Computer Methods and Programs in Biomedicine 282 (2026) 109374

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

A modular deep learning pipeline for stromal TILs scoring in breast cancer H&E slides☆

Shrief Abdelazeez [a,b,*], Faisal Ahmed [c,d], Laia Adalid [e,f], Krzysztof Siemion [a,g,h], Carlos Lopez [e,f,c], Marylene Lejeune [e,f,c], Hatem Rashwan [c], Anna Korzynska [a]

[a] Laboratory of Analysis and Processing of Microscopic Image, Nałęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Ks. Trojdena 4, Warsaw, 02-109, Poland
[b] System and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Giza, Cairo, Egypt
[c] Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, Spain
[d] NVISION Systems and Technologies SL, Barcelona, Spain
[e] Oncological Pathology and Bioinformatics Research Group, Institut d'Investigació Sanitària Pere Virgili, Tortosa, Spain
[f] Department of Pathology, Hospital de Tortosa Verge de la Cinta, Institut Català de la Salut, Tortosa, Spain
[g] Medical Pathomorphology Department, Medical University of Bialystok, Bialystok, Poland
[h] Faculty of Medical and Health Sciences, University of Radom, Radom, Poland

ARTICLE INFO

Keywords:
Tumor-infiltrating lymphocytes
Breast cancer
Deep learning
Digital pathology
H&E stained images
Quantitative pathology
TILs scoring
Convolutional neural networks

ABSTRACT

Background and Objective: Tumor-infiltrating lymphocytes (TILs) are an important indicator of immune activity in breast cancer, yet scoring them consistently on H&E slides remains challenging in routine pathology. This work presents a modular deep learning pipeline that delivers fully automated and continuous stromal TILs (sTILs) scores in line with the International Immuno-Oncology Biomarker Working Group (IIOBWG) guidelines.
Methods: The pipeline combines three components: a TIL segmentation model refined through pathologist-guided active learning, a robust stroma segmentation network based on an enhanced DeepLabV3+, and a lightweight regression module that learns how TILs distribute within stromal regions. A new adaptive aggregation strategy integrates patch-level predictions into a single, clinically meaningful score while accounting for heterogeneous infiltration.
Results: The system was evaluated on two independent datasets (60 and 112 WSIs) with expert-annotated ROIs, achieving strong agreement with pathologists (Pearson of 0.814; ICC of 0.808).
Conclusions: Importantly, the pipeline is interpretable: each stage produces human-readable outputs (stroma masks, TIL-in-stroma maps), and SegGradCAM visualizations confirm that predictions rely on biologically relevant tissue regions. These findings demonstrate the pipeline's potential as a reliable and clinically adaptable tool for standardized, fully automated TILs quantification in breast cancer pathology. The source code and pretrained models are publicly available at https://github.com/Shrief-Abdelazeez/TILs-Scoring.

1. Introduction

Breast cancer remains one of the most common malignancies among women worldwide and a leading cause of cancer-related mortality [1]. Tumor-infiltrating lymphocytes (TILs), a key component of the tumor microenvironment, play a critical role in orchestrating the immune response against cancer. Numerous studies have demonstrated that the density and spatial dispersion of TILs are associated with treatment response and patient survival, particularly in triple-negative and HER2-positive breast cancer subtypes [1,2]. Because pathologists require confidence in AI-assisted scoring, any automated method must also provide interpretable intermediate results that can be visually verified, rather than functioning as a black-box system. Interpretability has therefore become an important research direction in medical artificial intelligence to improve transparency and clinical trust in deep learning models [3,4].

☆ This article is part of a Special issue entitled: 'AI & Emerging Technologies in Dig. Health (ITIB2026)' published in Computer Methods and Programs in Biomedicine.

* Corresponding author at: Laboratory of Analysis and Processing of Microscopic Image, Nałęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Ks. Trojdena 4, Warsaw, 02-109, Poland.
E-mail address: sabdelazeez@ibib.waw.pl (S. Abdelazeez).

https://doi.org/10.1016/j.cmpb.2026.109374
Received 3 February 2026; Received in revised form 16 March 2026; Accepted 14 April 2026
Available online 18 April 2026
0169-2607/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

The manual assessment of TILs on hematoxylin and eosin (H&E)-stained slides, while clinically informative, is inherently subjective, labor-intensive, and prone to inter-observer variability [1]. With the growing adoption of whole slide imaging (WSI) and digital pathology, computational methods have been increasingly developed to automate or support TILs evaluation, aiming to improve reproducibility and subjectivity of quantification results [5].

The IIOBWG has introduced standardized guidelines for TILs evaluation in breast cancer to ensure reproducible and clinically manual assessments across studies and institutions [6]. According to these recommendations, TILs should be quantified as the percentage of stromal area infiltrated by mononuclear inflammatory cells within the borders of invasive tumors, and the degree of immune infiltration is commonly categorized as *low*, *moderate*, or *high*. While this categorical approach facilitates clinical adoption, it may overlook subtle differences in immune response intensity. In contrast, continuous TILs scoring offers finer granularity and has the potential to improve prognostic modeling and therapy-response stratification by capturing more nuanced variations in immune infiltration.

