# Detecting Cancer Metastases on Gigapixel Pathology Images

Yun Liu[1⋆], Krishna Gadepalli[1], Mohammad Norouzi[1], George E. Dahl[1], Timo Kohlberger[1], Aleksey Boyko[1], Subhashini Venugopalan[2⋆⋆], Aleksei Timofeev[2], Philip Q. Nelson[2], Greg S. Corrado[1], Jason D. Hipp[3], Lily Peng[1], and Martin C. Stumpe[1]

{liuyun,mnorouzi,gdahl,lhpeng,mstumpe}@google.com

[1]Google Brain, [2]Google Inc, [3]Verily Life Sciences,
Mountain View, CA, USA

**Abstract.** Each year, the treatment decisions for more than 230,000 breast cancer patients in the U.S. hinge on whether the cancer has metastasized away from the breast. Metastasis detection is currently performed by pathologists reviewing large expanses of biological tissues. This process is labor intensive and error-prone. We present a framework to automatically detect and localize tumors as small as $100 \times 100$ pixels in gigapixel microscopy images sized $100,000 \times 100,000$ pixels. Our method leverages a convolutional neural network (CNN) architecture and obtains state-of-the-art results on the Camelyon16 dataset in the challenging lesion-level tumor detection task. At 8 false positives per image, we detect 92.4% of the tumors, relative to 82.7% by the previous best automated approach. For comparison, a human pathologist attempting exhaustive search achieved 73.2% sensitivity. We achieve image-level AUC scores above 97% on both the Camelyon16 test set and an independent set of 110 slides. In addition, we discover that two slides in the Camelyon16 training set were erroneously labeled normal. Our approach could considerably reduce false negative rates in metastasis detection.

**Keywords:** neural network, pathology, cancer, deep learning

## 1 Introduction

The treatment and management of breast cancer is determined by the disease stage. A central component of breast cancer staging involves the microscopic examination of lymph nodes adjacent to the breast for evidence that the cancer has spread, or metastasized [3]. This process requires highly skilled pathologists and is fairly time-consuming and error-prone, particularly for lymph nodes with either no or small tumors. Computer assisted detection of lymph node metastasis could increase the sensitivity, speed, and consistency of metastasis detection [16].

⋆ Work done as a Google Brain Resident (g.co/brainresidency).
⋆⋆ Work done as a Google intern.

In recent years, deep CNNs have significantly improved accuracy on a wide range of computer vision tasks such as image recognition [14, 11, 19], object detection [8], and semantic segmentation [17]. Similarly, deep CNNs have been applied productively to improve healthcare (*e.g.,*[9]).

This paper presents a CNN framework to aid breast cancer metastasis detection in lymph nodes. We build on [23] by leveraging a more recent Inception architecture [20], careful image patch sampling and data augmentations. Despite performing inference with stride 128 (instead of 4), we halve the error rate at 8 false positives (FPs) per slide, setting a new state-of-the-art. We also found that several approaches yielded no benefits: **(1)** a multi-scale approach that mimics the human cognition of a pathologist's examination of biological tissue, **(2)** pre-training the model on ImageNet image recognition, and **(3)** color normalization. Finally, we dispense with the random forest classifier and feature engineering used in [23] and find that the maximum function is an effective whole-slide classification procedure.

**Related Work** Several promising studies have applied deep learning to histopathology. The Camelyon16 challenge winner [1] achieved a sensitivity of 75% at 8 FP per slide and a slide-level classification AUC of 92.5% [23]. The authors trained a Inception (V1, GoogLeNet) [20] model on a pre-sampled set of image patches, and trained a random forest classifier on 28 hand-engineered features to predict the slide label. A second Inception model was trained on harder examples, and predicted points were generated using the average of the two models' predictions. This team later improved these metrics to 82.7% and 99.4% respectively [1] using color normalization [4], additional data augmentation, and lowering the inference stride from 64 to 4. The Camelyon organizers also trained CNNs on smaller datasets to detect breast cancer in lymph nodes and prostate cancer biopsies [16]. [12] applied CNNs to segmenting or detecting nuclei, epithelium, tubules, lymphocytes, mitosis, invasive ductal carcinoma and lymphoma. [7] demonstrated that CNNs achieved higher F1 score and balanced accuracy in detecting invasive ductal carcinoma. CNNs were also used to detect mitosis, winning the ICPR12 [6] and AMIDA13 [22] mitosis detection competitions. Other efforts at leveraging machine learning for predictions in cancer pathology include predicting prognosis in non-small cell lung cancer [25].

