-bit grayscale tile counterpart. Tiles were excluded if their grayscale counterparts included a majority of pixels with values above 235 (at least nearly white) or fell below 15 (at least close to black). The higher grayscale limit was chosen so that staining would still register as background but light-colored tissue regions would not. About 55% of the generated tiles survived background sifting; this fraction was consistent across tile sizes. Each image subset had about 180,000 training tiles evenly split among the MSI-H, MSI-L, and non-tumor classes.

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. For each image subset, test tiles were prepared by decomposing each subset test image into overlapping tiles and sifting tiles based on background content and image entropy. In particular, majority-background tiles and tiles whose image entropies were not on or between the entropy rails were excluded. The same procedures — image rescaling, tile generation, and exclusion based on background content and entropy — were carried out on the PAIP2020 validation set. Values of the entropy rails were quite consistent — within 1% — across tile sizes ranging from 200 x 200 to 650 x 650; the values for the 400 x 400 tiles were 6.42 and 7.46. As will

be seen, the CAMELYON16 tiles behaved very differently at small tile sizes, and the spread between maximum and minimum entropy values can effectively limit the minimum usable tile size.

The CNN architecture employed in this study was selected to minimize the number of convolutional layers and consequent trainable parameter count. Three dropout layers mitigated the risk of overfitting to the small dataset. 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, softmax activation, and random horizontal and vertical flip data augmentation. More significant data augmentation resulted from the degree of tile overlap noted above. Source code for this model has been posted.[1]

Training and testing were carried out separately for each of the three tile subsets. In each case, the model was saved after each of the 75 training epochs. It was unclear *a priori* whether models producing the most accurate subtype classifications would also generate the best segmentations; therefore, segmentations were obtained using all models with classification accuracies exceeding 60%. In fact, as noted below, models exhibiting poor classification performance sometimes produced good segmentations. Best segmentation and classification performance were found to occur with 400 x 400 pixel tiles.

Following analysis by the CNN, the tiles of a candidate image have been sifted twice: first by the entropy rails prior to processing and then by the softmax activation function. Because of the high degree of overlap among tiles, the union of all tiles classified as MSI-H or MSI-L (whether or not the tumor-level classification is correct) were used to approximate the tumor region. The resulting segmentations, each based on an average of about 2000 test tiles, were assessed against the corresponding segmentation masks in terms of Jaccard similarity, precision, and recall. The Jaccard score quantifies the degree of overlap between the prediction P and the ground truth T:

$$\text{Jaccard score} = \frac{P \cap T}{P \cup T}$$

This metric is closely related to the Dice coefficient.

Precision represents the proportion of pixels classified as positive (i.e., as tumor pixels) that are, in fact, positive while recall 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*),

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

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

The Jaccard, precision, and recall scores obtained for masks produced with the best-performing models are shown in Table1. Scores on the validation set were comparable to those obtained for the different training/test subsets, particularly on a relative basis among the assessed metrics. Accordingly, the training subsets were generally representative and unbiased. Classification accuracies were obtained by majority vote from label counts corresponding to MSI-H and MSI-L classifications for each image. Better classification accuracy, 0.90 on the validation set, was achieved by combining label counts produced by the three models. This reflects the likelihood that the classification error attributable to each model has some degree of independence from the others, so at least some of the overall error is eliminated by averaging.

	Training Slides				*Validation Slides*		
	Mean Jaccard Score	Mean Precision	Mean Recall	Classif. Accuracy	Mean Jaccard Score	Mean Precision	Mean Recall
Set 1: model 27	0.69	0.75	0.91	0.89	0.64	0.72	0.84
Set 2: model 49	0.63	0.65	0.96	0.83	0.62	0.69	0.91
Set 3: model 11	0.65	0.71	0.89	0.89	0.58	0.77	0.74
Average	0.66	0.70	0.92	0.87	0.60	0.70	0.86

Table 1 – The best models, labeled by epoch number, were identified for each image subset and their segmentation masks compared against those prepared by expert pathologists. These models were tested on the validation set and produced roughly similar scores.

Surprisingly, successful segmentation was largely independent of proper image classification. The three images incorrectly classified by model 49, for example, had mean Jaccard, precision, and recall scores of 0.67, 0.77, and 0.85, respectively. This was true despite mapping with tiles of the dominant, and therefore incorrect, classification. Mapping with all tiles classified as either tumor subtype invariably produced lower-quality segmentations. Similarly, although the models producing the best segmentation metrics

also delivered the most accurate classifications, a few models exhibiting relatively poor classification performance generated unexpectedly good segmentations.

