Effect of Color-Normalization on Deep Learning Segmentation Models for Tumor-Infiltrating Lymphocytes Scoring Using Breast Cancer Histopathology Images

Arian Arab, Victor Garcia, Shuyue Guan, Brandon D. Gallas, Berkman Sahiner, Nicholas Petrick, Weijie Chen

Division of Imaging, Diagnostics, and Software Reliability
Office of Science and Engineering Laboratories
Center for Devices and Radiologic Health
United States Food and Drug Administration, Silver Spring, Maryland, United States
{arian.arab, victor.garcia, shuyue.guan, brandon.gallas, berkman.sahiner, nicholas.petrick, weijie.chen}@fda.hhs.gov

ABSTRACT

Studies have shown that the increased presence of tumor-infiltrating lymphocytes (TILs) is associated with better long-term clinical outcomes and survival, which makes TILs a potentially useful quantitative biomarker. In clinics, pathologists' visual assessment of TILs in biopsies and surgical resections result in a quantitative score (TILs-score). The Tumor-infiltrating lymphocytes in breast cancer (TiGER) challenge is the first public challenge on automated TILs-scoring algorithms using whole slide images of hematoxylin and eosin-stained (H&E) slides of human epidermal growth factor receptor-2 positive (HER2+) and triple-negative breast cancer (TNBC) patients. We participated in the TiGER challenge and developed algorithms for tumor-stroma segmentation, TILs cell detection, and TILs-scoring. The whole slide images in this challenge are from three sources, each with apparent color variations. We hypothesized that color-normalization may improve the cross-source generalizability of our deep learning models. Here, we expand our initial work by implementing a color-normalization technique and investigate its effect on the performance of our segmentation model. We compare the segmentation performance before and after color-normalization by cross validating the models on the three datasets. Our results show a substantial increase in the performance of the segmentation model after color-normalization when trained and tested on different sources. This might potentially improve the model's generalizability and robustness when applied to the external sequestered test set from the TiGER challenge.

Keywords: Deep Learning, Segmentation, Whole Slide Image, Color-Normalization

1. INTRODUCTION

Tumor-infiltrating lymphocytes (TILs) have been shown to be of predictive and prognostic importance in breast cancer patients [1]. In clinics, pathologists' visual assessment of TILs in biopsies and surgical resections of human epidermal growth factor receptor-2 positive (HER2+) and triple-negative breast cancers (TNBC) results in a quantitative score (TILs-score) ranging from 0 to 100 [2]. The TiGER challenge, organized by the Diagnostic Image Analysis Group of the Radboud University Medical Center in collaboration with the International Immuno-Oncology Biomarker Working Group, is the first public challenge for a fully automated assessment of computer-aided TILs-scoring algorithms in hematoxylin and eosin-stained (H&E) breast cancer whole slide images (WSI) [3]. In the TiGER challenge, the test data is reserved as a sequestered dataset where participants only get the performance results through a cloud computation platform without directly accessing the image data. Minimal information is provided to participants on the test data with the aim of studying generalizability and benchmarking different algorithm designs.

We participated in the TiGER challenge by designing an algorithm to segment the tissue into tumor-stroma regions, to detect the location of TILs, and to predict a TILs-score for breast cancer patients. In our algorithm design, the TILs score is calculated as the density of the TILs area within areas of tumor-associated stroma; the density ranges from 0 to 100. During model development, we realized the need for color-normalization to improve the cross-source generalizability of deep learning models. Thus, we are expanding on our initial work by implementing additional color-normalization steps in the model development pipelines to investigate the effects of color-normalization on the generalizability and robustness of the segmentation model we developed.

2. DATA

Whole slide image data and annotations were provided by the organizers of the TiGER challenge. A detailed description of the training data was provided in the TiGER challenge website [3]. For segmentation and detection, 195 WSIs of breast cancer patients were annotated by pathologists in selected regions of interests (ROIs). Annotations consist of masks of different tissue compartments and centroid positions of individual TILs. These WSIs were collected from three sources including 151 TNBC patients from the Cancer Genome Atlas Program (TCGA), 26 TNBC and HER2+ patients from Radboud University Medical Center (TC), and 18 TNBC and HER2+ patients from the Jules Bordet Institute (JB). Table 1 summarizes the distribution of data over the three sources (TCGA, TC, JB). All the slides have a resolution of 0.5 micrometers per pixel. The segmentation masks highlight 6 different tissue compartments (invasive tumor, tumor-associated stroma, in-situ tumor, healthy glands, necrosis not in-situ, and inflamed stroma) with label values from 1 to 6. Anything that does not fall into the above-mentioned categories will be labeled as the "rest" class with the label value of 7. Some parts of the ROIs are left without any annotations, those correspond to the label value of 0. Figure 1 shows example annotations for one ROI. To design a TILs-scoring algorithm, the key areas are the stroma, which includes tumor-associated stroma and inflamed stroma, and the invasive tumor regions. Hence, we focused on these and combined all the others and re-labeled them to 0, 1, 2, as illustrated in Figure 1.