## 2 Methods

Given a gigapixel pathology image (*slide*[1]), the goal is to classify if the image contains tumor and localize the tumors for a pathologist's review. This use case and the difficulty of pixel-accurate annotation (Fig. 2) renders detection and localization more important than pixel-level segmentation. Because of the large size of the slide and the limited number of slides (270), we train models using

[1] Each slide contains human lymph node tissue stained with hematoxylin and eosin (H&E), and is scanned at the most common high magnification in a microscope, "40X". We also experimented with 2- and 4-times down-sampled patches ("20X" and "10X").

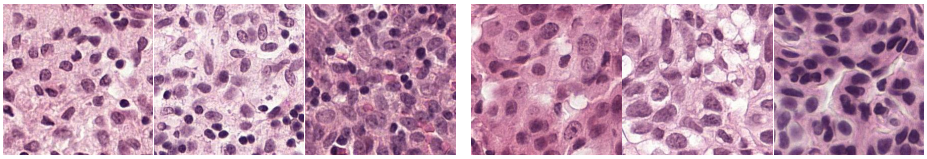

**Fig. 1. Left**: three tumor patches and **right**: three challenging normal patches.

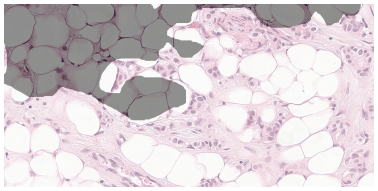

**Fig. 2.** Difficulty of pixel-accurate annotations for scattered tumor cells. Ground truth annotation is overlaid with a lighter shade. Note that the tumor annotations include both tumor cells and normal cells *e.g.,*white space representing adipose tissue (fat).

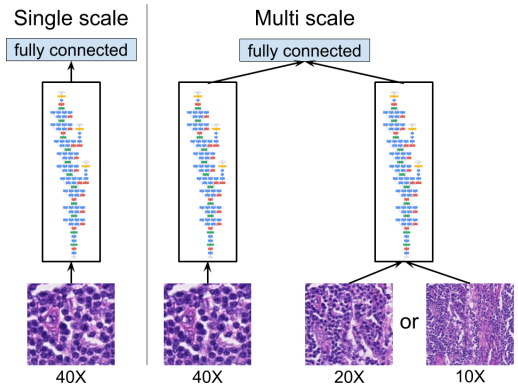


**Fig. 3.** The three colorful blocks represent Inception (V3) towers up to the second-last layer (PreLogit). *Single scale* utilizes one tower with input images at 40X magnification; *multi-scale* utilizes multiple (*e.g.,*2) input magnifications that are input to separate towers and merged.

smaller image *patches* extracted from the slide (Fig. 1). Similarly, we perform inference over patches in a sliding window across the slide, generating a tumor probability *heatmap*. For each slide, we report the maximum value in the heatmap as the slide-level tumor prediction.

We utilize the Inception (V3) architecture [20] with inputs sized $299 \times 299$ (the default) to assess the value of initializing from existing models pre-trained on another domain. For each input patch, we predict the label of the center $128 \times 128$ region. A 128 pixel region can span several tumor cells and was also used in [16]. We label a patch as tumor if at least one pixel in the center region is annotated as tumor. We explored the influence of the number of parameters by reducing the number of filters per layer while keeping the number of layers constant (*e.g., depth_multiplier* $= 0.1$ in TensorFlow). We denote these models "small". We also experimented with multi-scale approaches that utilize patches at multiple magnifications centered on the same region (Fig. 3). Because preliminary experiments did not show a benefit from using up to four magnifications, we present results only for up to two magnifications.

Training and evaluating our models was challenging because of the large number of patches and the tumor class imbalance. Each slide contains $10,000$

to $400,000$ patches (median $90,000$). However, each tumor slide contains 20 to $150,000$ tumors patches (median $2,000$), corresponding to tumor patch percentages ranging from 0.01% to 70% (median 2%). Avoiding biases towards slides containing more patches (both normal and tumor) required careful sampling. First, we select "normal" or "tumor" with equal probability. Next, we select a slide that contains that class of patches uniformly at random, and sample patches from that slide. By contrast, some existing methods pre-sample a set of patches from each slide [23], which limits the breadth of patches seen during training.