Our five-layer CNN performs favorably compared with U-Net applied to a single downscaled image. To make this assessment, the Keras platform was used to create a U-Net model configured to process 256 x 256 pixel images based on a frequently cited code example [20]. U-Net performs pixel-level binary classification based on a decision boundary, which is a hyperparameter of the architecture and ranges from 0 to 1. An initial value of 0.5 is common. For benchmarking purposes, we trained this model on 20 images from from the ISBI Challenge [21], a dataset of neuronal structures, rescaled to pixel dimensions of 256 x 256. U-Net is known to excel at segmenting neural tissue containing sharply defined structures with clear contrast [10]. To approximate the effect of tile overlap, various forms of data augmentation were employed: width and height shifts, shear, and zoom, all set at 0.05, and random horizontal flips as were utilized on the PAIP2020 training images. Tested on 10 ISBI images, this U-Net model delivered a mean Jaccard score of 0.90, a mean precision of 0.94, and a mean recall of 0.95.

The ISBI neuronal images featured well-defined patterns that were structurally similar across the dataset. This is not the case for the colorectal cancer images, on which the U-Net model performed poorly following the same training and test procedure (see Table 2). Performance improved as the decision boundary was reduced, but all metrics fall well below those achieved with the five-layer CNN on tiles drawn from much larger images. There is simply not enough anatomy visible in a 256 x 256 colorectal cancer image to support accurate segmentation. Also noteworthy is the size of the U-Net model at over 31 million parameters.

	Set 2 - U-Net (PAIP2020 Training Slides)		
	Mean Jaccard Score	Mean Precision	Mean Recall
Threshold = 0.5	0.48	0.76	0.53
Threshold = 0.05	0.53	0.74	0.62
Threshold = 0.001	0.56	0.69	0.70

Table 2 – Performance of U-Net model trained and tested on the second colorectal cancer image subset (29 training images, 18 test images).

Without modification, masks generated by tile overlap as described above have blocky edges with stepped features, the roughness of which depends on the tile size and the degree of overlap. Although image blurring is to be avoided, it is possible to smooth the edges while preserving their sharpness using morphological operations based on a structuring element or kernel (Fig. 2), which defines a neighborhood shape and size. Using a circular kernel to first shrink (“erode”) and then expand (“dilate”) white mask regions results in progressively rounder, smoother edges as the kernel size increases.

Segmentation maps resulting from softer-edged masks have less visual distraction — they are more user-friendly — but obviously what are ultimately aesthetic considerations cannot trump accuracy. Fortunately, as indicated in Table 3, the effect of smoothing on the accuracy metrics is minimal over a visually significant range of kernel sizes. More concerning is the loss of diagnostic visual elements that can occur when the kernel size becomes a significant fraction of the tile size (Fig. 2(d)), which imposes an upper limit on smoothing. Although Table 3 shows results for only one model, equivalent results were obtained for the best models of the other two image subsets.

	Set 2 - Model 49 (Training Slides)					
	Edge smoothing (kernel size = 150 pixels)	Edge smoothing (kernel size = 100 pixels)	Edge smoothing (kernel size = 50 pixels)	Un-smoothed	Isomorphic shrink 5%	No rails (un-smoothed)
Mean Jaccard Score	0.64	0.64	0.63	0.63	0.64	0.62
Mean Precision	0.67	0.66	0.66	0.66	0.70	0.68
Mean Recall	0.94	0.95	0.96	0.96	0.89	0.90

Table 3 – Effects of edge smoothing, isomorphic shrinkage, and use of entropy rails on similarity metrics. With the rails omitted, tiles were sifted based only on the amount of background,with majority-background tiles excluded.

Fig. 3 illustrates the practical benefit of high recall combined with at least acceptable precision and Jaccard scores. In the best case, 3(a), the segmentation includes the entire lesion and very little else. Even for the worst performer, 3(c), nearly all diagnostically relevant tissue is captured with spurious highlighting confined to regions immediately surrounding the lesion. Given this pattern, it seemed plausible that isomorphically shrinking the diagnostic mask regions might improve segmentation quality. As shown in Table 3, however, shrinking by 5% has minimal impact on overall (Jaccard) similarity while mean recall diminshes significantly. The effect of a 10% reduction is worse. The reason for this is the uneven distribution of misclassified pixels around a lesion; the error margin in some regions is larger than in others, so the beneficial and deleterious effects of an isomorphic size reduction largely cancel out.

Finally, Table 3 shows the improvement provided by sifting using entropy rails rather than simple background thresholding. While not dramatic, the effect — particularly on recall — is appreciable.

2.2 CAMELYON2016 Slides

The CAMELYON2016 dataset consists of whole-slide images provided in multilevel TIFF format. The dataset includes segmentation masks prepared by expert pathologists for 111 of these slide images,

which have an average uncompressed size of 88,816 x 55,352 pixels. While a few of the images feature large tumor regions, such as that shown in Fig. 1(b), the majority have small lesions that may themselves consist of archipelago-like clusters of minuscule features (see Fig. 4).

This necessitated a much larger rescaled image size. In order to be classified properly as a tumor tile, at least half the tile area must be occupied by tumor tissue; the tile must contain enough image information to permit the CNN to distinguish reliably among classes. 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. And finally, if possible, the rescaled image should be small enough to be stored and processed on a mobile device. Balancing these considerations ultimately led to a maximum dimension of 15,000 for the rescaled image.

