Performance Evaluation of Deep Learning-based Tumor-Infiltrating Lymphocyte Cell Detection Algorithms for Histopathology Whole Slide Images

Arian Arab, Lakshyana KC, Seyed Kahaki, Weijie Chen

Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, CDRH, FDA

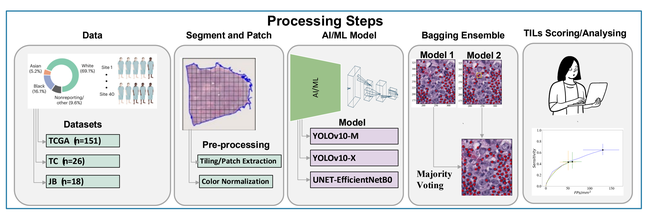
Abstract

- Tumor-infiltrating lymphocytes (TILs) are promising prognostic biomarkers for HER2+ and triple-negative breast cancer patients.
- Automated TILs detection using deep learning models holds clinical potential, but standardized assessment methods are lacking.
- In this study, we developed and evaluated two TILs detection models (a YOLOv10 and an EfficientNetB0-U-Net) using data from a public challenge (TiGER)[1].
- The TiGER dataset was divided into patient-level subsets for model development and testing, with the development set further split into five folds at the ROI level.
- Using development ROIs, the models were trained five times, using four folds for training and one for tuning.
- We used a bagging ensemble method (majority voting) to combine the five finetuned models' outputs.
- Using these finetuned models, we evaluated their performance on the hold-out test data for the cell detection task.
- We calculated several performance metrics at the patient-level. Using bootstrapped samples, we obtained confidence intervals for the selected metrics at the operating points determined during model's finetuning.

Purpose

- Cell detection by an AI model is commonly used in whole slide image (WSI)-based software as a medical device (SaMD). A standardized framework for evaluating these models can help streamline the premarket assessment of such SaMDs and facilitate their translation into clinical use.
- **The overall purpose of our project is to develop assessment methods for cell detection algorithms that are commonly used in many pathology devices, with the detection of Tumor-infiltrating lymphocytes (TILs) as a use case.**

Pipeline Overview



Data & Methods

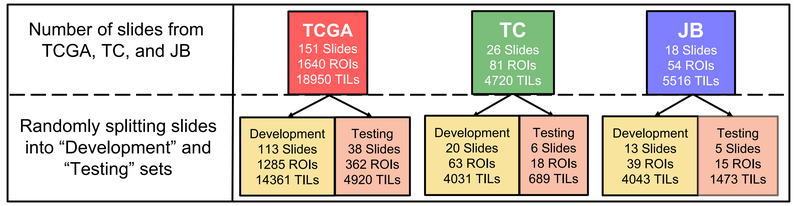
Development and Testing Data Split Strategy

TiGER data are from three sources:

TCGA: 151 Slides, Cell level annotations derived from BCSS and NuCLS datasets.

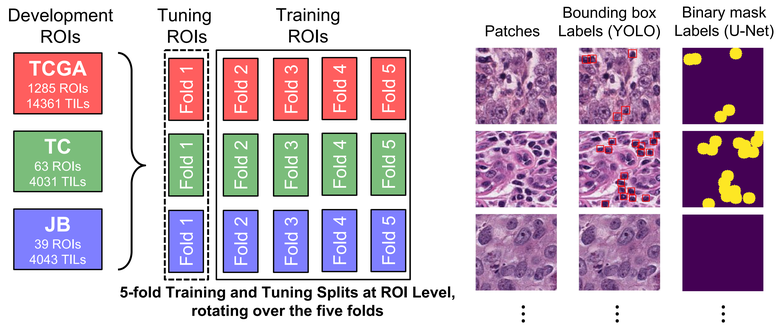
JB: 26 Slides, Cell level annotations made by a panel of board-certified breast pathologists.

TC: 18 Slidess Cell level annotations made by a panel of board-certified breast pathologists.