Over the past few years, several computational studies have explored machine learning and deep learning techniques for automated TILs quantification in histopathological images. Early works relied on handcrafted features and conventional classifiers to estimate immune infiltration levels [2,7], whereas recent approaches have utilized convolutional neural networks (CNNs) and attention-based architectures to segment lymphocytes or directly estimate TILs scores from image tiles [8–10]. These methods have shown promising correlations with expert visual assessments, demonstrating the feasibility of artificial intelligence (AI) in reducing subjectivity in TILs evaluation.

However, despite these advances, most existing methods remain limited to isolated tasks such as lymphocyte detection, stroma segmentation, or categorical TILs grading. Few approaches incorporate both tissue context and immune distribution to yield a continuous, stroma-constrained quantitative score. Moreover, most pipelines rely on fixed mathematical formulations for score aggregation, lack explicit explainability, or are not easily adaptable across datasets with varying image quality and staining conditions. This highlights the need for a unified, transparent framework in which each computational step, i.e., segmentation, localization, and scoring, produces human-understandable outputs that align with pathologists' reasoning.

To address these gaps, this study proposes a **modular deep learning pipeline** for continuous stromal TILs (sTILs) scoring in breast cancer H&E slides. The pipeline integrates complementary modules for TILs segmentation, stroma segmentation, and regression-based continuous score prediction within a unified framework designed for transparency and adaptability. The modular design enables independent optimization and fine-tuning of each component, facilitating generalization across datasets and clinical settings. Because each module outputs visually interpretable masks and heatmaps, the framework allows clinicians to inspect how the final score is derived, increasing its usability for real-world pathology workflows.

The main contributions of this work are summarized as follows:

- We present the first modular deep learning pipeline that integrates a regression-based scoring module capable of predicting continuous TILs scores by learning discriminative features from binary TILs and stroma maps, rather than relying on predefined mathematical ratios.
- The proposed pipeline incorporates a TILs segmentation module with active learning [11,12] to iteratively improve model performance and increase the precision of TIL detection through expert-guided refinement of difficult or ambiguous regions.
- A robust preprocessing and patch-level quality control strategy is employed to exclude irrelevant or unstained regions, enabling accurate scoring at both the ROI level, as defined by expert pathologists, and the whole-slide level.
- A novel adaptive aggregation strategy combining stromal area weighting and percentile scaling is introduced to translate patch-level predictions into biologically meaningful continuous scores.
- Comprehensive validation was performed across two independent patient subsets, including one used for internal validation, supported by SegGradCAM-based explainability analysis to ensure transparency and clinical trustworthiness.

The remainder of this paper is organized as follows. Section 2 reviews existing computational approaches for stromal TILs quantification. Section 3 presents the materials and methods, including dataset descriptions, preprocessing strategies, and the proposed modular pipeline. Section 4 reports and discusses the experimental results in detail. Section 5 outlines the main limitations of the study. Finally, Section 6 provides the overall conclusion together with key directions for future work, followed by the *Acknowledgments* section recognizing funding sources and contributions.

2. Related work

In recent years, deep learning has substantially advanced computational pathology, particularly in automating the assessment of tumor-infiltrating lymphocytes (TILs) in breast cancer. Although standardized guidelines exist for TILs evaluation [6], manual assessment on H&E-stained slides remains laborious and inherently subjective. This motivates the development of reproducible, automated quantification approaches.

Several computational studies have addressed this challenge. Saltz et al. [2] developed a deep learning framework using convolutional neural networks (CNNs) to identify and spatially map TILs across The Cancer Genome Atlas (TCGA) dataset. Their large-scale analysis demonstrated the prognostic significance of immune cell spatial organization but relied on patch-level classification without providing direct TILs scoring. Similarly, Lu et al. [13] employed an attention-based multiple instance learning (MIL) approach to predict molecular features and immune infiltration at the tissue level. While effective for slide-level inference, the method required weak labels and lacked spatial interpretability at the cellular or regional level. These approaches highlight the potential of AI for immune profiling but also show that most existing systems provide limited control over how predictions are generated, reducing their transparency for clinical use.

Other researchers explored weakly supervised learning for histopathological analysis. Campanella et al. [5] demonstrated the feasibility of large-scale WSI classification using MIL, but their framework did not address TIL-specific quantification. Bandi et al. [14] later introduced a context-aware model that incorporated neighboring tissue information to enhance TIL detection. Although this approach improved spatial awareness, it involved complex preprocessing and did not generate interpretable or continuous TILs scores. In addition, most weakly supervised pipelines do not explicitly separate stromal and tumoral compartments, which is essential for guideline-compliant sTILs scoring.

AI-assisted quantification has also been reported to enhance scoring consistency. Heindl et al. [15] showed that automated spatial analysis could improve the reproducibility of sTILs scoring and facilitate broader clinical adoption. Saltz et al. [2] similarly demonstrated high concordance between AI-predicted immune infiltration and pathologists' evaluations, underscoring the potential of computational systems to standardize TILs assessment across institutions.