Ninety of 111 annotated tumor-containing whole-slide images were selected for training and validation, and 20 of the remaining 21 annotated images served as the test set. Tiles were prepared at different sizes for the tumor and non-tumor portions of each image as described above. The criterion of fidelity dictated a maximum practical tile size of 400 x 400. As shown in Fig. 5, the spread between minimum and maximum entropy values increased substantially below size 200 x 200, which produced the best segmentations. For this dataset at the selected degree of image rescaling, smaller tumor tiles had insufficient visual diversity (presumably arising from insufficiently distinctive anatomic detail) to be well-characterized by the entropy criterion. As a consequence, fewer tiles were rejected during preprocessing.

To obtain roughly equal sets of tumor and non-tumor tiles after majority-background sifting, the tumor tiles at size 200 x 200 and above were overlapped by amounts ranging from 86% to 90% and the non-tumor tiles were overlapped by amounts ranging from 50% to 67%. At each size, enough tiles were removed at random from the class having the larger resulting population to equalize the tumor and non-tumor tile sets. These training sets ranged in size from 245,735 tiles of each class at 100 x 100 pixels to 21,576 tiles of each class at size 400 x 400.

The CNN architecture used for this binary classification task was unchanged from that described above but for training we used a binary cross-entropy loss function and sigmoid activation. Once again, models were saved after each training epoch but overfitting set in much earlier — generally after 25 epochs.

Best performance was observed with 200 x 200 pixel tiles. As shown in Table 4, the models that achieved highest classification accuracies also produced the best segmentations, but the similarity metrics other than recall were only fair. The large tile size relative to the size of tumor features in some images resulted in a few lesions escaping detection altogether. In other cases, the tumor area was fully captured or nearly so, resulting in high recall scores, but overall similarity suffered due to the large tile size; the tissue approximations, in other words, were coarse. Still, recall scores were above 0.9 for 41% of the images that received a score and above 0.5 for 88% of those images. In most cases, that is, the tumor regions were reasonably well covered despite the small feature sizes. Where feature sizes were large relative to the tile size (as in Fig. 1), performance was comparable to that achieved with the PAIP2020 dataset.

At 150 x 150 and 100 x 100 pixels, tile classification failed altogether. Despite training accuracies that exceeded 99%, none of the test tiles were classified as tumor and the resulting segmentation masks contained no white regions. That this might occur was suggested by the sudden increase in the spread between minimum and maximum entropy values in Fig. 5. With insufficient anatomic information in the tiles to distinguish between tumor and non-tumor tissue, the CNN seems to have overfit immediately to the training tiles.

	Test Slides			
	Mean Nonzero Jaccard Score	Mean Nonzero Precision	Mean Nonzero Recall	Classif. Accuracy
Model 14	0.30	0.34	0.74	0.99
Model 20	0.39	0.45	0.73	0.98
Average	0.35	0.40	0.74	0.99

Table 4 – Once again the best models, labeled by epoch number, were identified and their segmentation masks compared to ground-truth masks prepared by expert pathologists. Three (in the case of model 20) or four (in the case of model 14) of the segmentation masks showed no relevant features and received scores of zero for all metrics. These corresponded to images with tumor features that were small relative to the tiles size and spread out, so no tiles intercepted enough tumor tissue to trigger a positive classification.

[1] https://github.com/stevenjayfrank/A-Eye.

Discussion

The primary objective of this work is to make accurate tissue segmentation possible for telemedicine and other applications that utilize computationally limited devices and bandwidth-limited communication. In traditional clinical practice, a complex or ambiguous case may be reviewed by more than one pathologist; with access to immunohistochemistry and molecular information in addition to whole-slide images of H&E-stained biopsy samples, an expert team can manually create tissue segmentations and diagnose disease with extremely high accuracy. But not every biopsy can be evaluated — at least not immediately — by specialists at a comprehensive cancer center. 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*; a virtual team need not pore over the same image shoulder-to-shoulder. A screening pathologist, remote from the biopsy site and from other pathologists, might post an electronic request for additional perspectives — with links to the rescaled image and the lightweight model employed. Available responders, whether in the hospital or on an airplane, can comment or use their own

favored lightweight models to generate alternative segmentations. Such “expert crowdsourcing” can beneficially match immediate clinical needs with ready expertise.

Achieving this objective requires computational simplicity, limited storage requirements, and sufficient consistency and reliability across tissue types. Since the goal is to assist rather than outperform clinicians, recall represents the most important similarity metric. A high recall ensures that most or all of the diseased region is made visible in the segmentation. So long as precision and Jaccard scores are within reason (i.e., the tissue misclassified as diseased is not excessive), the segmentation will be useful as diagnostic guidance.

Tiling is a mechanically simple array operation. By preprocessing tiles as they are generated to exclude those with image entropies outside previously established boundaries, CNN classification accuracy is enhanced and fewer tiles need be classified. The spread between minimum and maximum entropy values is a critical parameter. As the gap widens, the fraction of tiles excluded — and hence the utility of the criterion for preprocessing — diminishes. Moreover, because of the logarithmic nature of the entropy function, relatively small increases or decreases in the gap width have significant effects on the proportion of tiles excluded. When the gap expands suddenly, the capacity to distinguish between tumor and nontumor tiles disappears. This abrupt change, which occurred with the smallest CAMELYON16 tiles but not similarly sized PAIP2020 tiles, reveals a close tie between image entropy and CNN performance.