To combat the rarity of tumor patches, we apply several data augmentations. First, we rotate the input patch by 4 multiples of 90°, apply a left-right flip and repeat the rotations. All 8 orientations are valid because pathology slides do not have canonical orientations. Next, we use TensorFlow's image library (*tensorflow.image.random_X*) to perturb color: brightness with a maximum delta of 64/255, saturation with a maximum delta of 0.25, hue with a maximum delta of 0.04, and contrast with a maximum delta of 0.75. Lastly, we add jitter to the patch extraction process such that each patch has a small x,y offset of up to 8 pixels. The magnitudes of the color perturbations and jitter were lightly tuned using our validation set. Pixel values are clipped to $[0, 1]$ and scaled to $[-1, 1]$.

We run inference across the slide in a sliding window with a stride of 128 to match the center region's size. For each patch, we apply the rotations and left-right flip to obtain predictions for each of the 8 orientations, and average the 8 predictions.

**Implementation Details** We trained our networks with stochastic gradient descent in TensorFlow [2], with 8 replicas each running on a NVIDIA Pascal GPU with asynchronous gradient updates and batch size of 32 per replica. We used RMSProp [21] with momentum of 0.9, decay of 0.9 and $\epsilon = 1.0$. The initial learning rate was 0.05, with a decay of 0.5 every 2 million examples. For refining a model pretrained on ImageNet, we used an initial learning rate of 0.002.

## 3 Evaluation and Datasets

We use the two Camelyon16 evaluation metrics [1]. The first metric, the area under receiver operating characteristic, (Area Under ROC, AUC) [10] evaluates *slide-level* classification. This metric is challenging because of the potential for FPs when $10^5$ patch-level predictions are obtained per slide. We obtained 95% confidence intervals using a bootstrap approach[2].

The second metric, FROC [5], evaluates *tumor detection and localization*. We first generate a list of coordinates and corresponding predictions from each heatmap. Among all coordinates that fall within each annotated tumor region, the highest prediction is retained. Coordinates falling outside tumor regions are FPs. We use these values to compute the ROC. The FROC is defined as the sensitivity at $0.25, 0.5, 1, 2, 4, 8$ average FPs per tumor-negative slide [16]. This

[2] Sample with replacement $n$ slides from the dataset/split, where $n$ is the number of slides in the dataset/split, and compute the AUC. Repeat for a total of 2000 bootstrap samples, and report the 2.5 and 97.5 percentile values.

metric is challenging because reporting multiple points per FP region can quickly erode the score. We focused on the FROC as opposed to the AUC because there are approximately twice as many tumors as slides, which improves the reliability of the evaluation metric. Similar to the AUC, we report 95% confidence intervals by computing the FROC over 2000 bootstrap samples of the predicted points. In addition, we report the sensitivity at 8 FP per slide ("@8FP") to assess the false negative rate.

To generate points for FROC computation, the Camelyon winners [23, 1] thresholded the heatmap to produce a bit-mask, and reported a single prediction for each connected component in the bit-mask. By contrast, we use a non-maxima suppression method similar to [6] that repeats two steps until no values in the heatmap remain above a threshold $t$: **(1)** report the maximum and corresponding coordinate, and **(2)** set all values within a radius $r$ of the maximum to 0. Because we apply this procedure to the heatmap, $r$ has units of 128 pixels. $t$ controls the number of points reported and has no effect on the FROC unless the curve plateaus before 8 FP. To avoid erroneously dropping tumor predictions, we used a conservative threshold of $t = 0.5$.

**Datasets** Our work utilizes the Camelyon16 dataset [1], which contains 400 slides: 270 slides with pixel-level annotations, and 130 unlabeled slides as a test set.[3] We split the 270 slides into train and validation sets (Supplement) for hyperparameter tuning. Typically only a small portion of a slide contains biological tissue of interest, with background and fat comprising the remainder (*e.g.,* Fig. 2). To reduce computation, we removed background patches (gray value $> 0.8$ [12]), and verified visually that lymph node tissue was not discarded.