More specialized frameworks have been proposed for distinct histological components. Makhlouf et al. [16] developed a two-stage deep learning pipeline that employed U-Net [17] for tumor and stroma segmentation and HoVer-Net [18] for immune cell detection. Although the model performed well for luminal breast cancer, it lacked validation across other molecular subtypes, limiting its clinical generalizability. Choi et al. [19] proposed a multi-model framework combining Faster

R-CNN [20] for lymphocyte detection and DeepLabV3 [21] for tissue segmentation. Despite achieving good correlation with three expert pathologists, the absence of external validation restricted its real-world applicability. Moreover, these multi-model approaches do not provide a unified, end-to-end scoring mechanism, making their outputs harder to interpret and compare across datasets.

Yosofvand et al. [22] introduced a two-stage architecture using U-Net for stromal segmentation and Mask R-CNN [23] for TIL detection. While high accuracy was reported, both training and evaluation were conducted on the same dataset, raising concerns about overfitting and lack of independent validation. Similarly, Kim et al. [24] applied self-supervised pretraining (SimCLR [25] and SwAV [26]) for feature extraction in histopathological images. Although the framework leveraged large-scale public datasets to improve representation learning, it relied exclusively on resized WSIs, which can lead to spatial information loss and hinder clinical interpretability.

More recently, Liu et al. [27] proposed MuTILs, a concept-bottleneck neural network that jointly performs lymphocyte detection and tumor-associated stroma segmentation in accordance with IIOBWG guidelines. MuTILs introduced interpretability through intermediate "concepts" and achieved strong results on the PanopTILs dataset. However, its evaluation was confined to HER2-positive and triple-negative breast cancer subtypes, limiting generalization to the full spectrum of molecular phenotypes. Furthermore, concept-bottleneck models still rely on fixed heuristic ratios for final scoring, whereas our approach learns a continuous scale directly from data and integrates an adaptive aggregation strategy.

In summary, previous research has demonstrated the promise of deep learning for TILs quantification but remains constrained by fragmented pipelines, limited subtype coverage, and scarce external validation. Most importantly, few approaches offer a combination of stroma-aware TIL localization, continuous score prediction, and interpretable intermediate outputs within a single, modular framework. Addressing these challenges, the present study proposes a modular deep learning pipeline that integrates TILs segmentation, stroma segmentation, and regression-based score prediction within a unified architecture. The proposed method directly predicts spatially and numerically continuous TIL assessment scores in the ROI from hematoxylin and eosin-stained images and complies with the IIOBWG guidelines. At the same time, it offers a scalable and interpretable solution for standardized TIL quantification across molecular subtypes and can support inference based on the entire WSI.

3. Materials and methods

3.1. Data description

3.1.1. Private data

The private dataset consists of hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) from breast cancer patients collected by the Oncological Pathology and Bioinformatics Research Group at the Institut d'Investigació Sanitària Pere Virgili, Tortosa, Spain. All slides were digitized at 20× magnification with a spatial resolution of 0.2433 μm/pixel. Expert pathologists manually identified and annotated the regions of interest (ROIs) using the GeoJSON format. These ROIs focused on tissue compartments including tumor, stroma, and other components such as necrosis and glands.

The dataset is divided into two subsets:

Subset 1 (60 WSIs): This subset includes 60 WSIs representing multiple breast cancer molecular subtypes: Luminal A, Luminal B (HER2+/-), HER2-enriched, and Triple Negative, as summarized in Table 1. ROIs were annotated for tumor, stroma, and other tissue classes. Among these, 10 ROIs were further annotated for tumor-infiltrating lymphocytes (TILs), including both lymphocytes and plasma cells.

Each ROI was visually evaluated by two expert pathologists independently and in a blinded manner, following the standardized recommendations of Salgado et al. [6]. For each ROI, both pathologists assigned a TILs score reflecting the extent of stromal immune infiltration, categorized as *Low*, *Moderate*, or *High*. This independent dual assessment minimized observer bias and ensured reproducibility of the reference annotations. The aggregated distribution of visually assigned TILs categories is reported in Table 1.

Subset 2 (112 WSIs): This subset contains 112 WSIs encompassing the major breast cancer molecular subtypes (Table 1). Each ROI was visually assessed by a single expert pathologist according to the IIOBWG standardized guidelines [6], establishing a gold standard for model evaluation.

This subset was exclusively reserved as an independent hold-out cohort for **internal validation** of the proposed pipeline to assess its generalization performance on unseen patient samples that were not used during model development. The distribution of TILs scores in this subset is also summarized in Table 1.

3.1.2. Public data

We used the TIGER dataset,[1] consisting of 151 H&E-stained WSIs of triple-TNBC in TIF format at a resolution of 0.5 μm/pixel. Each WSI includes ROIs and corresponding masks annotated with pixel-level information for seven tissue compartments: *invasive tumor*, *in-situ tumor*, *tumor-associated stroma*, *inflamed stroma*, *healthy glands*, *necrosis (non–in-situ)*, and remaining tissue areas. In our experiments, we considered *tumor-associated stroma* and *inflamed stroma* as *stroma*, and the remaining tissue regions as *background*, following the IIOBWG guidelines.