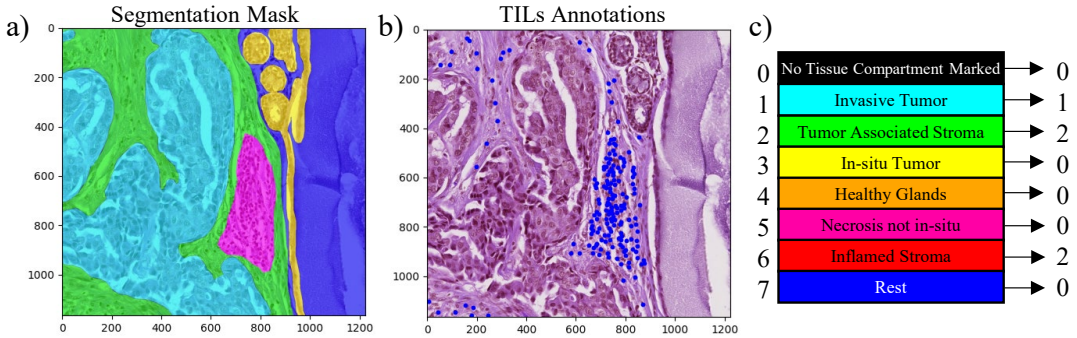

Figure 1. a) Example ROI showing pathologist-annotations of tissue segmentation masks and b) TILs annotations (blue dots). c) The legend shows the merge of annotations to stroma, tumor, and all the others.

Table 1. Distribution of data over the three source sites (TCGA, TC, JB).

	TCGA	TC	JB
No. of WSIs	151	26	18
No. of Segmentation Mask ROIs	151	81	54
Average Segmentation ROIs Size	2200 Pixels	1200 Pixels	1200 Pixels
No. of TILs ROIs	1744	81	54
Total No. of TILs annotated	20727	4728	5523

3. METHODS

To train and internally validate the segmentation model, we first split the WSI data from each source into development and test subsets at the patient level (slide level) by random sampling. After splitting the data, we extracted ROIs from the WSIs in the development dataset and split the ROIs further to training and tuning subsets. Figure 2 sketches this data split strategy. We use the sliding window technique to obtain training and tuning patches from the corresponding ROIs. For training, the patch size is 256×256 pixels with a stride of 128 pixels. For tuning, the patch size is 256×256 with a stride of 256 to ensure that none of the extracted tuning patches overlap. The tuning patches also exclude the boundary patches if they overlap. The extracted training patches are augmented spatially using D4 symmetry of a square (symmetry group of a square). As a result, each training patch is augmented eight times.

We first trained four segmentation models: (1) a model trained using TCGA training patches; (2) a model trained using

TC training patches; (3) a model trained using JB training patches; and (4) a model trained by pooling together all the training patches from these three sources. We then tested each model's performance across the test sets by cross validating the performance on each of the three data sources separately. To better account for the color differences across datasets, we repeated the previous steps by training four additional segmentation models, but this time, we first normalized the color of all the TCGA, TC, and JB slides using the Reinhard color-normalization technique [4]. The Dice score was used to compare the results.

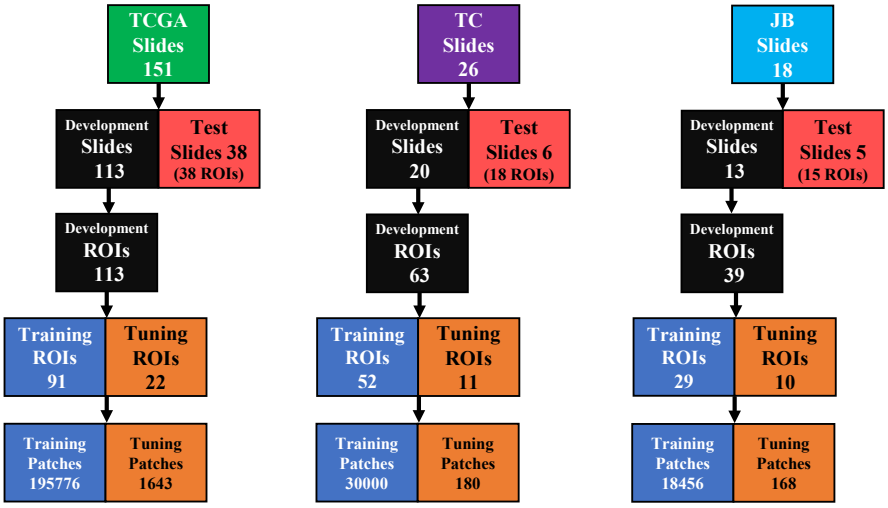

Figure 2. Data split strategy for training, tuning, and test sets. Data is first split at the slide level to create development and test sets. The development ROIs are then further split to create training and tuning sets.