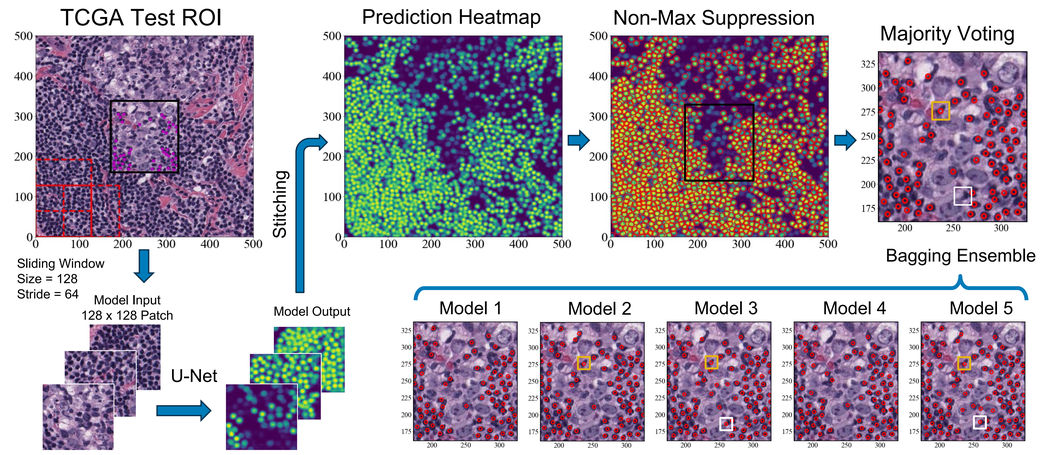
Model Training and Tuning

- YOLO model: An end-to-end detection model leveraging the YOLOv10 architecture for efficient TIL cell detection (M: medium version, X: extra-large version).
- U-Net model: A semantic segmentation model based on the U-Net architecture, using a lightweight EfficientNet-B0 as the backbone.



Model Testing

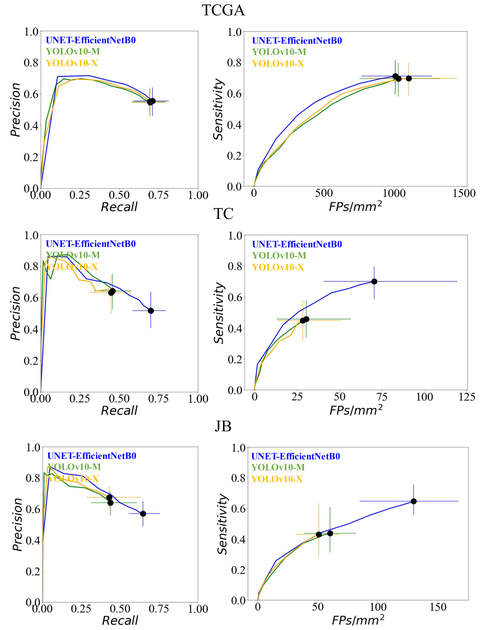
For each of the ROIs in the test set, we used the sliding window technique (strides of 64 pixels) to obtain patches of size 128×128 pixels. Models' outputs were obtained at the patch level and stitched together to obtain model's predictions at the ROI level. We used non-max suppression to obtain individual TILs locations and used a bagging ensemble method (majority voting) to combine predictions obtained from the five models trained using different training/tuning folds.



References:

1. https://tiger.grand-challenge.org/
2. Arab A. et al. Assessment of machine learning algorithms for TILs scoring using whole slide images: comparison with pathologists, Proc SPIE Medical Imaging 2024: Digital and Computational Pathology 12933, 285-289.

Results and Conclusion



Recall And Precision at the slide level for N slides

$Recall_i = \frac{tp_i}{tp_i+fn_i}$, $Precision_i = \frac{tp_i}{tp_i+fp_i}$

$Recall = \frac{1}{N}\sum_{i=1}^{N} Recall_i$ where the sums skip ROIs for which $Recall_i$=NA

$Precision = \frac{1}{N}\sum_{i=1}^{N} Precision_i$ where the sums skip ROIs for which $Precision_i$=NA

EfficientNetB0-U-Net			
	TCGA	TC	JB
Recall	0.71 [0.59-0.81]	0.7 [0.59-0.8]	0.65 [0.56-0.76]
Precision	0.56 [0.46-0.64]	0.52 [0.41-0.64]	0.57 [0.49-0.65]
FPs/mm2		70 [41-119]	130 [84-167]
YOLO10M			
Recall	0.69 [0.58-0.80]	0.46 [0.34-0.58]	0.44 [0.31-0.61]
Precision	0.55 [0.46-0.63]	0.64 [0.53-0.75]	0.64 [0.56-0.71]
FPs/mm2	1032 [756-1328]	30 [13-57]	60 [39-82]
YOLO10X			
Recall	0.70 [0.58-0.80]	0.45 [0.31-0.58]	0.43 [0.28-0.63]
Precision	0.55 [0.46-0.63]	0.63 [0.5-0.75]	0.68 [0.61-0.74]
FPs/mm2	1107 [801-1450]	28 [13-51]	51 [32-69]

- Binary metrics at one operating point can be ambiguous and performance curves (e.g., precision-recall curve or FROC curve) are more informative.
- We found in this study that multiple metrics ranked the models consistently – the U-Net model outperformed the YOLO models in this task using the TiGER datasets.

L. KC acknowledges funding by appointment to the Research Participation Program at the Center for Devices and Radiological Health administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the US Food and Drug Administration.