Breast Tumor Cellularity Assessment using Deep Neural Networks

Alexander Rakhlin
Neuromation OU
Tallinn, 10111 Estonia
rakhlin@neuromation.io

Aleksei Tiulpin
University of Oulu
Oulu 90220, Finland
aleksei.tiulpin@oulu.fi

Alexey A. Shvets
MIT
Boston, MA 02142, USA
shvets@mit.edu

Alexandr A. Kalinin
University of Michigan
Ann Arbor, MI 48109, USA
akalinin@umich.edu

Vladimir I. Iglovikov
ODS.ai
San Francisco, CA 94107, USA
iglovikov@gmail.com

Sergey Nikolenko
Neuromation OU, Estonia,
Steklov Mathematical Institute
at St. Petersburg, Russia
sergey@logic.pdmi.ras.ru

Abstract

Breast cancer is one of the main causes of death worldwide. Histopathological cellularity assessment of residual tumors in post-surgical tissues is used to analyze a tumor's response to a therapy. Correct cellularity assessment increases the chances of getting an appropriate treatment and facilitates the patient's survival. In current clinical practice, tumor cellularity is manually estimated by pathologists; this process is tedious and prone to errors or low agreement rates between assessors. In this work, we evaluated three strong novel Deep Learning-based approaches for automatic assessment of tumor cellularity from post-treated breast surgical specimens stained with hematoxylin and eosin. We validated the proposed methods on the BreastPathQ SPIE challenge dataset that consisted of 2395 image patches selected from whole slide images acquired from 64 patients. Compared to expert pathologist scoring, our best performing method yielded the Cohen's kappa coefficient of 0.69 *(vs.* 0.42 *previously known in literature) and the intra-class correlation coefficient of* 0.89 *(vs.* 0.83*). Our results suggest that Deep Learning-based methods have a significant potential to alleviate the burden on pathologists, enhance the diagnostic workflow, and, thereby, facilitate better clinical outcomes in breast cancer treatment.*

1. Introduction

Breast cancer is one of the most common cancer types diagnosed in women in the United States and worldwide [36]. Biopsies and histological assessment allow pathologists to analyze microscopic structures of breast tissues and, in particular, assess the cancer's aggressiveness.

Multiple options are available to manage and monitor the breast cancer treatment based on the information provided from the tumor's response to it. In addition to the treatment effect on the tumor size, the therapy may also alter the tumor's cellularity [8]. During anticancer therapy, the size of the tumor may remain the same, but the overall cellularity may be drastically reduced [30]. As a result, it makes the residual tumor cellularity an important factor in assessing the response treatment.

Currently, tumor cellularity is manually assessed by pathologists from hematoxylin and eosin (H&E)-stained slides [30]. The costs of such estimation are high, the process is tedious and subjective, and the quality and reliability might be also be affected by high inter-observer variability even among senior pathologists. This potentially may affect prognostic power assessment in clinical trials [39]. The subjectivity in visual tissue assessment motivates the use of computer-aided methods to improve the diagnosis accuracy, reduce human error and increase inter-observer agreement and reproducibility [27, 11]. Automated analysis of the H&E slide using computer vision could provide immediate benefits to patient care. Recent success in Deep Learning (DL) [22, 34], and in particular the advances in convolutional neural networks (CNN), have recently shown high potential in this realm [9].

In this work, we evaluate three DL-methods to score the cellularity of the breast tissue from histopathological images. In particular, our first approach employs a weakly-supervised segmentation model with Resnet-34 [13] encoder and Feature Pyramid Network (FPN)[24] and a second-stage regression network that predicts the cellularity score using the predicted segmentation maps. Our second approach is also based on segmentation, however,