3.2. Model architecture

The proposed pipeline comprises three main modules, designed to process 256 × 256 histopathology image patch and predict a continuous TILs score. These modules are: a TILs segmentation module, a stroma segmentation module, and a quantification module. The overall architecture is illustrated in Fig. 1.

TILs Segmentation Module: A binary semantic segmentation model specialized in identifying tumor-infiltrating lymphocytes (TILs), including lymphocytes and plasma cells, within the input patch. Active learning is integrated to mitigate annotation bias and improve detection precision. The module outputs a binary mask highlighting TIL-positive regions.

Stroma Segmentation Module: A parallel segmentation model trained to identify stromal regions in the same image. It outputs a binary mask corresponding to stroma tissue.

TILs-in-Stroma Map Generation (Intermediate Step): The outputs of the TILs and stroma segmentation modules are combined using a pixel-wise logical AND operation to generate a binary mask representing only the TILs located within the stroma. Although not a trainable module, this step plays a critical role in preparing the input for the regression model.

TILs Quantification Module: The TILs-in-stroma map is concatenated with the stroma mask, forming a two-channel input tensor. This is fed into a CNN-based regression model, which predicts a continuous TILs score for the input patch, reflecting the extent of immune cell infiltration in stromal regions.

Each of the three core modules (TILs segmentation, stroma segmentation, and regression) is independently trained using task-specific data and loss functions, as described in the next sections.

[1] https://tiger.grand-challenge.org/

Table 1
Distribution of breast cancer molecular subtypes and visual TILs score categories in the private dataset.

Category	Subset 1 (60 WSIs)	Subset 2 (112 WSIs)
Molecular Subtypes		
Luminal A	27	49
Luminal B (HER2+/-)	29	49
HER2-enriched	1	4
Triple Negative	3	10
Visual TILs Score Categories		
Low (≤10%)	44	73
Moderate (10%–40%)	12	26
High (>40%)	4	13

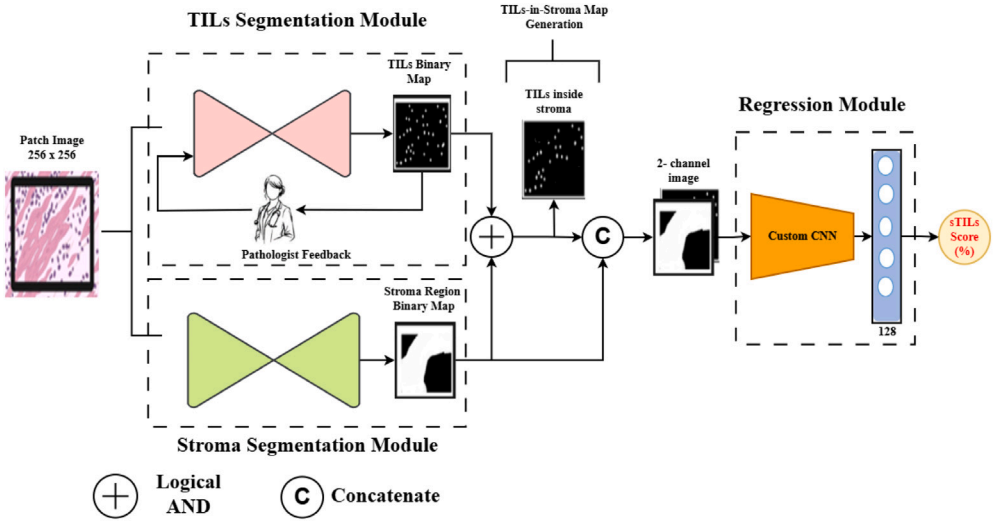


Fig. 1. Overview of the proposed model architecture, including segmentation modules for TILs and stroma, the intermediate TILs-in-stroma map generation step, and the regression module.

3.3. TILs segmentation module

The TILs segmentation module is responsible for detecting tumor-infiltrating lymphocytes (TILs)—including lymphocytes and plasma cells—in breast cancer histopathology images. This component isolates immune-related cellular structures within the stroma and provides the foundation for subsequent TILs quantification. To improve segmentation precision and mitigate annotation bias, a pathologist-guided active learning strategy was incorporated into the training pipeline.

Data Utilized: The dataset used for training this module consists of the 10 manually annotated ROIs described in Section 3.1.1. Each ROI was converted into a binary mask identifying all TILs. This curated collection served as the basis for both model training and evaluation.

Model Architectures: Three architectures were evaluated for TILs segmentation: UNet [28], Attention UNet [29], and TransUNet [30]. UNet provides a strong encoder–decoder baseline with skip connections for preserving fine detail. Attention UNet adds attention gates to better highlight sparse lymphocytes and suppress irrelevant regions. TransUNet combines CNN features with transformer-based global context, enabling richer spatial representation. All models were trained end-to-end with ImageNet-pretrained encoders to support stable convergence and improve feature quality.