Ultimately, the parameter of chief importance is tile size, since this quantity dictates the resolution of the segmentation — certainly if tiles are overlaid to form the map or mask, but even pixel-level resolution is progressively assessed at a tile level. Classification accuracy also depends strongly on tile size and its companion variable, the degree of image rescaling; together, these determine how much anatomy a tile will contain. Although it may seem tempting simply to maximize tile-level classification accuracy, this may fail to optimize segmentation resolution; some models were mediocre classifiers yet delivered very good segmentations, so peak classification accuracy may not guarantee the best segmentations. The approach we describe here requires systematic exploration of the parameter space for a particular tissue type and slide magnification, but once established, the resulting parameter values should prove robust across similar slides.

The foregoing discussion centered on whole-slide images because they are difficult to analyze computationally and unwieldy to transmit and store; as a result, the benefits of the described approach are most pronounced. Preprocessing with entropy rails is useful for any image too large to be processed directly by a CNN, however. Many medical images, from MRI scans to simple X-rays, are currently downsampled to satisfy CNN input requirements at the sacrifice of image detail. Tiling the original image rather than rescaling it preserves this detail. Indeed, if rescaling to an intermediate size is unnecessary, a hyperparameter — the degree of rescaling — is eliminated from consideration and the overall processing flow is simplified. Training and testing are further simplified by assessing the spread between minimum and maximum tile entropies. Those tile sizes with widening spreads can probably be omitted from consideration since they are unlikely to perform well. We believe that image tiling and sifting based on

entropy rails offers a versatile alternative to many automated image-analysis protocols now in use — an alternative that is readily implemented on mobile devices and whose constituent models and images are easily shared.

References

1. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L. & Erickson, B. J. Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *Journal of Digital Imaging* (2017) doi:10.1007/s10278-017-9983-4.
2. Havaei, M. *et al.* Brain tumor segmentation with Deep Neural Networks. *Med. Image Anal.* (2017) doi:10.1016/j.media.2016.05.004.
3. Zhang, W. *et al.* Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage* (2015) doi:10.1016/j.neuroimage.2014.12.061.
4. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2015). doi:10.1109/CVPR.2015.7298965.
5. Noh, H., Hong, S. & Han, B. Learning deconvolution network for semantic segmentation. in *Proceedings of the IEEE International Conference on Computer Vision* (2015). doi:10.1109/ICCV.2015.178.
6. 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 *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science* (2015).
7. Liu, L. *et al.* A survey on U-shaped networks in medical image segmentations. *Neurocomputing* (2020) doi:10.1016/j.neucom.2020.05.070.
8. Dong, H., Yang, G., Liu, F., Mo, Y. & Guo, Y. Automatic brain tumor detection and segmentation using U-net based fully convolutional networks. in *Communications in Computer and Information Science* vol. 723 (2017).
9. Iglovikov, V., Rakhlin, A., Kalinin, A. & Shvets, A. Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks. *bioRxiv* (2017) doi:10.1101/234120.
10. UNet: a convolutional network for biomedical image segmentation. https://hpc.nih.gov/apps/UNet.html.
11. 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) doi:10.1016/j.neucom.2016.01.034.
12. 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) doi:10.1016/j.compmedimag.2019.101686.

13. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2015). doi:10.1007/978-3-319-24574-4_28.
14. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S. O., Villena-Martinez, V. & Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* (2017).
15. Ibtehaz, N. & Rahman, M. S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *arXiv* (2019).
16. Frank, S. J. Resource-frugal classification and analysis of pathology slides using image entropy. *arXiv* (2020).
17. Yu, K. H. *et al.* Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. *J. Am. Med. Informatics Assoc.* (2020) doi:10.1093/jamia/ocz230.
18. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* (2018) doi:10.1038/s41591-018-0177-5.
19. Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - J. Am. Med. Assoc.* **318**, 2199–2210 (2017).
20. GitHub - zhixuhao/unet: unet for image segmentation. https://github.com/zhixuhao/unet.
21. About the 2D EM segmentation challenge | ISBI Challenge: Segmentation of neuronal structures in EM stacks. http://brainiac2.mit.edu/isbi_challenge/.