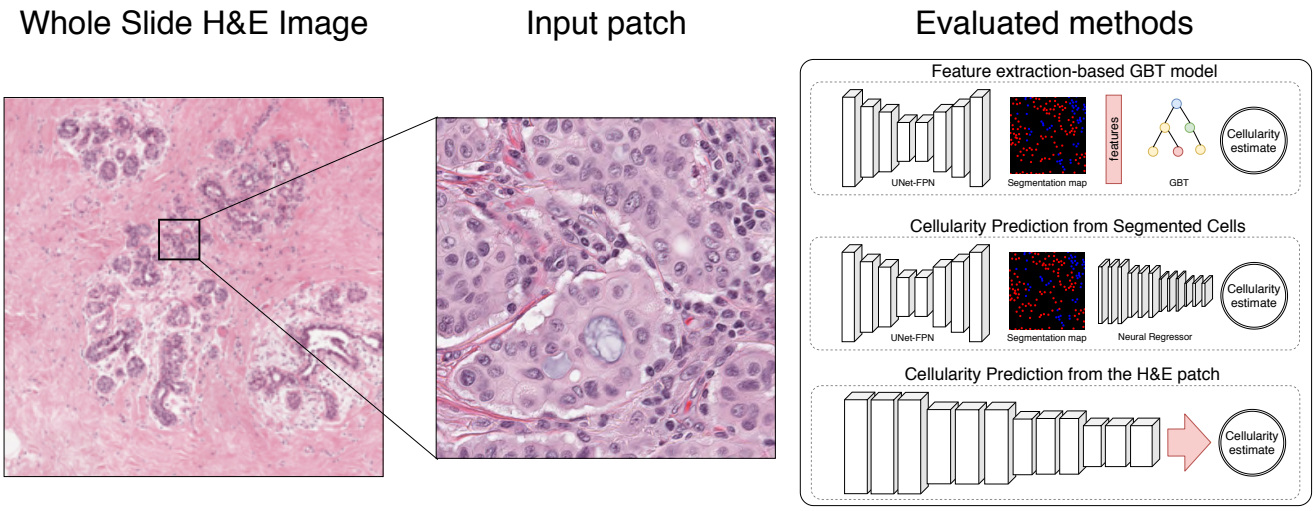


Figure 1: Generic description of the methods developed and evaluated in this study. Our first approach leverages segmentation model, feature extraction and gradient boosted trees. The second approach directly predicts the cellularity from the raw data. Finally, in our third setting, we combine the first and the second approach and used a deep convolutional neural network to predict the cellularity score from segmentation mask.

instead of using the segmentation maps directly, we extract various features from them and use the gradient boosting trees (GBT) [28] to predict the cellularity score. Finally, we also evaluate using H&E image patches directly to predict the cellularity score.

2. Related work

CNNs have recently been successfully applied to many tasks in biomedical image analysis, often outperforming conventional machine learning methods [9, 41, 35]. As such, they have successfully been utilized for digital pathology image analysis and have demonstrated great potential for improving breast cancer diagnostics [38, 4, 5, 32].

Although there are not many studies focusing directly on automated quantitative cellularity assessment, it has been shown that this task can be solved by first segmenting malignant cells and then computing the tumor's area [29]. Many efforts have been devoted to developing supervised and unsupervised methods for automated cell and nuclear segmentation and detection [44, 21]. Supervised segmentation models have superior performance but require hand-labeled nuclear mask annotations [44]. In these approaches, segmented nuclear bodies are used to extract features that are typically inspired by visual markers recognized by pathologists. Commonly used features describe morphology, texture, and spatial relationships among cell nuclei in tissue [29, 19].

The conventional approach most relevant to our work is by Peikari *et al.* [29] who proposed an automated cellularity assessment protocol. First, they used smaller patches, or regions of interest (RoI), extracted from whole slide images to segment all present cell nuclei. Then they extracted a number of predefined features from segmented nuclei and used support vector machines to distinguish lymphocytes and normal epithelial nuclei from malignant ones. Cellularity estimation was done using distinguished malignant epithelial figures for every RoI.

Alternatively, segmentation-free methods that directly estimate cellularity from histopathology imaging data and nuclei locations annotated by human observers were also shown promising. In particular, Veta *et al.* [43] proposed a deep learning-based method that leverages an information from a tumor's cells nuclei locations (centroids) and predicts the areas of individual nuclei and mean nuclear area without the intermediate step of nuclei segmentation. In particular, this approach was based on a 10-layer deep neural network predicting nuclear areas quantized into 20 histogram bins. The results showed that predicted measurements had substantial agreement with manual measurements, which suggests that it is possible to compute the areas directly from imaging data, without the intermediate step of nuclei segmentation. This is in spirit similar to one of our approaches, but we do not directly compare our methods to Veta *et al.* since we use different datasets and performance metrics.

Recent works by Akbar *et al.* [3, 2] have compared the conventional approach based on segmentation and feature extraction and direct applications of deep CNNs to image patches in both regression and classification settings. Overall, they showed that the DL-based approach outperformed hand-crafted features in both accuracy and intra-