Preprocessing Strategy: A unified preprocessing pipeline was applied throughout the study to ensure consistent experimental conditions across all training stages. To prevent data leakage and guarantee a fair evaluation of model performance, the dataset was split at the ROI (patient) level into 60% training, 20% validation, and 20% testing. The same held-out test set was preserved for both the initial training phase and the subsequent active learning stage.

Each ROI was partitioned into overlapping image patches of size 256 × 256 pixels using a stride of 64 pixels. This tiling strategy increases the effective sampling of local tissue patterns while preserving contextual continuity between neighboring patches. For each extracted patch, the corresponding binary TIL mask was generated from the annotation mask. Patches that did not contain TILs were excluded from the training set to reduce class imbalance and focus the learning process on biologically relevant regions.

To further refine cell-level annotations in densely populated areas, a watershed segmentation algorithm was applied to the binary masks. This step enabled separation of touching or overlapping lymphocytes, which commonly occur in regions with high immune infiltration, thereby improving the quality and accuracy of the training labels used for model learning.

Data Augmentation: Data augmentation was employed to improve model generalization and robustness to variations commonly encountered in digital pathology images. Augmentation was applied exclusively to the training set to avoid introducing artificial variability into validation and test data.

Offline augmentations were first applied to the training dataset to increase sample diversity, effectively expanding the dataset size by approximately four times. These transformations included horizontal and vertical flipping, 90° rotations, transpositions, affine transformations (including shifts, scaling, and rotations), Gaussian blurring, and brightness and contrast adjustments.

To simulate scanner-related variability observed across different acquisition systems, a custom magnification augmentation strategy was implemented. This operation downsampled image patches followed by upscaling them back to the original resolution, thereby mimicking variations in optical magnification and scanner resolution that may occur across institutions.

In addition, stain variability was addressed using the RandStainNA method [31], which was applied as an online stain augmentation

during training. This technique randomly perturbs stain distributions in hematoxylin and eosin images, enabling the model to learn features that are robust to staining variations.

Representative examples of the applied augmentation strategies are provided in the Supplementary Material (Figure S1). By introducing controlled variability during training, this augmentation protocol improves the model's ability to generalize to unseen data while preserving the morphological characteristics of lymphocytes and surrounding cellular context.

Active Learning Integration: To refine annotations and improve TIL detection performance, a pathologist-guided active learning strategy was implemented at the ROI level. The objective of this step was to correct systematic model errors and improve annotation quality using targeted expert feedback.

The active learning procedure consisted of a single refinement iteration. First, the best-performing model obtained from the initial training stage was applied to all ROIs to generate predicted TIL segmentation masks. These predictions were subsequently post-processed using watershed segmentation to separate touching lymphocytes and enable cell-level analysis.

Potential model errors were then identified using a false positive (FP) and false negative (FN) detection framework. FP masks corresponded to structures incorrectly detected as lymphocytes by the model, whereas FN masks indicated lymphocytes present in the ground truth but missed by the model. These FP and FN masks were converted into GeoJSON format and overlaid on the corresponding whole-slide images using QuPath, allowing efficient expert review within a digital pathology environment.

Pathologists inspected these candidate regions and provided feedback by confirming true lymphocytes among FP detections and removing incorrectly annotated structures from FN regions that did not correspond to true TILs. The review was performed at the ROI level rather than the patch or tile level to preserve spatial context and to avoid the substantial annotation burden associated with inspecting a large number of extracted patches. In practice, a single ROI typically contains many overlapping tiles, which would significantly increase annotation time if reviewed individually. Presenting model predictions directly on the whole-slide image also aligns with the standard workflow used by pathologists in digital pathology software, facilitating efficient verification and correction of candidate detections. On average, expert review required approximately one hour per ROI.

Based on this validated feedback, corrected annotation masks were generated by updating the original ground-truth masks. These refined masks were subsequently reprocessed using watershed segmentation to maintain accurate separation of neighboring lymphocytes. The updated annotations were then incorporated into the training dataset, and the model was retrained using the same training protocol as the initial stage.

Since the primary objective of the active learning step was to correct annotation inconsistencies and improve label quality, a single refinement iteration was considered sufficient once the expert-reviewed annotations were integrated into the training dataset.

An overview of the complete active learning workflow is illustrated in Fig. 2.

Explainability: To provide interpretable outputs and support verification by pathologists, SegGradCAM [32] was applied to visualize features influencing model predictions. Because predicted masks often contain multiple TILs, each instance was first isolated and processed separately. SegGradCAM heatmaps were generated for each instance and then combined using a pixel-wise maximum to create a global interpretability map. This instance-aware approach provides fine-grained insights into the model's decision-making process, highlighting relevant cellular structures and aiding interpretation in densely populated regions.

3.4. Stroma segmentation module

The stroma segmentation module identifies stromal regions surrounding tumor areas and provides the spatial context required for accurate TILs quantification. The stroma comprises both tumor-associated and inflamed stromal regions, and both are considered as part of the total stromal area used for downstream scoring.