**Additional Evaluation: NHO-1** We digitized another set of 110 slides (57 containing tumor) from H&E-stained lymph nodes extracted from 20 patients (86 biological tissue blocks[4]) as an additional evaluation set. These slides came with patient- or block-level labels. To determine the slide labels, a board-certified pathologist blinded to the predictions adjudicated any differences, and briefly reviewed all 110 slides.

## 4 Experiments & Results

To perform slide-level classification, the current state-of-the-art methods apply a random forest to features extracted from a heatmap prediction [1]. Unfortunately, we were unable to train slide-level classifiers because the 100% validation-set AUC (Table 1) rendered internal evaluation of improvements impossible. Nonetheless, using the maximum value of each slide's heatmap achieved AUCs $> 97\%$, statistically indistinguishable from the current best results.

For tumor-level classification, we find that the connected component approach [23] provides a $1-5\%$ gain in FROC when the FROC is modest ($< 80\%$), by masking FP regions. However, this approach is sensitive to the threshold (up

[3] The test slides labels were released recently as part of the training dataset for Camelyon17. We used these labels for evaluation, but not for parameter tuning.

[4] A tissue block can contain multiple slides that vary considerably at the pixel level.

| Input & model size | Validation FROC | Validation @8FP | Validation AUC | Test FROC | Test @8FP | Test AUC |
|---|---|---|---|---|---|---|
| 40X | 98.1 | 100 | 99.0 | **87.3** (83.2, 91.1) | **91.1** (87.2, 94.5) | **96.7** (92.6, 99.6) |
| 40X-pretrained | 99.3 | 100 | 100 | **85.5** (81.0, 89.5) | **91.1** (86.8, 94.6) | **97.5** (93.8, 99.8) |
| 40X-small | 99.3 | 100 | 100 | **86.4** (82.2, 90.4) | **92.4** (88.8, 95.7) | **97.1** (93.2, 99.8) |
| ensemble-of-3 | - | - | - | **88.5** (84.3, 92.2) | **92.4** (88.7, 95.6) | **97.7** (93.0, 100) |
| 20X-small | 94.7 | 100 | 99.6 | **85.5** (81.0, 89.7) | **91.1** (86.9, 94.8) | **98.6** (96.7, 100) |
| 10X-small | 88.7 | 97.2 | 97.7 | 79.3 (74.2, 84.1) | 84.9 (80.0, 89.4) | **96.5** (91.9, 99.7) |
| 40X+20X-small | 94.9 | 98.6 | 99.0 | **85.9** (81.6, 89.9) | **92.9** (89.3, 96.1) | **97.0** (93.1, 99.9) |
| 40X+10X-small | 93.8 | 98.6 | 100 | 82.2 (77.0, 86.7) | 87.6 (83.2, 91.7) | **98.6** (96.2, 99.9) |
| Pathologist [1] | - | - | - | 73.3* | 73.3* | **96.6** |
| Camelyon16 winner [1, 23] | - | - | - | 80.7 | 82.7 | **99.4** |

**Table 1.** Results on Camelyon16 dataset (95% confidence intervals, CI). Bold indicates results within the CI of the best model. "Small" models contain 300K parameters per Inception tower instead of 20M. -: not reported. *A pathologist achieved this sensitivity (with no FP) using 30 hours.

to $10 - 20\%$ variance), and can confound evaluation of model improvements by grouping multiple nearby tumors as one. By contrast, our non-maxima suppression approach is relatively insensitive to $r$ between 4 and 6, although less accurate models benefited from tuning $r$ using the validation set (*e.g.,* 8). Finally, we achieve 100% FROC on larger tumors (macrometastasis), indicating that most false negatives are comprised of smaller tumors.

Previous work (*e.g.,* [24, 9]) has shown that *pre-training* on a different domain improves performance. However, we find that although pre-training significantly improved convergence speed, it did not improve the FROC (see Table 1: 40X *vs.* 40X-pretrained). This may be due to a large domain difference between pathology images and natural scenes in ImageNet, leading to limited transferability. In addition, our large dataset size ($10^7$ patches) and data augmentation may have enabled the training of accurate models without pre-training.

Next, we studied the effect of *model size*. Although we were originally motivated by improved experiment turn-around time, we surprisingly found that slimmed-down Inception architectures with only 3% of the parameters achieved similar performance to the full version (Table 1: 40X *vs.* 40X-small). Thus, we performed the remaining experiments using this smaller model.