Figures

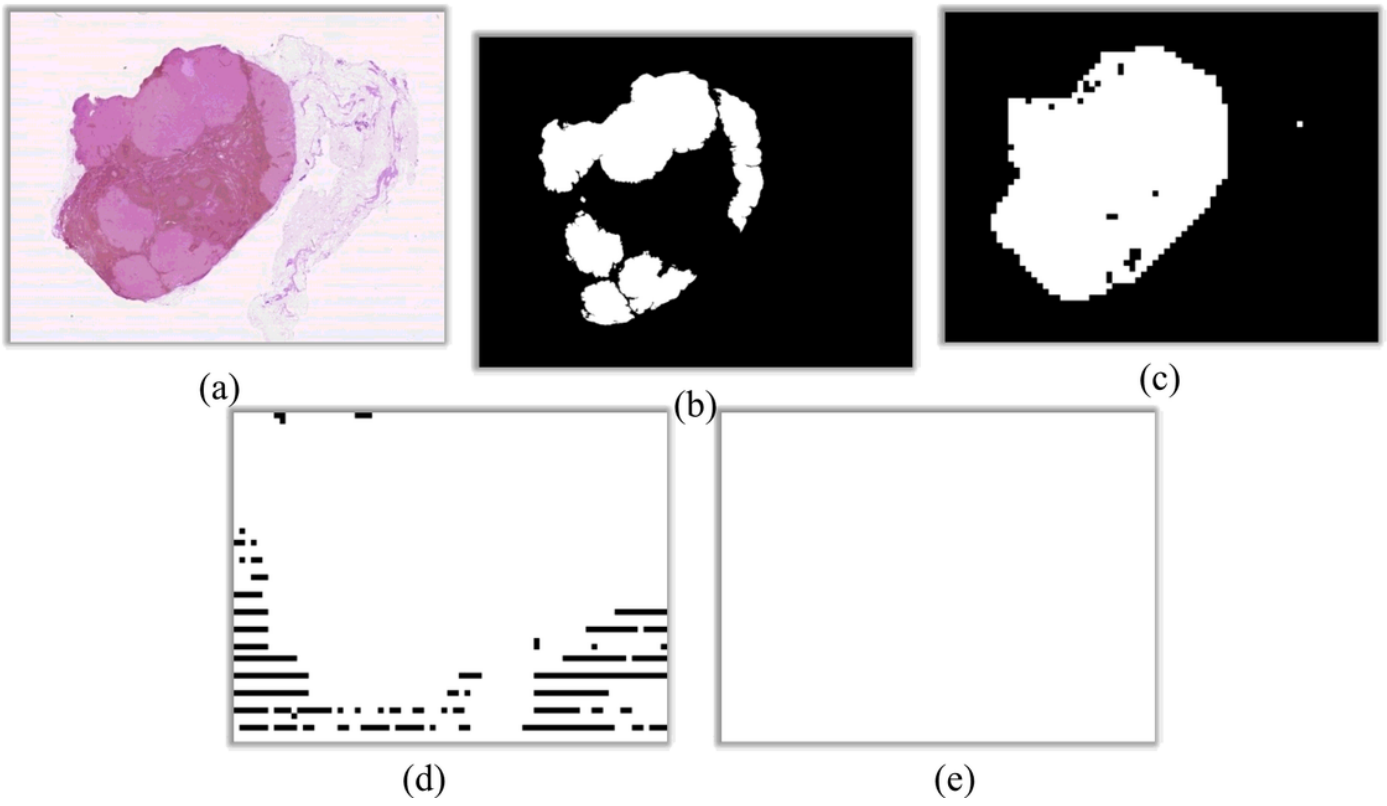

Figure 1

Effect of sifting with entropy rails. (a) Source histology image of metastatic lymph node tissue from CAMELYON16 dataset. (b) Ground-truth segmentation mask for the image in (a). When applied to the image, the segmentation mask occludes non-tumor tissue regions. (c) Location map of overlapping tiles created from the image in (a) and having entropies within the rails established for the training image set. The white region represents the union of the qualifying tiles and exceeds, but approximates, the tumor region. (d, e) Location map of overlapping tiles, created from the image in (a), sifted using background thresholding. In (d), only tiles with background regions constituting less than 10% of the tile image were retained; when the permissible background fraction is raised to 50%, as in (e), no image regions are excluded. The horizontal bands in (d) track subtle stripe artifacts in the source image (a); these are eliminated by sifting with entropy rails.