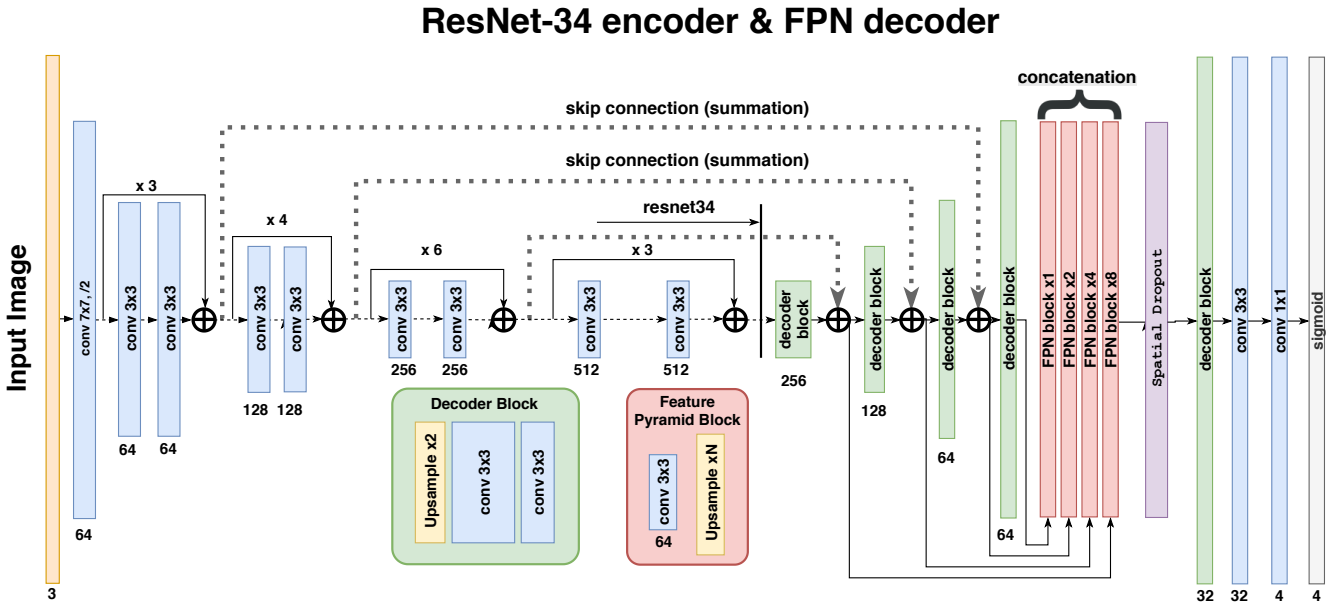


Figure 2: Encoder-decoder segmentation network architecture with Resnet-34 encoder and feature pyramid network decoder. Spatial Dropout 2D is added after multi-layer concatenation.

class correlation (ICC) with expert pathologist annotations. Specifically, their best result was achieved by using a pre-trained Inception [40] model that reached ICC of 0.83 and 0.81 with two expert pathologists. In this study we evaluate even wider range of DL-based approaches, including segmentation-based and segmentation-free, in both regression and classification settings. We provide appropriate performance comparisons with previously reported results[1]. All the methods developed in this study are fully automatic and do not require any involvement of the human annotators at the test time.

3. Methods

In this study, we propose and evaluate three different methods. The first two methods are based on the nuclei segmentation and the third method leverages the raw image without preceding segmentation step. Graphical illustration of our approach is presented in Figure 1

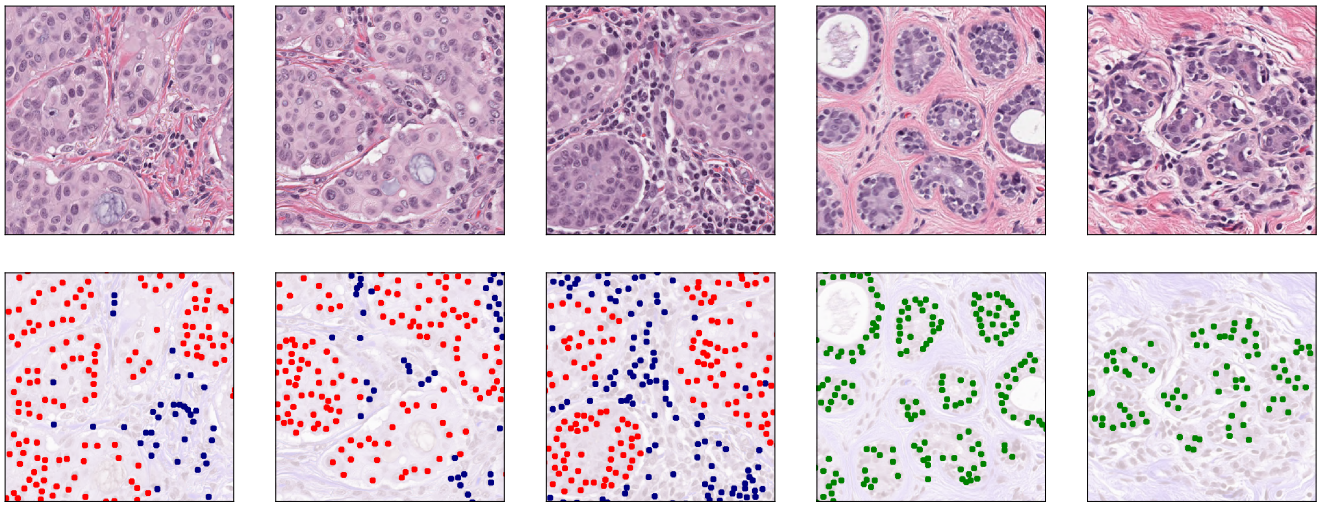
3.1. Segmentation

Network Architecture. Most modern segmentation architectures inherit the encoder-decoder architecture similar to U-Net [33], where convolutional layers in the contracting branch (encoder) are followed by an upsampling branch that brings segmentation back to the original image size (decoder). In addition, skip connections are used between contracting and upsampling modules to help the localization information propagate through the complex multilayer structure and eventually improve segmentation accuracy [33]. U-Net and architectures inspired by this idea have produced state of the art results in various segmentation problems, and many improvements for the architecture and its training protocols have recently been proposed. In particular, Iglovikov *et al.* [15] used batch normalization [17] and exponential linear unit (ELU) as the primary activation function and an ImageNet pre-trained VGG-11 network [37] as an encoder. Liu *et al.* [25] proposed an hourglass-shaped network (HSN) with residual connections, which is also very similar to the U-Net architecture. Rakhlin *et al.* [31] used the Resnet-34 network [14] as the encoder and the Lovász-Softmax loss function [6] along with Stochastic Weight Averaging (SWA) [18] for training.

In our proposed architecture, the segmentation module also inherits the U-Net architecture. The contracting branch (encoder) of our model is based on the Resnet-34 [14] network architecture where we have introduced several useful modifications. In particular, we have replaced ReLU activations with ELU that does not saturate gradients and keeps the output close to zero mean and have changed order of batch normalization [17] and activation layers. In Section 4 we compare encoders initialized with random *He's initialization* [12] and pretrained on ImageNet.

