Development of Automatic Tumor-Infiltrating Lymphocytes Scoring System for Breast Cancer

Amartya A. Mandal, Ashley Alva, Scott Edwards, Sumeet Gadagkar, May D. Wang *Fellow, IEEE*
Georgia Institute of Technology, Atlanta, GA, United States

***Abstract*—Tumor-infiltrating lymphocytes (TIL), a surrogate for the host immune response against tumor cells, has been identified as a prognostic indicator for the detection of triple-negative and HER2-positive breast cancer. Recent studies have demonstrated that the predictive and prognostic value of visually scored TILs in breast cancer as well as in other types of cancer, making TILs a potential bio-marker for real-world clinical use. However, inter-observer discordance due to the lack of standardization continues to impede the clinical application of TIL. In this paper, we propose an automated TIL-quantifying deep learning algorithm for predicting the TIL scores of digital whole-slide images. The proposed framework consists of three stages: (1) detection of lymphocytes and plasma cells using U-Net, (2) segmentation of invasive tumor and tumor-associated stroma using HookNet, and (3) TIL detection using MaskRCNN. The TIL score is then calculated based on the detected TILs and target tissues. We conduct experiments on TIGER dataset to demonstrate the effectiveness of our proposed scoring system with detailed evaluation of each stage. Moreover, we develop an application prototype with graphic user interface to translate the proposed research to potential clinical use. We believe that a standardized and automated TIL scoring algorithm can improve the prognostic impact of TIL to promote the adoption of quantitative TIL scoring into the cancer staging criteria and real-world clinical practice.**

***Index Terms*—Breast Cancer, Tumor-infiltrating Lymphocytes, Risk Prediction, Image Segmentation, Whole Slide Image**

I. Introduction

By the end of 2020, 7.8 million women had been diagnosed with breast cancer within the previous five years, making breast cancer the most prevalent cancer in the world [1]. For breast cancer patients, molecular sub-type is one of the most important factors need to be considered when defining the treatment strategy with consequences for patient prognosis. Existing studies have shown that Her2-positive and triple-negative breast cancers (TNBC) have the poorest prognosis, highlighting the importance of identifying prognostic and predictive bio-markers for improving patient management and prognosis [2].

In recent years, researchers and clinicians have paid increasing attention to tumor micro-environment (TME) with a particular emphasis on the interaction between tumor cells and the host immune system. Specifically, tumor-infiltrating lymphocytes (TILs), a surrogate for the host immune response against tumor cells, has been proven to be an important bio-marker for cancer patients. In breast cancer, identifying and measuring TILs can improve target treatments for patients, especially immunotherapy. Prior research has also demonstrated a correlation between high TIL densities and a favorable prognosis, specifically in TNBC and Her2-positive breast cancer patients [3]. To improve patient management and prognosis, it is strongly recommended that TILs should be incorporated into standard clinical practice for both early and advanced TNBC and HER2-positive BC [4].

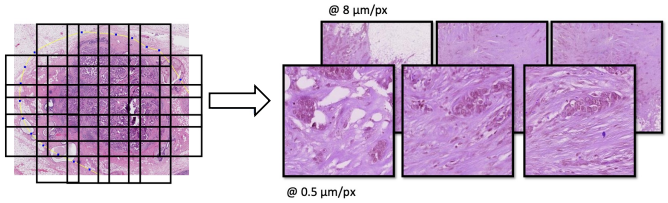

Fig. 1. Example of WSI at different magnifications. The lower magnification usually contains more context or spatial information, while the higher magnification contains more fine-grained and detailed information. Multi-magnification input is designed to incorporate both fine-grained and contextual information in model prediction, analogous to how pathologists zoom in and out for local and global information while examining whole slide images.

One major barrier preventing the adoption of TIL in clinic is the inter-observer discordance due to the lack of standardization. Manual quantification has advantages in pattern recognition, but it has limitations in quantitative assessment and assessing spatial distribution patterns with diagnostic value. Despite the existence of a standardization guideline [5] in clinical practices, observer variability affecting the accuracy and reproducibility of TILs remains a significant barrier to their widespread adoption in research and clinical settings [6]. Consequently, it is becoming increasingly crucial to develop a TIL-quantifying system that can facilitate the analysis of complicated spatial patterns and provide objective metrics for rigorous validation [7].