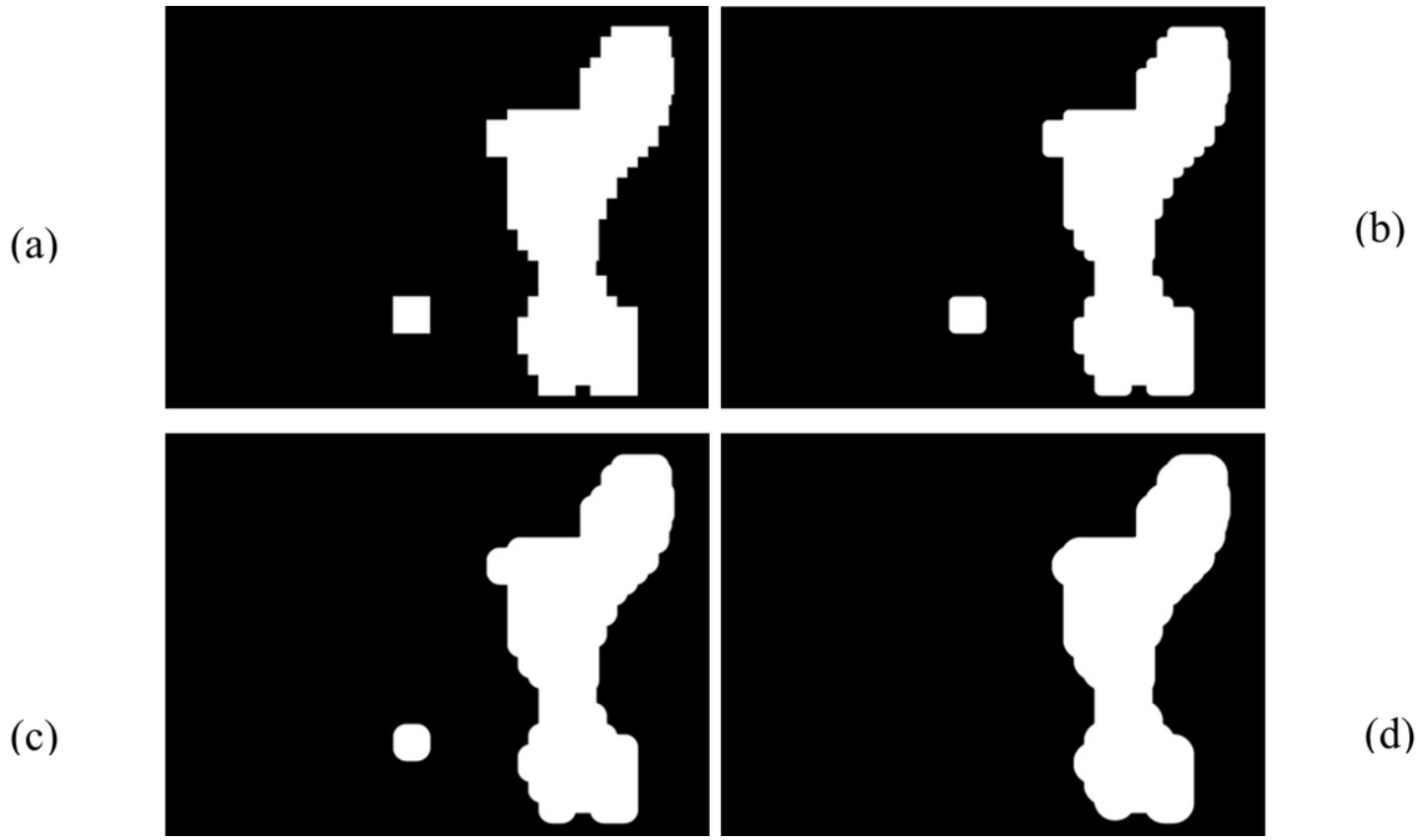

Figure 2

Smoothing the masks. Unsmoothed masks (a) exhibit considerable edge roughness. Edge smoothing using a circular structuring element of varying sizes - (b) 50 pixels, (c) 100 pixels, (d) 150 pixels - produces different degrees of smoothing without blur. (d) Too large a structuring element can eliminate important features altogether. The isolated square in (a) illustrates the size of a 400 x 400 pixel tile relative to the rescaled image.

