Week 4

Tumor-Infiltrating Lymphocytes (TILs) Using the TIGER Grand Challenge

1. The Problem

Tumor-infiltrating lymphocytes (TILs) are immune cells present within and around tumor tissue and play a critical role in the body's response to cancer. In breast cancer, particularly in HER2-positive and Triple Negative breast cancer (TNBC), TILs have been shown to be strong biomarkers for predicting patient prognosis and response to therapy. Traditionally, TIL assessment is performed manually by pathologists using histopathology slides stained with hematoxylin and eosin (H&E). This process involves visually estimating the proportion of immune cells within tumor-associated stroma.

However, manual TIL scoring presents several challenges. It is time-consuming, subjective, and prone to inter-observer variability, meaning that different pathologists may assign different scores to the same slide. Furthermore, the increasing volume of histopathology data in clinical practice makes manual evaluation impractical for large-scale or routine use. These limitations hinder the consistent use of TILs as a biomarker in clinical decision-making.

The TIGER (Tumor-Infiltrating lymphocytes in breast cancer) Grand Challenge addresses this problem by promoting the development of automated computer vision and artificial intelligence (AI) algorithms capable of accurately detecting immune cells, segmenting relevant tissue regions, and computing reliable TIL scores from whole-slide images. The challenge focuses on HER2-positive and TNBC cases, where TILs are most clinically relevant and impactful.

2. Why the Problem Matters

Breast cancer is the most commonly diagnosed cancer worldwide and a leading cause of cancer-related mortality among women. Despite advances in screening and treatment, outcomes vary widely depending on tumor subtype, stage, and biological characteristics. Molecular subtyping divides breast cancer into categories such as Luminal A, Luminal B, HER2-enriched, and Triple Negative, each with different treatment strategies and prognoses.

HER2-positive and Triple Negative breast cancers are particularly aggressive and associated with poorer survival outcomes. For these subtypes, immune system involvement plays a significant role in determining treatment response. Numerous studies have demonstrated that higher levels of TILs are associated with improved response to chemotherapy and immunotherapy, as well as better long-term survival.

Reliable TIL assessment therefore has major clinical implications:

- It helps oncologists predict patient prognosis.
- It supports personalized treatment planning, especially for immunotherapy.
- It can reduce unnecessary exposure to aggressive treatments such as chemotherapy.

However, the lack of standardized, reproducible, and scalable methods for TIL scoring limits its routine clinical use. Automating TIL assessment using AI can address these limitations by providing fast, objective, and consistent measurements. This not only improves diagnostic accuracy but also

enhances the potential of precision medicine by enabling better patient stratification and treatment selection.

The TIGER challenge is significant because it provides a standardized platform for evaluating AI-based solutions and validating their clinical utility on large, independent datasets, including data from clinical trials. This bridges the gap between research and real-world clinical application.

3. The Approach

The TIGER challenge defines a structured, multi-step approach to automated TIL assessment. Instead of directly predicting a TIL score from an image, the task is decomposed into three core components:

3.1 Detection of Immune Cells

The first step involves detecting lymphocytes and plasma cells within H&E-stained whole-slide images. These cell types constitute the main population of tumor-infiltrating lymphocytes. Accurate cell detection is essential because errors at this stage propagate to subsequent steps, affecting the final TIL score.

This task typically involves training deep learning models, such as convolutional neural networks (CNNs), to identify and localize immune cells at the pixel or patch level. Techniques such as object detection, instance segmentation, and classification are commonly employed.

3.2 Segmentation of Tissue Compartments

The second step is to segment relevant tissue regions, particularly invasive tumor and tumor-associated stroma. TIL scoring is performed specifically within stromal regions adjacent to tumor tissue, excluding areas such as necrosis, normal tissue, or in-situ carcinoma.

Segmentation models classify each pixel or region into tissue categories, enabling the system to focus immune cell counting within clinically meaningful compartments. This step ensures that the TIL score reflects immune infiltration in biologically relevant regions rather than across the entire slide indiscriminately.