In recent years, computer-aided image analysis has been investigated on automated cell segmentation and identification tasks to standardize TIL scoring in digitized hematoxylin and eosin (H&E)-stained slides [6], [8]. Specifically, recent advancements in machine learning and deep learning algorithms have been widely applied pathological imaging segmentation and detection task and largely improved the model performance. Acs et al. [6] proposed an automated TIL algorithm for independent prognosis identifier in melanoma using traditional digital imaging analysis methods. Chou et al. [8] established an automated TIL scoring system with machine learning methods improve the prognostic impact of TIL. Comparing with existing work, our proposed framework leverage the advantage of deep learning algorithms to facilitate the pattern recognition and quantification from clinical big

TABLE I
SUMMARY OF THE TOTAL NUMBER OF IMAGES IN EACH SECTION WITH DETAILED DATA SPLIT IN TRAIN, VALIDATION, AND TEST SET.

Dataset	Total	Train	Validation	Test
WSIROIS	195	117	39	39
WSIBULK	93	75	15	3
WSITILS	82	50	16	16

data. Moreover, we develop an application with graphic user interface (GUI) of proposed algorithm to promote the clinical adoption.

In this study, we propose a multi-stage automated TIL-quantifying deep learning framework to predict the TIL scores of digital whole-slide images in breast cancer. First, we develop a deep learning model using U-Net to identify tumor bulk regions in the whole slide image. We then implement a multi-resolution semantic segmentation framework using HookNet to segment invasive tumor and tumor-associated stroma in tissue bulk patches. Lastly, we perform TIL detection using Mask R-CNN to calculate the TIL score. Besides, we develop an application prototype with graphic user interface to translate the proposed research to potential clinical use. Our experiment results on TIGER dataset demonstrate the effectiveness of our proposed scoring system with detailed evaluation of each stage. This approach contributes to the standardized and automated TIL scoring system development to accelerate the adoption in real-world clinical practice.

II. Data Description

We conduct experiment on Tumor InfiltratinG lymphocytes in breast cancER (TIGER) (https://tiger.grand-challenge.org), a public clinical dataset with TILs in H&E breast cancer slides. TIGER dataset contains a total of 195 Her2+ and TNBC breast cancer whole-slide images (WSIs) with manual annotations by a panel of board-certified breast pathologists. WSIs are essentially a container with the same image at different resolutions and each can contain up to $100k \times 100k$ pixels (see Fig. 1). The dataset is divided into three sections that correspond to the multi-stage tasks: (1) whole-slide images with manual annotations in regions of interest (WSIROIS), (2) whole-slide images with coarse manual annotation of the tumor bulk (WSIBULK), and (3) whole-slide images with coarse manual annotation of the tumor bulk (WSITILS).

A. WSIROIS

This section contains 195 WSIs of breast cancer, including biopsies and surgical resections with manually annotated regions of interest (ROI). These annotations consist of both polygons indicating distinct tissue compartments and points indicating lymphocytes and plasma cells. Each ROI has been subdivided into seven regions: (1) invasive tumor, (2) tumor-associated stroma, (3) in situ tumor, (4) inflamed stroma, (5) healthy glands, (6) non-in situ necrosis, and (7) the remainder. Prior research has demonstrated that TILs from regions (1) to (4) have a predictive nature. In addition, these four ROIs contain lymphocyte and plasma cell annotations in the form of bounding boxes. The cells were annotated using point

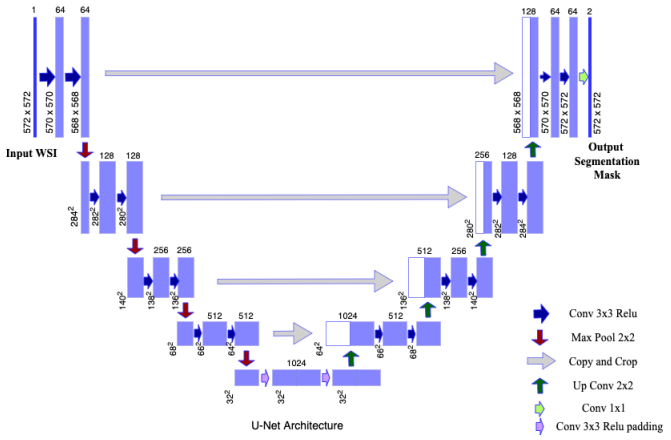

Fig. 2. Overview of U-Net architecture for Stage I.

annotations, and 8×8 micron squared bounding boxes were then constructed by centering on the point annotation. We leverage the expert annotations in this section to train the first-stage deep learning based detection algorithm for lymphocytes and plasma cells.