We also experimented with a *multi-scale* approach inspired by pathologists' workflow of examining a slide at multiple magnifications to get context. However, we find no performance benefit in combining 40X with an additional input at lower magnification (Fig. 3). However, these combinations output smoother heatmaps (Fig. 4), likely because of translational invariance of the CNN and overlap in adjacent patches. These visual improvements can be deceptive: some of the speckles in the 40X models reveal small non-tumor regions surrounded by tumor.

Figures 1 and 3 highlight the variability in the images. Although the current leading approaches report improvements from *color normalization*, our experi-

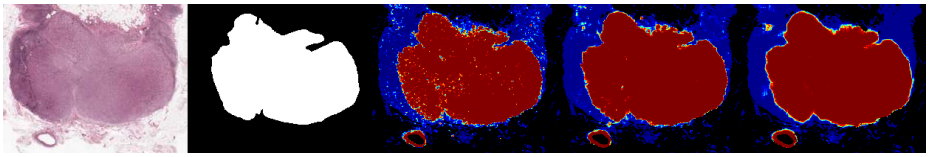

**Fig. 4.** Left to right: sample image, ground truth (tumor in white), and heatmap outputs (40X-ensemble-of-3, 40X+20X, and 40X+10X). Heatmaps of 40X and 40X-ensemble-of-3 look identical. The red circular regions at the bottom left quadrant of the heatmaps are unannotated tumor. Some of the speckles are either out of focus patches on the image or non-tumor patches within a large tumor.

ments revealed no benefit (Supplement). This could be explained by our extensive data augmentations causing our models to learn color-invariant features.

Finally, we experimented with *ensembling* models in two ways. First, averaging over predictions across the 8 rotations/flips yielded a few percent improvement in the metrics. Second, ensembling across independently trained models yield additional but smaller improvements, and gave diminishing returns after 3 models.

**Additional Validation** We also tested our models on another 110 slides that were digitized on different scanners, from different patients, and treated with different tissue preparation protocols. Encouragingly, we obtained an AUC of **97.6 (93.6, 100)**, on-par with our Camelyon16 test set performance.

**Qualitative Evaluation** We discovered tumors in two "normal" slides: 086 and 144. Fortunately, the challenge organizers confirmed that both were data processing errors, and the patients were unaffected. Remarkably, both slides were in our training set, suggesting that our model was relatively resilient to label noise. In addition, we discovered an additional 7 tumor slides with incomplete annotations: 5 in train, 2 in validation (Supplement).

**Limitations** Our errors were related to out-of-focus tissues (macrophages, germinal centers, stroma), and tissue preparation artifacts. These errors could be reduced by better scanning quality, tissue preparation, and more comprehensive labels for different tissue types. In addition, we were unable to exhaustively tune our hyperparameters owing to the near-perfect FROC and AUC on our validation set. We plan to further develop our work on larger datasets.

## 5 Conclusion

Our method yields state-of-the-art sensitivity on the challenging task of detecting small tumors in gigapixel pathology slides, reducing the false negative rate to a quarter of a pathologist and less than half of the previous best result. We further achieve pathologist-level slide-level AUCs in two independent test sets. Our method could improve accuracy and consistency of evaluating breast cancer cases, and potentially improve patient outcomes. Future work will focus on improvements utilizing larger datasets.