Data Utilized: Model development was performed using the public dataset described in Section 3.1.2. To assess generalization beyond the training domain, the trained models were further evaluated on an independent in-house dataset consisting of 60 manually annotated ROIs from breast cancer WSIs (Section 3.1.1). This two-stage evaluation enabled a rigorous assessment of cross-domain robustness.

Preprocessing & Training Strategy: Initial inspection of the binary annotation masks in the public dataset revealed that 48 of the 151 annotated ROIs were rotated relative to the horizontal axis. When such regions are extracted, the resulting bounding rectangles contain large non-informative black margins. These artifacts introduce irrelevant background pixels that may negatively affect the learning process and bias the segmentation model. To mitigate this issue, the rotated ROIs were excluded from the dataset, resulting in 103 valid samples used for model training and evaluation.

To prevent data leakage and ensure reliable performance estimation, the dataset was partitioned at the ROI (patient) level into 80% training and 20% testing subsets. Within the training subset, stratified 5-fold cross-validation was applied to optimize model training and hyperparameter selection. For each fold, 80% of the ROIs were used for training and the remaining 20% for validation, ensuring that all samples contributed to the validation process across folds.

During training, each ROI was divided into patches of size 256×256 pixels extracted on the fly with 50% overlap (128 pixels). This overlapping tiling strategy increases spatial sampling of stromal structures while preserving contextual continuity between neighboring patches. Patches containing only background pixels were discarded to focus the learning process on biologically relevant tissue regions.

Data Augmentation: Data augmentation was applied during training to increase sample diversity and improve the robustness of the segmentation model to variations commonly encountered in histopathological images. Augmentation was performed on the fly to generate diverse training samples while preserving the original dataset distribution.

The augmentation strategy included geometric transformations such as horizontal and vertical flips as well as random rotations. Photometric transformations were also applied, including brightness adjustments and hue–saturation–value (HSV) shifts, allowing the model to learn features that are invariant to staining intensity variations. In addition, resolution scaling was introduced to simulate variability in scanner magnification and acquisition settings across different institutions.

Representative examples of the applied augmentation strategies are provided in the Supplementary Material (Figure S2). These transformations introduce controlled variability during training, enabling the model to better generalize to unseen tissue appearances while preserving the morphological characteristics of stromal structures.

Proposed Model Architecture: Stroma segmentation was performed using an enhanced DeepLabV3+ architecture [21] equipped with Squeeze-and-Excitation (SE) blocks [33]. SE modules were integrated at two key locations in the decoder: (1) after feature concatenation following the upsampling stage, and (2) after the final convolutional layer before the final upsampling step.

This design enables channel-wise recalibration of feature responses, allowing the decoder to emphasize informative stromal patterns while suppressing irrelevant background signals. By adaptively weighting feature channels, the network can better capture complex stromal textures and structural variations commonly observed in histopathological images. The overall architecture of the proposed model is illustrated in Fig. 3.

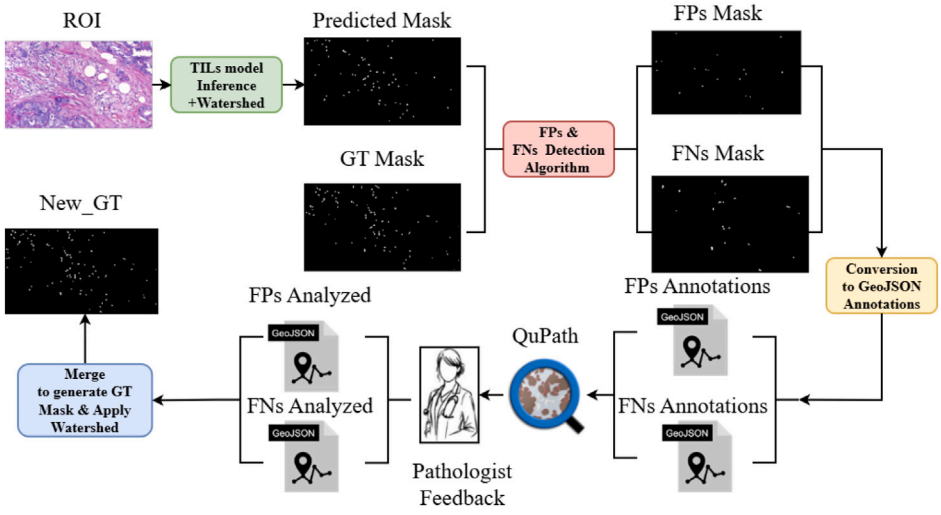


Fig. 2. Overview of the active learning integration process, illustrating the inference, annotation conversion, FP/FN detection, pathologist feedback, and updated ground truth mask generation.