B. WSIBULK

The WSIBULK section includes 93 WSIs of biopsies and surgical resections of TNBC and Her2+ breast cancer tissues with coarse annotations of the "tumor bulk" region, indicating regions containing invasive tumor cells. The bulk regions of the tumor were manually annotated so that all cancer cells belonging to the invasive portion of the tumor are confined to these regions. This section is used to train the second-stage bulk segmentation model in order to identify the bulk tumor region in the WSI.

C. WSITILs

This section differs from the others in that there are no manual annotations other than one TIL score per slide. It contains 82 WSIs of TNBC and Her2+ cancer tissue biopsies and surgical resections. This section tests the entire multi-stage framework of the TIL-quantification system.

D. Dataset Split

Table I presents a summary of the total number of images in each section, as well as the number of images used for training, validation, and testing.

III. Methodology and System Design

The proposed multi-stage framework for automatic TIL-quantifying system contains three networks: (1) Stage I: U-Net based detection model to extract the tumor bulk region for detection of lymphocytes and plasma cells, (2) Stage II: HookNet based segmentation model for invasive tumor and tumor-associated stroma, and (3) Stage III: Mask R-CNN based quantification model for an automated TILs score per slide, based on the output of detection and segmentation.

A. Stage I: Tumor Bulk Detection

At this stage, we identify the bulk of the tumor in order to detect lymphocytes and plasma cells, which are the primary cell types considered to be TILs. As stated previously, a WSI contains the same image at different resolutions. At the step of tumor bulk detection, the image must be passed at its lowest resolution to reduce computation time and resource consumption.

For tumor bulk region segmentation, we employ U-Net [9], a commonly used convolutional neural network (CNN) in biomedical applications. Specifically, we modify the original U-Net architecture by using padding in the last convolutional layer to maintain the exact shape of the segmented mask as the input image, as shown in the figure 2. The WSIs are resized to 324×324 at a magnification of 8 μm/px from the WSI images. As input to U-Net, the tumor bulk masks are then tiled at both high and low magnifications. The architecture of U-Net contains two paths. The first path is the contraction path (encoder), which captures the context of the input image. The second path is the decoder, which employs transposed convolutions to enable precise localization. Thus, it is an end-to-end fully convolutional network (FCN) that contains only convolutional layers and no dense layers, allowing it to accept images of any size, which improves the robustness for our application given the various tumor bulk size.

In addition, morphological image transformations are performed as a post-processing step to account for discontinuities and remove noise from the predicted masks. The predicted tissue bulk is subjected to image morphological transformations of closing, erosion, and dilation to produce an accurately segmented continuous bulk region. The output of the WSI segmentation and morphological transformations is the mask of the bulk tumor region, which is cropped and fed to the tissue segmentation (Stage II) via the tiling process.

For the training and optimization details, we combine Dice and binary cross-entropy loss as the objective function for U-Net. Similar to the majority of medical imaging datasets, all three datasets suffer from data imbalance problems. To address this issue, we employ weighted Dice loss [10] since it attributes equal weight to false positives and false negatives, which is more resistant to the data-imbalance problem.

B. Stage II: Tissue segmentation

For tissue segmentation, the tissue bulk is segmented into tissues of interest (i.e., invasive tumor and tumor-associated stroma), which are the primary tissue compartments considered when identifying relevant regions for the TILs. HookNet [11] is designed to work with inputs present in multiple resolutions, making it a good candidate for our data. Specifically, one branch takes input at lower magnification with more context or spatial information, named as the context branch; the other branch takes input at a higher magnification concentric to the context input, containing more details due to higher resolution, named as the target branch.