## References


1. Camelyon 2016. https://camelyon16.grand-challenge.org/, accessed: 2017-01-17
2. Abadi, M., et al.: TensorFlow (2015)
3. Apple, S.K.: Sentinel lymph node in breast cancer: Review article from a pathologists point of view. J. of Pathol. and Transl. Medicine 50(2), 83 (2016)
4. Bejnordi, B.E., et al.: Stain specific standardization of whole-slide histopathological images. IEEE Trans. on Medical Imaging 35(2), 404–415 (2016)
5. Bunch, P.C., et al.: A free response approach to the measurement and characterization of radiographic observer performance. Appl. of Opt. Instrum. in Medicine VI pp. 124–135 (1977)
6. Cireşan, D.C., et al.: Mitosis detection in breast cancer histology images with deep neural networks. Int. Conf. on Medical Image Comput. and Comput. Interv. (2013)
7. Cruz-Roa, A., et al.: Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. SPIE medical imaging (2014)
8. Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Comput. Vis. and Pattern Recognit. (2014)
9. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J. of the Am. Medical Soc. 316(22), 2402–2410 (2016)
10. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143(1), 29–36 (1982)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. Int. Conf. on Machine Learning (2015)
12. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J. of Pathol. Informatics 7 (2016)
13. Kothari, S., et al.: Pathology imaging informatics for quantitative analysis of whole-slide images. J. of the Am. Medical Informatics Assoc. 20(6), 1099–1108 (2013)
14. Krizhevsky, A., et al.: Imagenet classification with deep convolutional neural networks. Adv. in Neural Inf. Process. Syst. pp. 1097–1105 (2012)
15. van der Laak, J.A., et al.: Hue-saturation-density model for stain recognition in digital images from transmitted light microscopy. Cytometry 39(4), 275–284 (2000)
16. Litjens, G., et al.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci. Reports 6 (2016)
17. Long, J., et al.: Fully convolutional networks for semantic segmentation (2015)
18. Pitié, F., Kokaram, A.: The linear monge-kantorovitch linear colour mapping for example-based colour transfer (2007)
19. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. of Comput. Vis. 115(3), 211–252 (2015)
20. Szegedy, C., et al.: Going deeper with convolutions. Comput. Vis. and Pattern Recognit. (2015)
21. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude (2012)
22. Veta, M., et al.: Assessment of algorithms for mitosis detection in breast cancer histopathology images. Medical image analysis 20(1), 237–248 (2015)
23. Wang, D., et al.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)
24. Yosinski, J., et al.: How transferable are features in deep neural networks? Adv. in Neural Inf. Process. Syst. (2014)
25. Yu, K.H., et al.: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. 7 (2016)

# 6 Supplement

## 6.1 Dataset Details

| | Number of Slides | | | Number of Patches (M) | | | Number of Tumors | |
|---|---|---|---|---|---|---|---|---|
| Dataset/split | Normal | Tumor | Total | Normal | Tumor | Total | Macro | Micro |
| Camelyon-Train | 127 | 88 | 215 | 13+8.9* | 0.87 | 23 | 81 | 345 |
| Camelyon-Validation | 32 | 22 | 54 | 3.8+2.3* | 0.28 | 6.4 | 14 | 58 |
| Camelyon-Test | 80 | 50 | 130 | | | | 40 | 185 |
| NHO-1 $\star$ | 53 | 57 | 110 | | | | | |

**Table 2.** Number of slides, patches (in millions), and tumors in each dataset/split. We excluded "Normal" slide 144 because preliminary experiments uncovered tumors in this slide. Later experiments also uncovered tumors in "Normal" 086, but this slide was used in training for the results presented in this paper. In addition, Test slide 049 was an accidental duplication by the organizers (Tumor 036), and was not used for evaluation. Tumor sizes: macrometastasis (macro, $> 2000\mu m$), micrometastasis (micro, $> 200$ & $\leq 2000\mu m$). *normal patches extracted from the tumor slides. $\star$: additional evaluation set with slide-level labels only.

## 6.2 Soft Labels

Our experiments used binary labels: a patch is positive if at least one pixel in the center 128 x 128 region is annotated as tumor. We also explored an alternative "soft label" approach in preliminary experiments, assigning as the label the fraction of tumor pixels in the center region. However, we found that the thresholded labels yielded substantially better performance.

## 6.3 Image Color Normalization

As can be seen in Fig. 1 & 3, the (H&E) stained tissue vary significantly in color. These differences arise from differences in the underlying biological tissue, physical and chemical preparation of the slide, and scanner adjustments. Because reducing these variations have improved performances in other automated detection systems [4, 13], we experimented with a similar color normalizing approach. However, we have not found this normalization to improve performance, and thus we detail our approach for reference only. This lack of improvement likely stems from our extensive color perturbations encouraged our models to learn color-insensitive features, and thus the perturbations were unnecessary.