The basic segmentation model was a U-Net with an InceptionV3 backend [5], which uses ImageNet pretrained weights to initialize its weights and biases. We also rescaled the RGB values of the training patches by mapping the range [0, 255] to [-1, 1]. We added a Dropout layer with the rate of 0.3 before an output SoftMax layer to avoid overfitting. A compound loss function of Dice Loss and Categorical Focal Loss is used in model training. ADAM was chosen as the optimizer with a fixed learning rate of 0.0001. The hyperparameters of the model were manually tuned between train/tune cycles. We trained each model for 20 epochs with a batch-size of 32. We then saved the best model weights that gave the highest Dice score on its corresponding tuning patches.

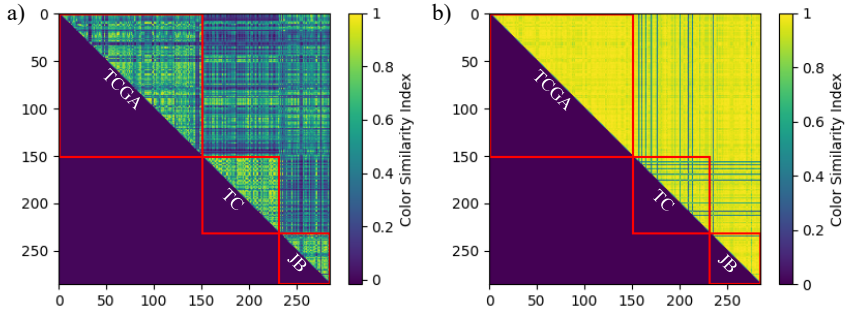

Figure 3. a) Color similarity matrix between all 286 ROIs before color-normalization. b) Color similarity matrix after normalizing all the ROIs using Reinhard color-normalization technique giving the best match between pairs of ROIs.

Color-normalization: A color similarity index between two images can be designed by calculating the correlation and comparing the histograms [6], obtained from the RGB channels of the images. This index quantifies the mismatch between the colors of two images. By calculating the color similarity index between pairs of ROIs from the TCGA, TC, and JB sources, we created a color similarity matrix. As can be seen from Figure 3a, there is a large mismatch between the colors of the ROIs from these three sources. We determined the reference image for Reinhard color-normalization by selecting the ROI which results in a color similarity matrix with the most similarity between the colors of the paired ROIs, Figure 3b.

To implement Reinhard's method [4], we first transferred the source and the target image from RGB color space to LAB color space [7]. We then calculated the mean (μ) and the standard deviation (σ) of the source and the target image in the LAB color space ($\mu_{source}, \sigma_{source}, \mu_{target}, \sigma_{target}$). Using the following equation, we transferred the source image:

$$\text{transferred\_image} = (\text{source\_image} - \mu_{\text{source}}) \frac{\sigma_{\text{target}}}{\sigma_{\text{source}}} + \mu_{\text{target}}$$

We then transformed the transferred image from the LAB color space back to the RGB color space. In Figure 4, three examples' ROIs from TCGA, TC, and JB are given, before and after applying the color-normalization technique.

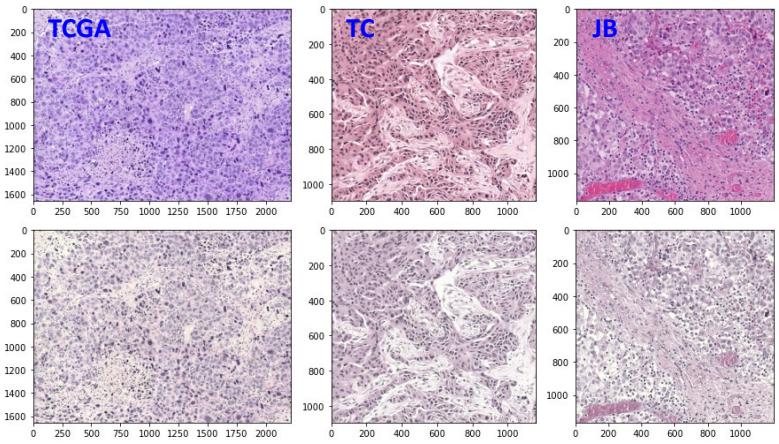

Figure 4. The top row of ROIs are selected ROIs from the three sources (TCGA, TC, JB) before color-normalization. The bottom row of ROIs are the same ROIs after normalizing the color using the Reinhard method.