At a suitable point in the encoder-decoder path, the feature maps from the context branch are "hooked" or concatenated to the target branch. In order to guarantee pixel alignment during this hooking, spatial resolution of the feature maps of both context and target branches are used and is defined as $SRF = 2^d r$, where d is the depth in the encoder-decoder model and r is the resolution of the input patch measured in $\mu m/px$. To determine the depth at which hooking is possible, a SRF ratio is defined between a pair of feature maps

$$\frac{SRF_C}{SRF_T} = 2^{d_c - d_T} * \frac{r_C}{r_T}, \tag{1}$$

where d_C and d_T are the depths for the context and target branch, respectively. Hooking can take place when the SRF ratio is 1, which ensures pixel-wise alignment between the two feature maps. The HookNet model then outputs two segmentation maps, one from the context branch and the other from the target branch.

For the training details, each convolutional layer performs a valid 3×3 convolution with stride of one, followed by max-pooling layer with a 2×2 down-sampling factor. For the up-sampling path, nearest-neighbor 2×2 up-scaling is utilized followed by convolutional layers. The central idea behind this model is that fine-grained and contextual information can be combined by concatenating feature maps across branches, similar to how pathologists zoom in and out while examining whole slide images.

For the model optimization, a combined loss function is used to optimize the model by back propagating the loss from both the context branch and the target branch. The Stage-II loss function L_2 is then formulated as $L_2 = \lambda L_T + (1 - \lambda) L_C$, where L_T and L_C are the losses from the target and context branch, respectively, and $\lambda \in [0, 1]$. Specifically, $\lambda = 1$ indicates that the target branch loss dominates the loss function, while $\lambda = 0.5$ means that both are given equal importance.

C. Stage III: TILs Quantification

For the final TIL-quantifying, we implement a Mask R-CNN model [12] with a ResNet-50 backbone. Mask R-CNN is a popular framework for deep learning that generates object bounding boxes, classes, and masks from a given input. First, it generates proposals about the regions where there might be an object based on the input image. Second, it predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on the first stage proposal. Both stages are then connected to the main feature extractor of Mask R-CNN, the backbone structure. Due to the complexity of the images used for TIL detection, ResNet, and more specifically ResNet-50, performs the best in our case. ResNet is also utilized to solve the problem of vanishing gradients by permitting gradients to flow directly from later layers to initial layers through skip connections.

For the training details, the Mask-RCNN architecture allows us to compute three different losses: (1) the classification loss L_{cls} that quantifies the error with respect to the predicted labels of the detected object, (2) the bounding box regression loss L_{box} that measures the error relative to the predicted bounding boxes for detected objects, and lastly, (3) the mask loss L_{mask} that measures the error between the predicted and

TABLE II
SEGMENTATION RESULTS FOR EACH TISSUE CLASS IN STAGE II.

Tissue Class	Dice score
Invasive Tumor	0.63
Tumor Stroma	0.61
In-situ Tumor	0.58
Healthy Glands	0.57
Necrosis	0.69
Inflamed Stroma	0.54
Rest	0.74
Overall (Avg.)	**0.62**

actual segmentation maps. The training objective function L_3 can then be represented as the sum of these losses, where $L_3 = L_{cls} + L_{box} + L_{mask}$.

For the TILs quantification, we compute a TIL score by the definition of the ratio between the area of detected TILs and the area of the target tissues. In the final mask, the invasive tumor, inflamed tissue, tumor-associated stroma, and in-situ tumor masks from the tissue segmentation step are combined. The lymphocyte detection results are compiled into a mask where non-zero entries correspond to the presence of lymphocytes. These masks are then multiplied to produce a binary image containing TILs in the target regions. Both of these mask images are derived from the same layer of image magnification, so their pixel resolutions are both $0.5\mu m/px$. This property enables the use of raw pixels when comparing areas. The final TIL score is defined as:

$$TIL_{score} = \frac{\#NonzeroPixels_{LymphocyteMask}}{\#NonzeroPixels_{TumorTissueMasks}} \quad (2)$$

IV. RESULTS

A. Stage I: Tumor Bulk Detection

The Dice score, also known as the Dice-Sørensen coefficient, is a common evaluation metric for image segmentation that we employ to evaluate our bulk tumor detection results for stage I. As defined in the following Eq. (3), the dice coefficient is a measure of overlap of the predicted mask and the ground truth,

$$Dice_{score} = 2 \times \frac{|A| \cap |B|}{|A| + |B|}, \quad (3)$$

where A is the ground truth mask and B is the predicted mask.