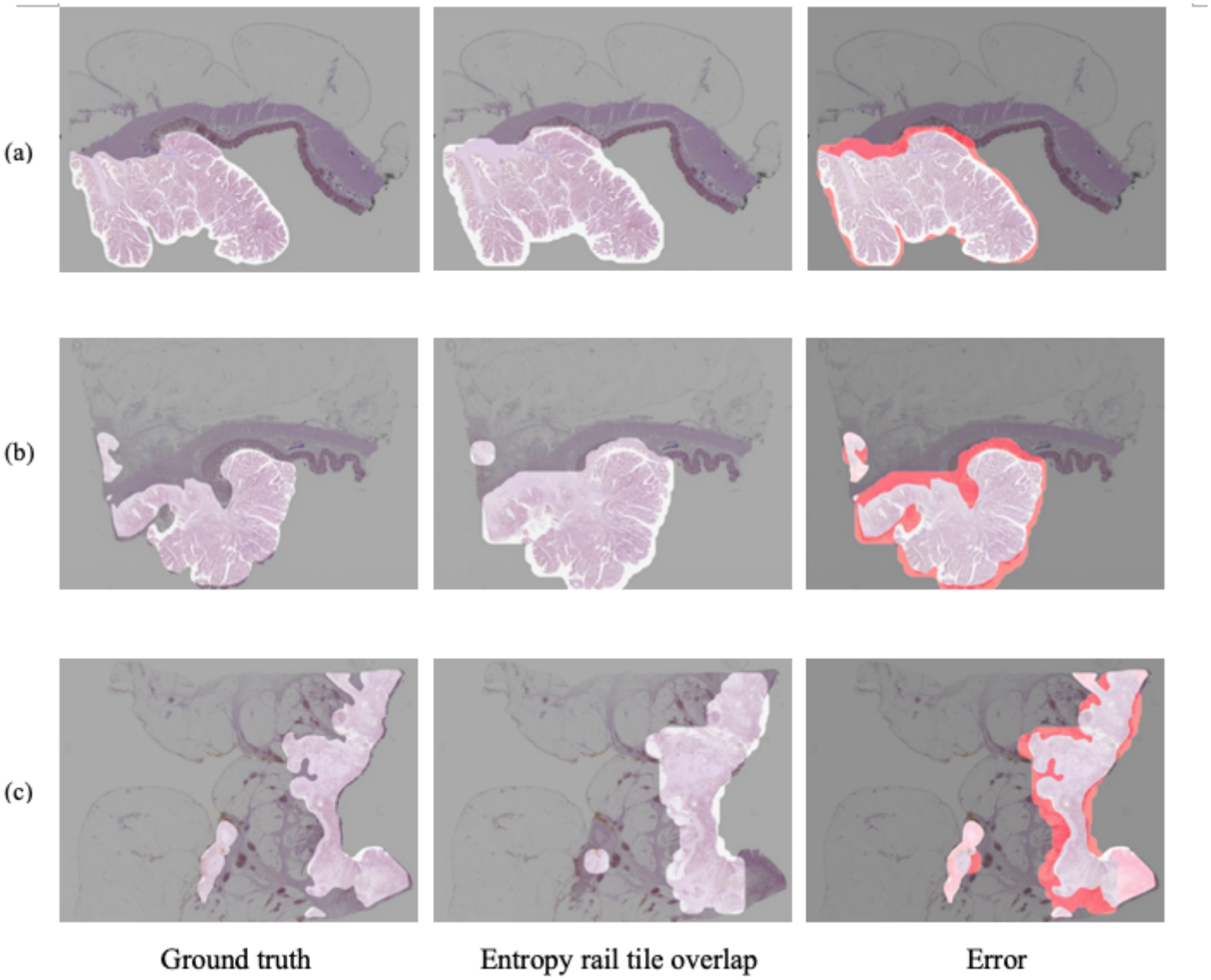

Figure 3

From Best to Worst. Ground truth segmentation maps with lesions highlighted, corresponding segmentations generated using entropy rails, and the error (colored red) associated with the latter is shown for three segmentations produced using model 49. The metrics in each case were: (a) Jaccard = 0.86, precision = 0.86, recall = 1.0; (b) Jaccard = 0.72, precision = 0.74, recall = 0.95; (c) Jaccard = 0.55, precision = 0.68, recall = 0.74. Image (c) was the worst performer across all three metrics. Of the 18 test images in this subset, 9 had recall scores of at least 0.99 and recall for all but 3 exceeded 0.90.