The performance of our models on the test set was obtained by sliding a window of size 256×256 pixels with stride of 128 pixels across each of the ROIs of the test set. By taking the maximum probability in the overlapped regions of the sliding window, the final probability map was obtained. We then calculated the mean Dice score between the tumor, stroma, and the rest classes.

4. RESULTS

Figure 5a shows the results of the four segmentation models on the test ROIs before color-normalization. There is a substantial drop in performance when a model is tested on data from a different source than that used in model training. For example, the JB model's Dice score on JB test set is 0.67 but reduces to 0.21 and 0.44 when applied to the test data from TC and TCGA, respectively. A similar trend is observed for the TC and TCGA models with the TCGA model showing slightly better generalizability than the TC and JB trained models. Figure 5b shows the results of the four segmentation models on the test ROIs after applying the Reinhard color-normalizing technique. Not only does the overall Dice score increase, but the TC and JB models also generalize better on the remaining test data. These results demonstrate that color-normalization of WSIs of H&E-stained slides scanned by different scanners can improve a deep learning segmentation model's generalizability to unseen data from other sites.

a)

	Test ROIs		
	Mean Dice Score		
ROI Level	TCGA 38	TC 18	JB 15
Trained Models: TCGA	0.78	0.60	0.65
Trained Models: TC	0.42	0.79	0.36
Trained Models: JB	0.44	0.21	0.67
Trained Models: All	**0.78**	**0.79**	**0.71**

b)

	Test ROIs		
	Mean Dice Score		
ROI Level	TCGA 38	TC 18	JB 15
TCGA	0.80	0.60	0.62
TC	0.63	0.82	0.65
JB	0.61	0.76	0.73
All	**0.80**	**0.84**	**0.73**

Figure 5. a) Performance of the 4 models developed using training patches from TCGA, TC, JB, and All patches on the test ROIs. b) Performance of the 4 models developed using training patches from TCGA, TC, JB, and All patches on the test ROIs after Reinhard color-normalization.

5. DISCUSSION AND CONCLUSION

Generalizability is a major concern in deploying deep learning based computational pathology (cPATH) models in the clinic [8]. Color variation of whole slide images from different scanners and clinical sites is one of the important factors impacting generalizability. This study demonstrates that color-normalization is useful in improving the generalizability of deep learning cPATH models. We investigated the effects of Reinhard color-normalization on the deep learning-based segmentation model performance. In our study design, we partitioned our data at the slide level to avoid data leakage between the development dataset and the test dataset. Our cross-source cross-validation testing shows that color-normalization substantially improves the segmentation performance for a model trained on data from one source and tested on another.

In our future work, we will apply the developed models to the external test set sequestered by the challenge organizer on the cloud server to confirm our findings. Further, we will explore data augmentation to enhance the model generalizability to unseen data. We note that the Dice scores were calculated by pooling the confusion matrices (TP, TN, FP) of all the test ROIs together to calculate one overall Dice score, as recommended by the challenge organizer. This method lacks an assessment of the variability of performance over different analysis units (*e.g.*, slides or patients). In a future study, we will analyze performance by averaging the Dice scores of testing slides and calculating the variability in the Dice estimates to better understand performance variability.

6. ACKNOWLEDGMENT

We thank Dr. Ravi Samala for useful discussions. This work was supported in part by the FDA Office of Chief Scientist and the Critical Path program. This project was supported in part by an appointment to the ORISE Research Participation Program at the Center for Devices and Radiological Health, U.S. Food and Drug Administration, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and FDA/CDRH.

7. REFERENCES

1) C. Denkert et al., "Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy", The Lancet Oncology, 19, 2018, 40-50.
2) R. Salgado et al., "The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014", Ann Oncol. 2015 Feb;26(2):259-71
3) https://tiger.grand-challenge.org/
4) Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., “Color transfer between images”, IEEE Computer Graphics and Applications, 2001 21 (5), 34 - 41.
5) qubvel/segmentation_models: Segmentation models with pretrained backbones. Keras and TensorFlow Keras. (github.com)
6) https://docs.opencv.org/3.4/d8/dc8/tutorial_histogram_comparison.html
7) https://docs.opencv.org/3.4/de/d25/imgproc_color_conversions.html
8) Wouter Bulten et. Al., Artifical intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge, Nature Medicine, 28, 154-163 (2022).