Multiple magnifications and image sizes are utilized to calculate the Dice score for tissue segmentation. The optimal image size and magnification is determined based on these results and the available computational resources. The optimal magnification was determined to be 8μm/px for all image sizes utilized during training. We have achieved a Dice score of 0.74 in the testing set with visualization examples from the test images shown in Fig. 3. In addition, after morphological image transformations, the predicted masks can achieve an accurate tumor bulk region that closely resembles the ground truth masks. An example WSI with all the morphological transformations from the test dataset is shown in Fig. 4.

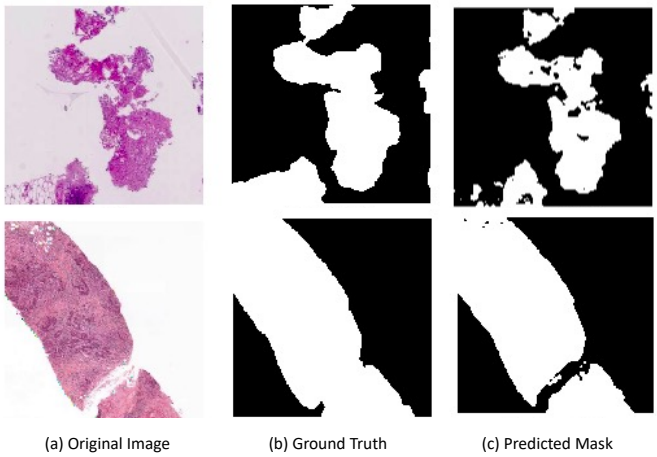

Fig. 3. Examples of tissue segmentation results using U-Net in Stage I.

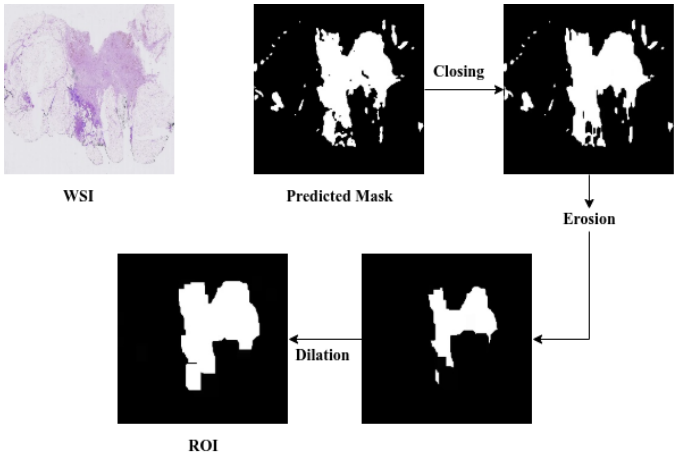

Fig. 4. Examples of morphological image transformation results in Stage I.

B. Stage II: Tissue Segmentation

In Stage II, we use HookNet to present the segmentation of invasive tumor and tumor-associated stroma as the main tissue compartments considered when identifying relevant regions for the TILs. The tissue segmentation result of seven tissue classes is shown in Table II. Meanwhile, we also visualize an example of ground truth and predicted segmentation results from two slides for reference (see Fig. 5).

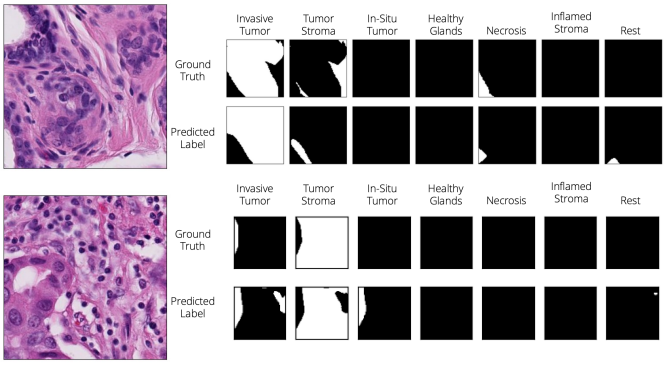

Fig. 5. Examples of tissue segmentation results of seven types in Stage II.