3.3 Computation of the TIL Score

The final step combines immune cell detection and tissue segmentation outputs to compute a single TIL score per slide. This score represents the proportion or density of immune cells within the tumor-associated stroma and is intended to be clinically interpretable and prognostically meaningful.

Participants are encouraged to follow established clinical guidelines, such as those proposed by the International TIL Working Group, or to develop alternative scoring strategies based on spatial analysis, morphological patterns, or Immunoscore-like approaches. The challenge promotes innovation in both algorithm design and biomarker definition, as long as the resulting score demonstrates strong prognostic value.

4. The Dataset

The TIGER challenge uses large-scale, curated datasets consisting of H&E-stained whole-slide images (WSIs) from breast cancer patients, specifically focusing on HER2-positive and Triple Negative

subtypes. These datasets are provided through the Grand-Challenge.org platform and are divided into training, validation, and test sets.

4.1 Image Data

The core data consists of high-resolution histopathology slides, which capture detailed cellular and tissue-level structures. These images are typically gigapixel-scale, requiring efficient processing techniques such as patch-based analysis and multi-scale modeling.

4.2 Annotations

The datasets include expert annotations for:

- Lymphocytes and plasma cells (for detection tasks)
- Tissue compartments such as invasive tumor, stroma, and background (for segmentation tasks)

These annotations serve as ground truth for training and evaluating AI models. They are created by experienced pathologists following standardized protocols, ensuring high-quality reference labels.

4.3 Clinical Data

In addition to image data, the final evaluation dataset includes associated clinical outcome information, such as survival data and treatment response. This enables the assessment of the prognostic value of automated TIL scores, not just their technical accuracy.

Importantly, part of the test dataset originates from a phase III clinical trial and is not directly accessible to participants, ensuring unbiased evaluation and preventing overfitting to known data.

5. Challenges and Observations

5.1 Variability in Histopathology Images

One of the major challenges in automated TIL assessment is the inherent variability in histopathology slides. Differences in staining intensity, scanner type, tissue preparation, and slide quality introduce significant variation across images. These factors can negatively affect model performance if not properly addressed through normalization, augmentation, and robust training strategies.

5.2 Complexity of the Tumor Microenvironment

The tumor microenvironment is highly complex, with diverse cell types, tissue structures, and morphological patterns. Distinguishing lymphocytes from other small, round cells (such as fibroblasts or tumor cells) is challenging, particularly in densely packed regions. Similarly, accurately segmenting tumor boundaries and stromal regions requires models to capture subtle histological cues.

5.3 Class Imbalance and Rare Patterns

Immune cells may be sparse in some regions, leading to class imbalance in training data. Rare morphological patterns, such as specific immune infiltration structures or tumor architectures, are underrepresented, making it difficult for models to generalize across all cases.

5.4 Generalization Across Institutions

Models trained on data from one institution or scanner may not perform equally well on data from other sources. Ensuring cross-domain generalization is essential for clinical deployment. The TIGER challenge addresses this by including multi-institutional data and independent test sets, encouraging the development of robust and transferable models.

5.5 Interpretability and Clinical Trust

For AI models to be adopted in clinical practice, their outputs must be interpretable and trustworthy. Clinicians require not only a numerical TIL score but also visual evidence, such as highlighted immune cells and segmented tissue regions, to verify and understand model decisions. Designing systems that provide transparent and explainable results is therefore a critical consideration.

5.6 Evaluation Beyond Technical Accuracy

A key observation from the TIGER challenge is the importance of evaluating models not only on technical metrics (such as detection accuracy or segmentation overlap) but also on clinical relevance, specifically the prognostic value of the computed TIL score. A technically accurate model that does not correlate with patient outcomes is of limited clinical utility. This dual evaluation framework ensures that models are both scientifically sound and clinically meaningful.