To address the limited size of the BreastPathQ Cancer Cellularity Challenge dataset, we utilized two regulariza-

[1]It is worth noting that the official BrestPathQ challenge results have been reported only as a score distribution. Each team know their own results only, and ours belong to the right end of the distribution, but, unfortunately, we are not able to provide a comparison of our approach with other participants. http://spiechallenges.cloudapp.net/competitions/14#learn_the_details



(a) (b) (c) (d) (e)

Figure 3: Light micrograph of a histologic specimen of breast tissue stained with hematoxylin and eosin (top). The bottom row shows nuclei segmentation masks synthesized from weak labels: *Malignant* — red, *Normal* — green, *Lymphocyte* — blue.

tion techniques: (1) data augmentation and (2) spatial 2D dropout incorporated into the upsampling branch [42].

The upsampling branch is implemented as a Feature Pyramid Network (FPN) [23], reconstructing high-level semantic feature maps at four scales simultaneously. We implement a feature pyramid block as a convolutional layer with 64 activation maps followed by upsampling to the original resolution with upsampling rate of 8, 4, 2, or 1 depending on the feature map depth (see Fig. 2). In Section 4, we compare the performance of standard and FPN decoders.

We concatenate upsampled maps into a single layer of $64 \times 4 = 256$ maps and add after it a spatial 2D dropout layer, which acts as a regularizer and prevents coadaptation of the network weights, but unlike conventional dropout it drops out not individual neurons but rather entire activation maps. Throughout the work, we use dropout rate 0.5, randomly dropping 128 out of 256 activation maps.

Finally, the output of the model is a 4-channel sigmoid layer that assigns every pixel with four values from 0 to 1 that represent the probabilities of belonging to the *Normal*, *Lymphocyte*, *Malignant*, and *Background* classes.

Loss functions. Binary cross entropy (BCE), while convenient for training, does not directly translate into Jaccard index, the metric commonly used to evaluate segmentation accuracy. Hence, as the loss function we use

$$L^c(w) = (1-\alpha)\mathrm{BCE}^c(w) - \alpha J^c(w), \qquad (1)$$

a weighted sum of BCE and the soft Jaccard loss for class c [15, 31, 16]. In this work, we set $\alpha = 0.15$, a value found via cross-validation. The soft Jaccard loss is defined as

$$J^c(w) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i^c \hat{y}_i^c}{y_i^c + \hat{y}_i^c - y_i^c \hat{y}_i^c}\right), \qquad (2)$$

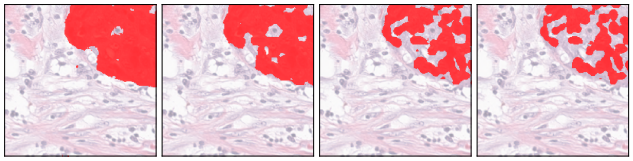
where w are network parameters, y_i^c is the binary label for pixel i and class c, $\hat{y}_i^c$ is the predicted probability of c for pixel i, and N is the total number of pixels. The total loss function is a weighted sum of class losses:

$$L(w) = \frac{1}{V}\sum_{c=1}^{4} L^c(w) v^c, V = \sum_{c=1}^{4} v^c, \qquad (3)$$

where v^c is a loss weight for class c. In this work we weigh *Normal*, *Lymphocyte*, and *Background* as 1 and *Malignant*, the class of primary importance in our problem, as 4.

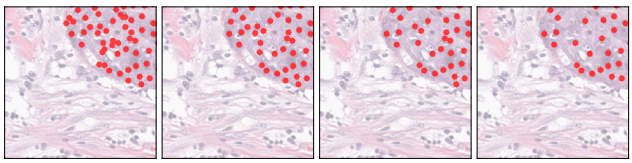
3.2. Cellularity estimation from segmented cells

In this subsection, we describe the method for cellularity assessment that leverages the output of the trained segmentation network Fig. 2. We feed the segmented output into a Resnet-34 CNN model. The model automatically learns deep features from the 4-channel segmentation input and regresses it onto continuous cellularity score using continuous regression loss (L_2). In this approach, the segmentation model acts as a filter the aim of which is to extract only the information about the cell morphology. We hypothesized that this structured approach makes our method similar to methods employed by expert pathologists, makes it transparent and less sensitive to data acquisition settings.



(a) 0.04 (b) 0.16 (c) 0.32 (d) 0.50

Figure 4: Segmentation results at thresholds (a) - (d) of *Malignant* channel superimposed with the original image. Masks generated after thresholding were used for feature extraction.



(a) 0.02 (b) 0.08 (c) 0.16 (d) 0.24

Figure 5: Nuclei blobs detected from the *Malignant* segmentation maps using the Laplacian of Gaussian method at thresholds (a)-(d). The blobs were used for feature extraction.