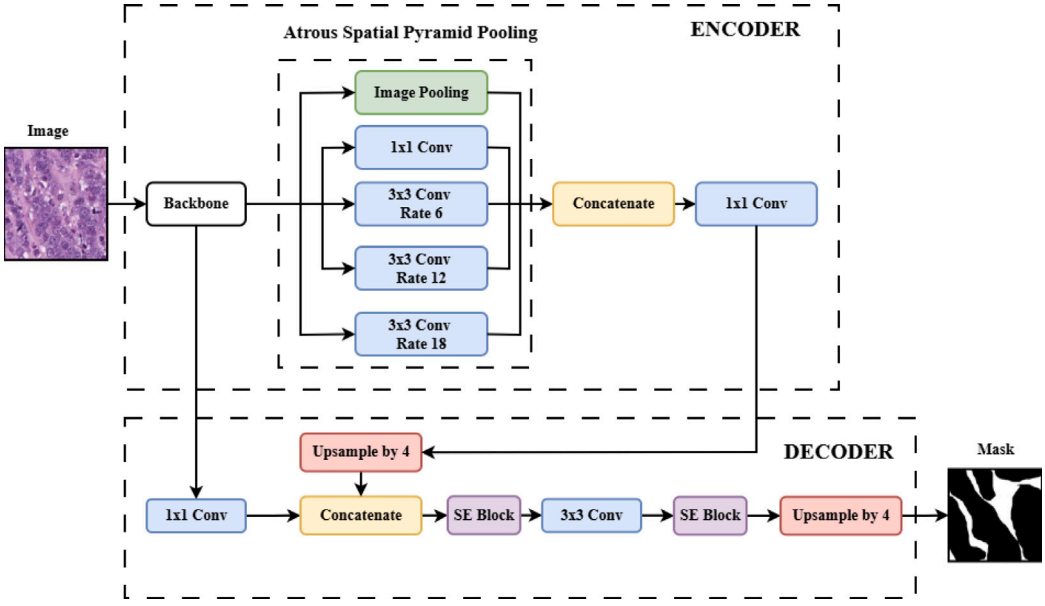


Fig. 3. Overview of the proposed DeepLabV3+ architecture enhanced with Squeeze-and-Excitation (SE) blocks at two positions in the decoder.

Four encoder backbones were incorporated into the SE-DeepLabV3+ framework to explore the trade-offs between accuracy, computational cost, and generalization: **ResNet101** [34], **ResNeSt101** [35], **MobileViT** [36], and **EfficientViT** [37].

These architectures were selected, as they are the most known backbones used in the literature in order to assess performance across a spectrum of model complexities. An overview of the decoder-enhanced architecture is shown in Fig. 3.

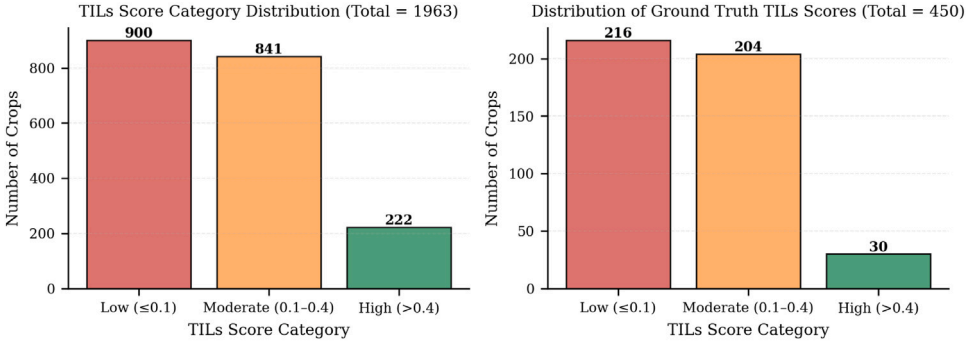
Evaluation Strategy: Evaluation was conducted in two stages. First, each of the five models obtained from 5-fold cross-validation was tested on the held-out public test ROIs to establish in-domain performance. Second, the same models were evaluated directly on the 60 private ROIs without fine-tuning. This external validation enabled objective assessment of cross-domain generalization and guided selection of the optimal backbone for integration into the TILs scoring pipeline.

Explainability: To interpret model predictions and enhance transparency, SegGradCAM [32] was applied to the predicted masks. This gradient-based technique highlights spatial regions that contribute most to the model's classification of stromal tissue. The resulting heatmaps allow verification that the model focuses on histologically relevant structures and provide an additional layer of interpretability for downstream clinical evaluation.

3.5. TILs quantification module

The TILs quantification module predicts a continuous TILs score at the tile level by leveraging spatial information derived from binary masks. The model receives a two-channel input: (1) a binary mask of TILs located within stromal regions, and (2) a binary mask of the stroma itself. This representation ensures that the regression model focuses exclusively on lymphocytes residing within stromal tissue, thereby capturing the biologically relevant context required for accurate TILs scoring.

Data Utilized: This module was developed using 60 ROIs from the private dataset described in Section 3.1.1. Each ROI was assigned a visual TILs score by an expert pathologist following the IIOBWG guidelines, and these ROI-level continuous scores served as ground truth for regression.



(a) Distribution of automatically generated patch-level TILs scores.