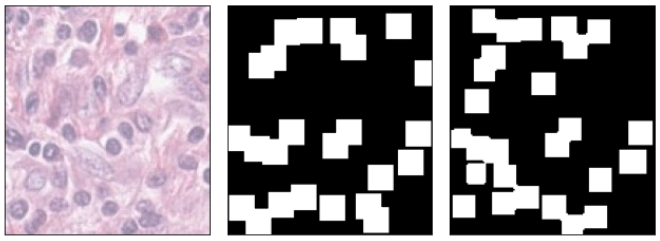

Fig. 6. Example for detection results of TILs in Stage III.

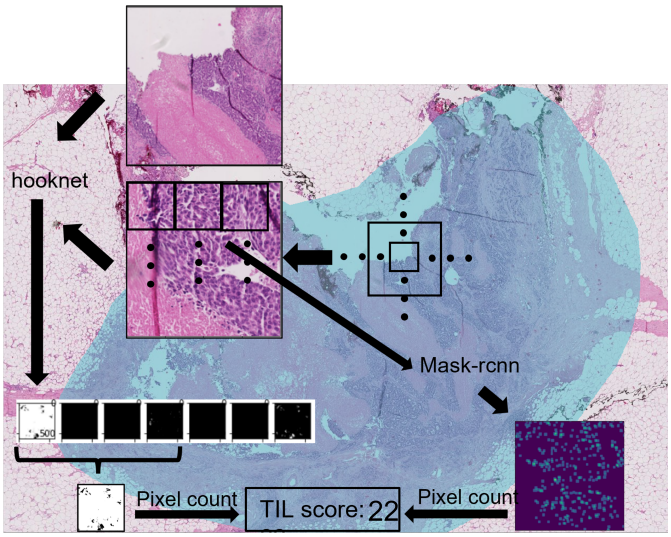

Fig. 7. Case study of the proposed multi-stage automatic TIL scoring system.

C. Stage III: TILs Quantification

For the TILs quantification, we have been able to achieve a Dice score of 0.73 on an independent test set using Mask R-CNN. We also include one visualization example of the detection results from a test image as shown in Fig. 6. For the final TIL score generation, we choose one case study to include the whole multi-stage quantification framework towards the final score prediction as shown in Fig. 7.

D. System Application

Figure 8 shows a screenshot of the GUI from our automated TIL-quantifying system. The primary component of the user interface is a window that allows the user to zoom and pan through an entire slide image. Next, there are options to enable overlays displaying the results of the tissue region segmentation, including the lymphocyte detection bounding boxes. Additionally, there are buttons for activating preset mask combinations. Currently, these are "isolate tumor bulk", which engages the invasive tumor, tumor stroma, and in situ-tumor to make the tumor easily identifiable; "show TIL contribution", which outlines the tissues and lymphocytes that contribute to the TIL score; and "reset all", which deactivates all overlays. The GUI displays the final TIL score in the lower right corner. This graphical user interface enables the clinician to obtain a quick TIL score upon loading an image and to explore how that TIL score is calculated.

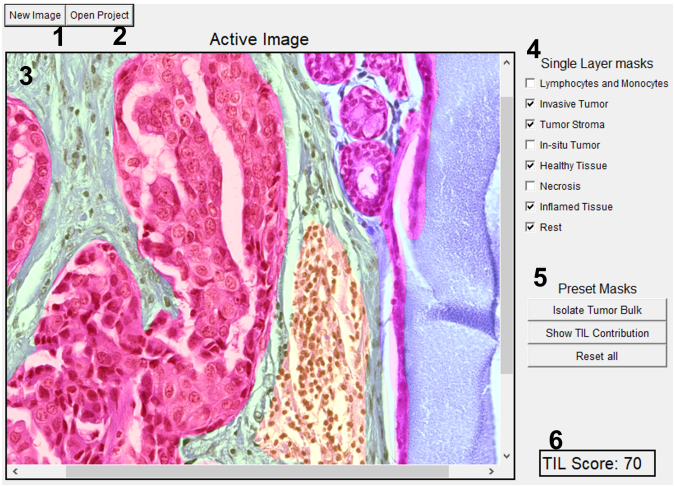

Fig. 8. Screenshot of our automated TIL-quantifying system with GUI. As annotated in the figure, 1 and 2 are buttons to load and generate images; 3 is the active image frame; 4 indicates boxes to activate image overlays; 5 is a preset button to quickly engage overlay combinations; and 6 is the output TIL score.