3.3. Feature extraction-based cellularity estimation

The second type of model is Gradient Boosted Trees (GBT) [20] in regression mode (L_2 loss). The general idea and handcrafted features are borrowed from the second place solution for 2017 Kaggle contest for Sea Lion Population Count in aerial imaginary [26]. The authors would like to thank Konstantin Lopuhin for valuable discussion they had while incorporating his method. In this study, GBT operates on a vector of hand-crafted features extracted from nuclei segmentation maps, including:

- activations and their areas aggregated over segmentation maps with different thresholds; for every segmentation map in *Normal*, *Lymphocyte*, *Malignant* and for 7 thresholds 0.02, 0.04, 0.08, 0.16, 0.24, 0.32, 0.5, we obtain 2 values: total area above threshold and total activation above threshold (see Fig. 4 for an illustration);
- using the Laplacian of Gaussian (LoG) method as implemented in the OpenCV library [7], we find blobs in segmentation maps at 6 thresholds: 0.02, 0.04, 0.08, 0.16, 0.24, 0.5; for each threshold we find the number of blobs and total activation in blob centers (Fig. 5);
- total activation for every channel, computed as a sum of the activations at every pixel after sigmoid.

In total, we obtain $3 \times (7 \times 2 + 6 \times 2 + 1) = 81$ features to train the GBT model.

3.4. Cellularity estimation from the raw images

The third type of model is a deep convolutional network implemented in regression or classification settings (L_2 or categorical cross-entropy loss functions respectively). These models do not use intermediate segmentation and predict the cellularity score immediately from the microscopic image. For classification, we categorize cellularity into 101 class using regular bins with thresholds $0.00, 0.01, \ldots, 1.00$. The idea of direct regression of an image into continuous value using CNN is not new. In particular, it was implemented in [16] where the authors use CNN to predict bone age from radiograph.

3.5. Evaluation metrics

We assessed the results using several metrics. The main evaluation metric is the mean squared error (MSE) between the cellularity score obtained in our experiments and ground truth provided by an expert pathologist.

In order to make our results comparable with previous work, we also report Cohen's kappa coefficient agreement and the intra-class correlation coefficient (ICC) between expert and automated methods, similar to [29]. In all experiments, we find our results superior to our predecessors; however, the cellularity score itself in [29] is evaluated based on binning it into four categories of 0–25%, 26–50%, 51–75%, and 76–100%. Such 4-class categorization is relatively coarse and, in our opinion, does not represent a suitable evaluation metric for continuous cellularity estimation that is our goal in this work.

4. Experiments and results

4.1. Data

The data used in this study had been acquired from the Sunnybrook Health Sciences Centre with funding from the Canadian Cancer Society and was made available for the BreastPathQ challenge sponsored by the SPIE, NCI/NIH, AAPM, and the Sunnybrook Research Institute [29].

In our experiments we used $2,395$ patches of 512×512 pixels in size, extracted from 96 haematoxylin and eosin (H&E) stained whole slide images (WSI) acquired from 64 patients. Each patch in the training set has been assigned a tumor cellularity score by an expert pathologist. In Figure 6, we present a distribution of the cellularity scores in the dataset.

Besides the image data, we used the annotations (X and Y coordinates) to identify lymphocytes, malignant epithelial, and normal epithelial cell nuclei in the additional 153 patches. Using these weak annotations, we generated the segmentation masks that were used in our experiments. Here, at each XY location, we simply fit a blob of 15 pixels in diameter. In Figure 3 we present the generated masks for various classes.

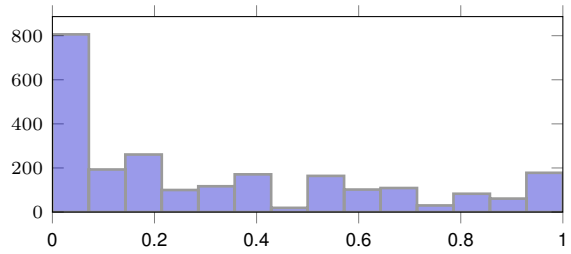


Figure 6: Cellularity score distribution.

4.2. Weakly-supervised cell segmentation

In this study, characteristic features of the data present a serious challenge for developing segmentation models: 1) the data features no segmented cells, only their coordinates (that is why we use semi-supervised segmentation); 2) annotated nuclei are present only in 154 microscopic images, each containing 0-50 malignant cells. However, cell segmentation is not a distinct goal of this study. As mentioned in Section 1 and in [8, 30, 10], cellularity within the tumor area is assessed by estimating the percentage area of the overall tumor bed comprised of invasive tumor cells. Aggregated area of individual invasive cell areas serves as a proxy and does not represent the ultimate cellularity value. Cellularity is affected by cell density, localization, and tissue structure. We use segmented cells essentially as an interpretable visualization of an invasive tumor within the tumor bed.

In our ablation studies, we evaluated our model in four different settings to find how different design choices influence the segmentation accuracy and generalization. Namely, we compared the model as described in Section 3 with a standard U-Net decoder against an FPN decoder, and with the encoder initialized randomly against the encoder initialized with weights pretrained on ImageNet. In all settings, the model was trained for 150 epochs with the Adam optimizer and gradually decreasing learning rate from 10^{-4} to 10^{-5}.

To obtain training patches, we downscaled the microscopy images $\times 2$ times, randomly cropped a 256×256 area, and rescaled pixel values from $[0, 255]$ to $[-1, 1]$. As mentioned previously, segmentation targets were generated as 4-channel masks with round blobs, 15 pixels in diameter (the characteristic nucleus size), drawn in the nuclei centers. During training, we dynamically augmented images with vertical and horizontal flips, rotation, gamma, hue, and saturation utilizing the *Albumentations* library [1].