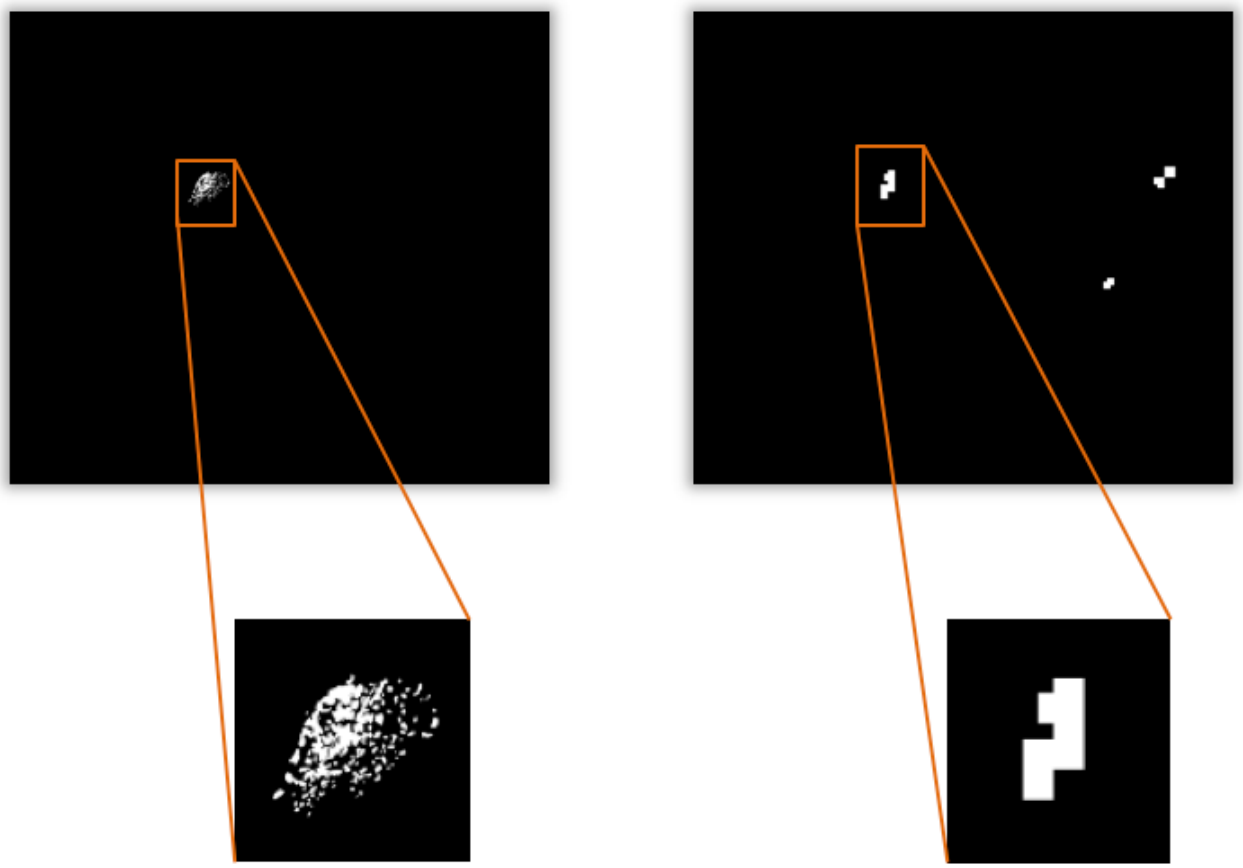

Figure 4

Tumor features in representative CAMELYON16 image. Left, ground-truth 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 produced by the best-performing model. Only the denser regions of the tumor were recognized and segmented properly, and the size of a tile - an example of which appears toward the upper right - limits the ability both to detect and represent tumor features.

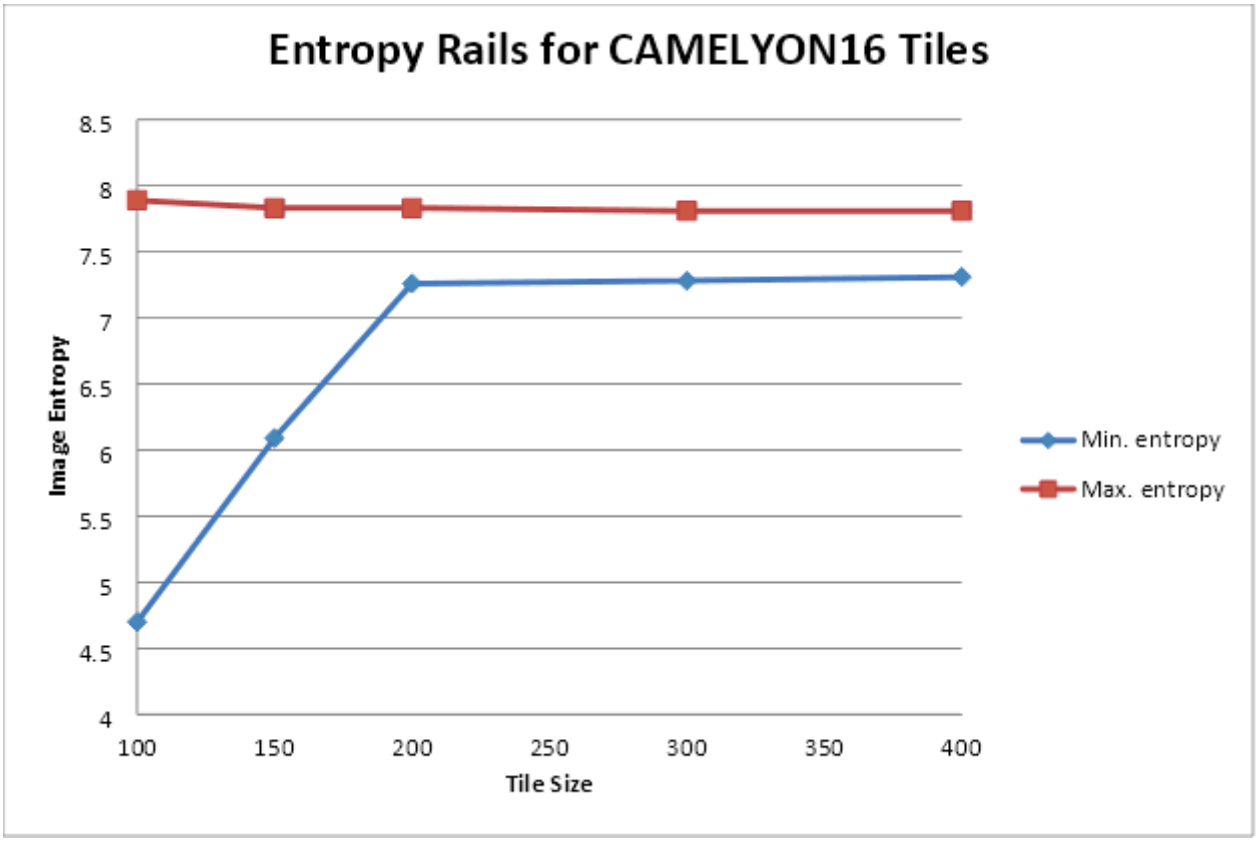

Figure 5

Entropy rails for CAMELYON16 tiles. At size 200 x 200 and above, tiles generated from tumor regions fell within a narrow band of entropy values. The gap widened considerably below this size.