V. Conclusion

In this paper, we develop an automated deep learning system for predicting the TIL scores of digital WSI. The proposed framework includes three stages: (1) detection of lymphocytes and plasma cells, (2) segmentation of invasive tumor and tumor-associated stroma, and (3) detection of TIL. The final score is then determined based on the TILs and target tissues detected. The experiment results on TIGER dataset demonstrate the effectiveness of our proposed algorithm. Meanwhile, we create a prototype application with GUI to translate the proposed research into a potential clinical application. The contribution of this study is to develop a standardized and automated system that can enhance the prognostic value, thereby facilitating the incorporation of quantitative TIL scoring into clinical practice.

References

[1] T. Ghazal, "Breast cancer prediction empowered with fine-tuning," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–9, 2022.

[2] L. Pusztai, J. Foldi, A. Dhawan, M. P. DiGiovanna, and E. P. Mamounas, "Changing frameworks in treatment sequencing of triple-negative and her2-positive, early-stage breast cancers," *The Lancet Oncology*, vol. 20, no. 7, pp. e390–e396, 2019.

[3] C. Denkert, S. Loibl, A. Noske, M. Roller, B. Muller, M. Komor, J. Budczies, S. Darb-Esfahani, R. Kronenwett, C. Hanusch *et al.*, "Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer," *J Clin Oncol*, vol. 28, no. 1, pp. 105–113, 2010.

[4] K. El Bairi, H. R. Haynes, E. Blackley, S. Fineberg, J. Shear, S. Turner, J. R. De Freitas, D. Sur, L. C. Amendola, M. Gharib *et al.*, "The tale of tils in breast cancer: a report from the international immuno-oncology biomarker working group," *NPJ Breast Cancer*, vol. 7, no. 1, pp. 1–17, 2021.

[5] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. L. Baehner, F. Pénault-Llorca *et al.*, "The evaluation of tumor-infiltrating lymphocytes (tils) in breast cancer: recommendations by an international tils working group 2014," *Annals of oncology*, vol. 26, no. 2, pp. 259–271, 2015.

[6] B. Acs, F. S. Ahmed, S. Gupta, P. F. Wong, R. D. Gartrell, J. Sarin Pradhan, E. M. Rizk, B. Gould Rothberg, Y. M. Saenger, and D. L. Rimm, "An open source automated tumor infiltrating lymphocyte algorithm for prognosis in melanoma," *Nature communications*, vol. 10, no. 1, pp. 1–7, 2019.

[7] F. Klauschen, K.-R. Müller, A. Binder, M. Bockmayr, M. Hägele, P. Seegerer, S. Wienert, G. Pruneri, S. De Maria, S. Badve *et al.*, "Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning," in *Seminars in cancer biology*, vol. 52. Elsevier, 2018, pp. 151–157.

[8] M. Chou, I. Illa-Bochaca, B. Minxi, F. Darvishian, P. Johannet, U. Moran, R. L. Shapiro, R. S. Berman, I. Osman, G. Jour *et al.*, "Optimization of an automated tumor-infiltrating lymphocyte algorithm for improved prognostication in primary melanoma," *Modern Pathology*, vol. 34, no. 3, pp. 562–571, 2021.

[9] 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*. Springer, 2015, pp. 234–241.

[10] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced nlp tasks," *arXiv preprint arXiv:1911.02855*, 2019.

[11] M. Van Rijthoven, M. Balkenhol, K. Siliņa, J. Van Der Laak, and F. Ciompi, "Hooknet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Medical Image Analysis*, vol. 68, p. 101890, 2021.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.