In the first series of experiments, we evaluated segmentation quality as an important intermediate metric for the evaluation of our methods. The segmentation performance as a function of the decoder and initialization is shown in Table 1. As we can see, the model with the feature pyramid decoder and encoder pretrained on ImageNet achieved significantly higher and more stable Jaccard index on the validation set than the alternatives. Figure 7 shows an example of generated segmentation masks in the Malignant channel and nuclei blobs reconstructed with the Laplacian of Gaussian method.

Table 1: Segmentation results: the Jaccard index for different decoders and initializations.

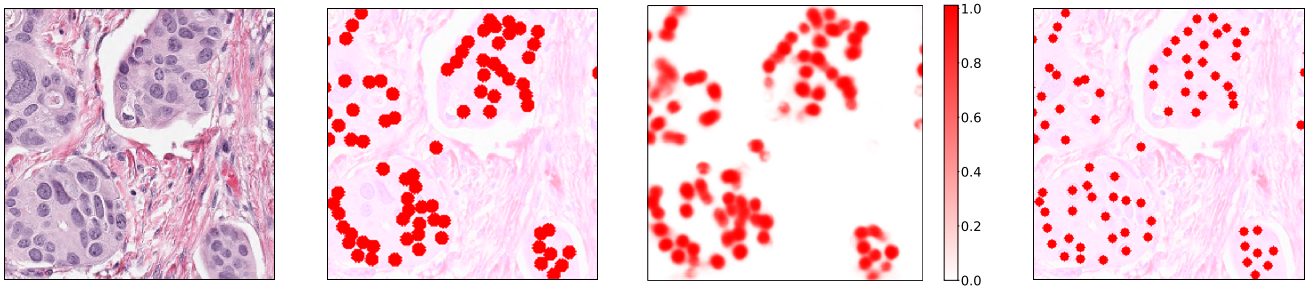
Initialization	Standard decoder	FPN decoder
Random	0.35	0.47
ImageNet	0.50	**0.53**

Table 2: Cellularity MSE with 95% confidence intervals for the segmentation-based (first row) and for the end-to-end methods. Our results demonstrate the importance of ImageNet pre-training. C in the parentheses indicates classification, R – regression and S – segmentation.

Model	Initialization	
	Random	ImageNet
GBT	0.023 [0.019-0.026]	0.022 [0.019-0.026]
Resnet34 (SR)	0.013 [0.011-0.015]	0.013 [0.011-0.015]
ResNet34 (R)	0.015 [0.013-0.018]	0.011 [0.010-0.012]
ResNet50 (R)	0.025 [0.022-0.028]	0.011 [0.009-0.012]
Xception (R)	0.017 [0.015-0.020]	**0.010 [0.009-0.012]**
Xception (C)	**0.012 [0.010-0.014]**	**0.010 [0.009-0.012]**

Table 3: Cellularity Kappa (4 class binning) and Intra-Class Correlation Coefficient (ICC) with 95% confidence intervals for the segmentation-based (1^{st} and 2^{nd} rows) and for the methods predicting cellularity directly, without segmentation. All the models here utilize ImageNet pre-training. C in the parentheses indicates classification, R – regression and S – segmentation.

Model	Metric	
	Kappa	ICC
GBT	0.571 [0.520-0.622]	0.787 [0.744-0.823]
Resnet34 (SR)	0.658 [0.604-0.704]	0.865 [0.835-0.891]
ResNet34 (R)	0.649 [0.599-0.700]	0.868 [0.840-0.892]
ResNet50 (R)	0.652 [0.603-0.701]	0.867 [0.844-0.894]
Xception (R)	0.669 [0.616-0.713]	0.881 [0.853-0.904]
Xception (C)	**0.689 [0.642-0.734]**	**0.883 [0.858-0.905]**
Peikari et al. [29]	0.38-0.42	0.75 [0.71-0.79]
Akbar et al. [3]	—	0.83 [0.79-0.86]



(a) Image (b) Mask (c) Prediction (d) LoG

Figure 7: Examples of the generated segmentation masks in the *Malignant* channel. Left to right: (a) original patch; (b) ground truth segmentation superimposed on the original image; (c) activation map; (d) nuclei blobs reconstructed from the activation map with the LoG method.

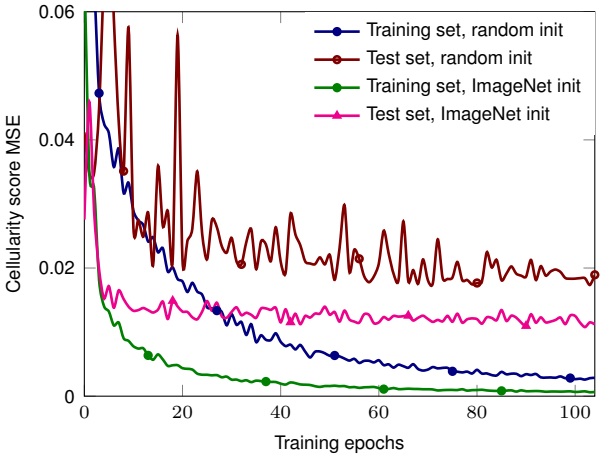


Figure 8: Cellularity MSE evolution during training.