First, we separate color and intensity information by mapping the raw RGB values to a Hue-Saturation-Density (HSD) space [15], and then normalize each component separately. This maps each color channel $(I_R, I_G, I_B) \in [0, 255]^3$ to a corresponding "light density values" via $D_\nu = -\ln((I_\nu + 1)/257), \nu \in \{R, G, B\}$,

followed by applying an HSI color space transformation, with $D = (D_R + D_B + D_G)/3$ being the intensity, and $c_x = \frac{D_R}{D} - 1$ and $c_y = (D_G - D_B)/(\sqrt{3} \cdot D)$ denoting the Cartesian coordinates that span the two-dimensional hue-saturation plane. We chose the HSD mapping over a direct HSI mapping of RGB values [15], because it is more compatible with the image acquisition physics and yields more compact distributions in general.

Next, we fit a single Gaussian to the color coordinates $(c_x, c_y)_i$ of the pixels in all tissue-containing patches, i.e. compute their empirical mean $\mu = (\mu_x, \mu_y)^T$ and covariance $\Sigma \in \mathcal{R}^{2\times2}$, and then determine the transformation $T \in \mathcal{R}^{2\times2}$ of the covariance $\Sigma$ to a reference covariance matrix $\Sigma^R$ using the Monge-Kantorovitch approach presented in [18]: $T = \Sigma^{-1/2}\left(\Sigma^{1/2}\Sigma_R\Sigma^{1/2}\right)\Sigma^{-1/2}$. Subsequently, the color values are normalized by applying the mapping:

$$\begin{bmatrix} c'_x \\ c'_y \end{bmatrix} = T\left(\begin{bmatrix} c_x \\ c_y \end{bmatrix} - \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}\right) + \begin{bmatrix} \mu_x^R \\ \mu_y^R \end{bmatrix}. \tag{1}$$

Intensity values, $D_i$, are normalized in the same manner, i.e. by applying the one-dimensional version of Equation 1 in order to transform the empirical mean and variance of all patch intensities to a reference intensity mean and variance.

As reference means and variances for the color and intensity component, respectively (i.e. $\mu_v^R, \Sigma^R$ for color), we chose the component-wise medians over the corresponding statistical moments of all training slides.

Finally, we map the normalized $(c'_x, c'_y, D')$ values back to RGB space by first applying the inverse HSI transform [15], followed by inverting the non-linear mapping, i.e. by applying $I_\nu = \exp(-D_\nu) \cdot 257 - 1$ to each component $\nu \in \{R, G, B\}$.

We applied this normalization in two ways. First, we applied this at inference only, by testing a model ("40X-small" in Table 1) on color-normalized slides. Unfortunately, this resulted in a few percent drop in FROC. Next we trained two models on color-normalized slides, both with and without the color perturbations. We then tested these models on color-normalized slides. Neither approach improved the performance.

### 6.4 Sample Results

**Tumor slides with incomplete annotations** At the outset, 11 tumor slides were known to have non-exhaustive pixel level annotations: 015, 018, 020, 029, 033, 044, 046, 051, 054, 055, 079, 092, and 095. Thus, we did not use non-tumor patches from these slides as training examples of normal patches. Over the course of our experiments, we discovered several more such cases that we verified with a pathologist: 010, 025, 034, 056, 067, 085, 110.

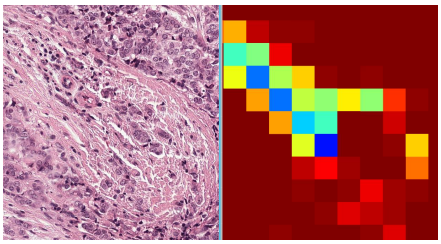

**Fig. 5. Left**: a patch from a H&E-stained slide. The darker regions are tumor, but not the lighter pink regions. **Right**: the corresponding predicted heatmap that accurately the tumor cells while assigning lower probabilities to the non-tumor regions.

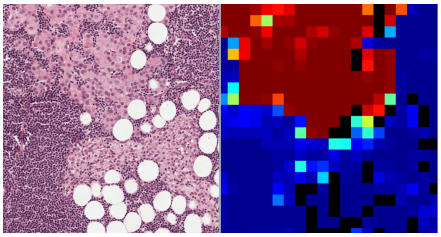

**Fig. 6. Left**: a patch from a H&E-stained slide, "Normal" 086. The larger pink cells near the top are tumor, while the smaller pink cells at the bottom are macrophages, a normal cell. **Right**: the corresponding predicted heatmap that accurately identifies the tumor cells while ignoring the macrophages.