4.3. Segmentation-based cellularity assessment

Prediction from the segmented cells. As mentioned previously, we used the output of the segmentation model as input for the cellularity regressor and then trained this cascade end-to-end. We froze the segmentation model and stack its 4-channel output with a randomly initialized Resnet-34 in the regression setting. We trained the regression part with cellularity targets and MSE loss until convergence. Then we unfroze segmentation weights and fine-tuned both modules in an end-to-end fashion, as a single model. We repeated this experiment with Resnet-34 pretrained on ImageNet. In the latter case, we excluded the background channel from segmentation output to comply with the vanilla Resnet-34 architecture that has a 3-channel input.

In these experiments, we found that after fine-tuning the accuracy of segmentation itself slightly decreases, while the accuracy of the overall cellularity scoring increases. This is in line with [10], which found that perfect segmentation of nuclei figures does not ensure better classification of malignant objects from breast cancer tissues. This finding suggests that the two branches of future work, tumor bed segmentation and cellularity assessment, are relatively independent.

Feature extraction-based method. In this series of experiments, we extracted the 81 features from segmentation masks as discussed in Section 3 and trained the LightGBM [20] regression model with mean squared error (MSE) objective. The model was trained for 600 epochs with learning rate 0.01. The maximum tree depth was set to 5; the number of leaves, to 8. These parameters have been selected through cross-validation.

We report LightGBM accuracy in Table 2 and show the resulting feature importance on Fig. 9. Feature importance was calculated based on the total gain of the loss function from the splits formed according to this feature. As expected, all highly important features come from the *Malignant* channel. The most important feature is the total activation above 0.5 threshold, and the second and third most important features are the activations above 0.32 and 0.24 thresholds, as expected since activations at different theshholds are highly correlated, and the segmentation quality at threshold 0.5 was the best, so the feature based on this mask is a natural candidate for the most important feature. Activations at lower thresholds provide additional value, but a big part of the information that they contain has already been conveyed via the 0.5 threshold feature. Interestingly, malignant cell count (detected at threshold 0.24) is only the 9th feature in order of importance.

4.4. Direct cellularity assessment from the raw images

In our final experiments, we evaluated several deep neural architectures that take the original microscopy images

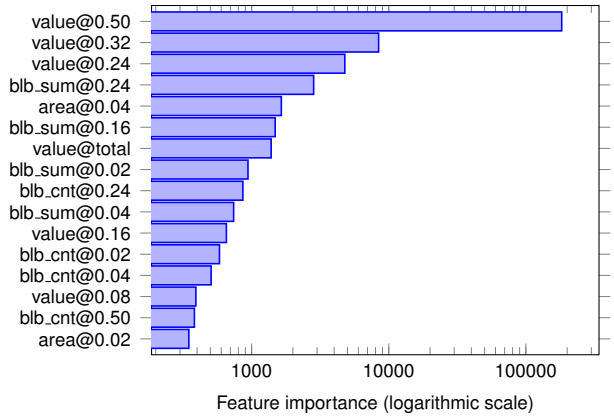


Figure 9: GBT top feature importance.

as input and output the cellularity score without intermediate segmentation. Similarly to previous experiments, we trained the models with random *He initialization* [12] or initialized them with weights pretrained on ImageNet. In all cases, ImageNet initialization was superior to random, and the overall accuracy was slightly better than for the models with intermediate segmentation. The Xception model implemented in a *classification* setup with random initialization performed slightly better than its counterparts (MSE 0.012 vs. 0.017-0.025). Although small, this difference could possibly be attributed to the known regression-to-mean problem of the continuous regression with L_2 loss (e.g., see a well-explained example for a colourization application in [45]). The mean squared error of cellularity prediction as a function of the training epoch for different initializations is shown in Figure 8. All the performance evaluation metrics are presented in Table 2 and Table 3.

4.5. Discussion

As we can see in Table 2 and Table 3, direct cellularity assessment method slightly outperforms the segmentation-based approach, where the regression module works on top of the segmentation feature extractor. We believe that performance improves due to two main reasons. First, segmentation models were not trained on accurate segmentation masks but rather on approximate masks generated from weakly supervised labels. Second, the cellularity score depends not only on the tumor masks but also on a broader set of features, some of which could be lost during the segmentation step.

While we note the record results of our end-to-end models, we believe that the modular form of the prediction pipeline provides benefits that more than compensate for this small difference in the final score.

The segmentation-based approach has two significant advantages: generalizability and interpretability. In practice, the data used for medical imaging tasks comes from different hospitals and is collected by different hardware. Images may differ in quality, level of noise, color and brightness distributions. In [16], the authors proposed to use segmentation to clean and standardize the data, which helps with overall robustness and performance of various task-specific models.

Better interpretability is achieved by the fact that we can visually verify the quality of the intermediate step, i.e., segmented tumors. Furthermore, the decision trees model allows to estimate the feature importance for every feature based on the information gain. If segmented tumors are correct, and the most informative features make intuitive sense, we obtain additional confidence in our model, which is very important in the medical setting.

5. Conclusion

In this paper, we evaluate three automatic methods to assess the cellulalarity of residual breast tumors in H&E stained samples. Our first method leverages the weakly-supervised segmentation masks as inputs for deep CNN. We believe that this method will be more generalizable and robust towards the data acquisition and easier to interpret.

Our second method that leverages feature extraction from the weakly-supervised segmentation mask yields the highest score among the all previously published feature extraction-based methods [3, 29].

Finally, the third method in this study is an end-to-end approach that predicts the cellularity score without any intermediate segmentation step. Although it is attractive and produces the best results it lacks interpretability of the segmentation-based methods and could perform best due to the dataset bias.

The main limitation of this study is the dataset size and the weak labels for the segmentation model. We think that given a bigger dataset and good quality annotations, segmentation-based approach could produce better results that less deviate from the end-to-end trained models.

Acknowledgements

The work of Sergey Nikolenko was supported by the Russian Foundation for Basic Research grant no. 18-54-74005. The authors thank the anonymous reviewer #2 for the constructive suggestions that helped to improve the article.

References

[1] E. K. V. I. I. A. Buslaev, A. Parinov and A. A. Kalinin. Albumentations: fast and flexible image augmentations. *ArXiv e-prints*, 2018. 6

[2] S. Akbar, M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel. Determining tumor cellularity in digital slides using

resnet. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810U, 2018. 2

[3] S. Akbar, M. Peikari, S. Salama, A. Y. Panah, S. Nofech-Momes, and A. L. Martel. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. *bioRxiv*, page 571190, 2019. 2, 6, 8

[4] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6):e0177544, 2017. 2

[5] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. van der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 2

[6] M. Berman, A. Rannen Triki, and M. B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5

[8] Detailed pathology methods for using residual cancer burden. https://www.mdanderson.org/education-and-research/resources-for-professionals/clinical-tools-and-resources/clinical-calculators/calculators-rcb-pathology-protocol2.pdf. 1, 6

[9] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141), 2018. 1, 2

[10] L. E. Boucheron, B. Manjunath, and N. Harvey. Use of imperfectly segmented nuclei in the classification of histopathology images of breast cancer. pages 666–669, 03 2010. 6, 7

[11] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132, 2015. 1

[12] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3, 8

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[15] V. Iglovikov and A. Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018. 3, 4

[16] V. I. Iglovikov, A. Rakhlin, A. A. Kalinin, and A. A. Shvets. Paediatric bone age assessment using deep convolutional neural networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 300–308. Springer, 2018. 4, 5, 8

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 3

[18] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 3

[19] A. A. Kalinin, A. Allyn-Feuer, A. Ade, G.-V. Fon, W. Meixner, D. Dilworth, S. S. Husain, J. R. de Wett, G. A. Higgins, G. Zheng, et al. 3D shape modeling for cell nuclear morphological analysis and classification. *Scientific Reports*, 8, 2018. 2

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017. 5, 7

[21] D. Komura and S. Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018. 2

[22] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015. 1

[23] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 4

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 1

[25] Y. Liu, D. Minh Nguyen, N. Deligiannis, W. Ding, and A. Munteanu. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sensing*, 9(6):522, 2017. 3

[26] K. Lopuhin. Noaa fisheries steller sea lion population count. https://www.kaggle.com/c/noaa-fisheries-steller-sea-lion-population-count/discussion/35422, 2017, online; accessed April 18, 2019. 5

[27] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B. A. Zehnbauer, K. Lister, and R. Parwaresch. Breast carcinoma malignancy grading by bloom–richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Modern pathology*, 18(8):1067, 2005. 1

[28] A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013. 2

[29] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel. Automatic cellularity assessment from post-treated breast surgical specimens. *Cytometry Part A*, 91(11):1078–1087, 2017. 2, 5, 6, 8

[30] R. Rajan, A. Poniecka, T. L. Smith, Y. Yang, D. Frye, L. Pusztai, D. J. Fiterman, E. Gal-Gombos, G. Whitman,

R. Rouzier, et al. Change in tumor cellularity of breast carcinoma after neoadjuvant chemotherapy as a variable in the pathologic assessment of response. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 100(7):1365–1373, 2004. 1, 6

[31] A. Rakhlin, A. Davydow, and S. Nikolenko. Land cover classification from satellite imagery with u-net and lovász-softmax loss. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 3, 4

[32] S. Robertson, H. Azizpour, K. Smith, and J. Hartman. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*, 2017. 2

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[34] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015. 1

[35] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018. 2

[36] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, 68(1):7–30, 2018. 1

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[38] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2560–2567. IEEE, 2016. 2

[39] W. F. Symmans, F. Peintinger, C. Hatzis, R. Rajan, H. Kuerer, V. Valero, L. Assad, A. Poniecka, B. Hennessy, M. Green, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *Journal of Clinical Oncology*, 25(28):4414–4422, 2007. 1

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[41] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific Reports*, 8:1727, 2018. 2

[42] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015. 4

[43] M. Veta, P. J. Van Diest, and J. P. Pluim. Cutting out the middleman: measuring nuclear area in histopathology slides without segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 632–639. Springer, 2016. 2

[44] F. Xing and L. Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Reviews in Biomedical Engineering*, 9:234–263, 2016. 2

[45] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *Lecture Notes in Computer Science*, page 649666, 2016. 8