(b) Distribution of expert-validated patch-level TILs scores.

Fig. 4. Comparison of automated and expert-validated TILs score distributions at the patch level.

Data Preparation: Because pathologist-derived TILs scores are provided at the ROI level, whereas the regression model operates on fixed-size image patches, a dedicated preparation pipeline was implemented to generate patch-level training samples.

For each ROI, 500 random crops of size 256 × 256 pixels were extracted. Corresponding stroma masks were obtained directly from manual annotations, while TILs masks were generated using the best-performing segmentation model refined via active learning (Section 3.4). This approach was necessary because only 10 ROIs contained manual TILs annotations, and high-precision predictions enabled meaningful dataset expansion.

Crops with empty stroma masks were discarded. For each remaining crop, a TILs-inside-stroma mask was created by intersecting the predicted TILs mask with the stroma mask. A continuous TILs score was then computed following the area-based metric in [24], defined as the ratio of TILs area to total stromal area within the crop. These scores were subsequently categorized into low, moderate, and high infiltration levels based on the Salgado criteria.

To ensure balanced representation, up to 15 crops per category were sampled from each ROI, resulting in a maximum of 45 crops per ROI. When fewer valid samples were available for a category, all were included. Due to the predominance of non-inflamed stroma, the resulting dataset was naturally skewed toward low infiltration. In total, 1963 valid crops were retained. Each crop was stored as both an RGB image and a two-channel NumPy array containing the TILs-inside-stroma and stroma masks. The distribution of automated patch-level scores is shown in Fig. 4(a).

Expert Validation of Patch-Level Score: To assess the reliability of the automatically generated patch-level TILs scores, an expert pathologist independently reviewed 450 randomly selected crops. For each crop, the expert counted the number of TILs, and a new TILs score was computed using the same area-based metric. These scores were categorized into low, moderate, and high infiltration levels according to the Salgado criteria. The distribution of expert-validated scores is presented in Fig. 4(b).

Given the non-normal distribution of scores, Spearman's rank correlation was used to compare automated and expert-derived scores. A strong positive correlation was observed ($r = 0.762$, $p = 1.3164 \times 10^{-86}$), confirming the reliability of the automated scoring pipeline.

Data Preprocessing: The final dataset of 1963 crops was split into 80% for training and 20% for validation. Because the dataset was highly imbalanced — with the majority of samples belonging to the low-infiltration category — a stratified splitting strategy was used to preserve the original category proportions in both subsets. The resulting class distributions for the training and validation sets are shown in Supplementary Material (Figure S3).

Proposed Model Architecture: To predict continuous TILs scores from the two-channel input, a lightweight convolutional neural network (CNN) was designed (Fig. 5). The model accepts a 256 × 256 × 2 tensor and processes it through a compact sequence of convolutional blocks:

- A 3 × 3 convolution with 32 filters (ReLU), followed by max pooling.
- A second convolutional block with 64 filters and max pooling.
- A third convolutional block with 128 filters, followed by global average pooling.

The resulting feature vector is passed through a dense layer with 128 units and ReLU activation, followed by dropout (rate = 0.3) to reduce overfitting. A final dense layer with a single sigmoid-activated neuron outputs a normalized continuous TILs score in the range $[0, 1]$. This architecture was intentionally kept lightweight to prevent overfitting given the moderate dataset size, while still capturing relevant spatial relationships.

3.6. Evaluation pipeline

To obtain an ROI-level TILs score from local patch predictions, we developed a dedicated aggregation pipeline that integrates patch-wise scores using an adaptive strategy. As illustrated in Fig. 6, each ROI was tiled into non-overlapping 256 × 256 patches, consistent with the regression model's input size. Patches lacking stromal tissue were excluded to ensure that the global score reflects only anatomically and clinically meaningful regions.

Each valid patch was processed by the regression model to produce a continuous TILs score. Let the resulting vector of local TILs scores of patch images within the ROI be

$$\mathbf{S} = [s_1, s_2, \ldots, s_N],$$

where N denotes the number of stroma-containing patches. For each patch, a stromal area weight w_i was computed as the ratio of stromal pixels to the total patch area, forming the vector

$$\mathbf{w} = [w_1, w_2, \ldots, w_N].$$

To characterize the overall infiltration density within the ROI, we defined the proportion of moderately or highly infiltrated patches as

$$T = \frac{|\{s_i \in \mathbf{S} \mid s_i > 0.1\}|}{N}.$$

Based on this metric, the global ROI-level TILs score G was computed using an adaptive aggregation rule:

$$G(\mathbf{w}, \mathbf{S}, T) = \begin{cases} \frac{\mathbf{w}^\top \mathbf{S}}{\sum_{i=1}^{N} w_i}, & \text{if } T \le T_r, \\ P_x(\mathbf{S}), & \text{if } T > T_r, \end{cases}$$

where T_r is a tunable infiltration-density threshold and $P_x(\mathbf{S})$ denotes the xth percentile of the local scores (e.g., P_{90}, P_{95}, P_{99}).

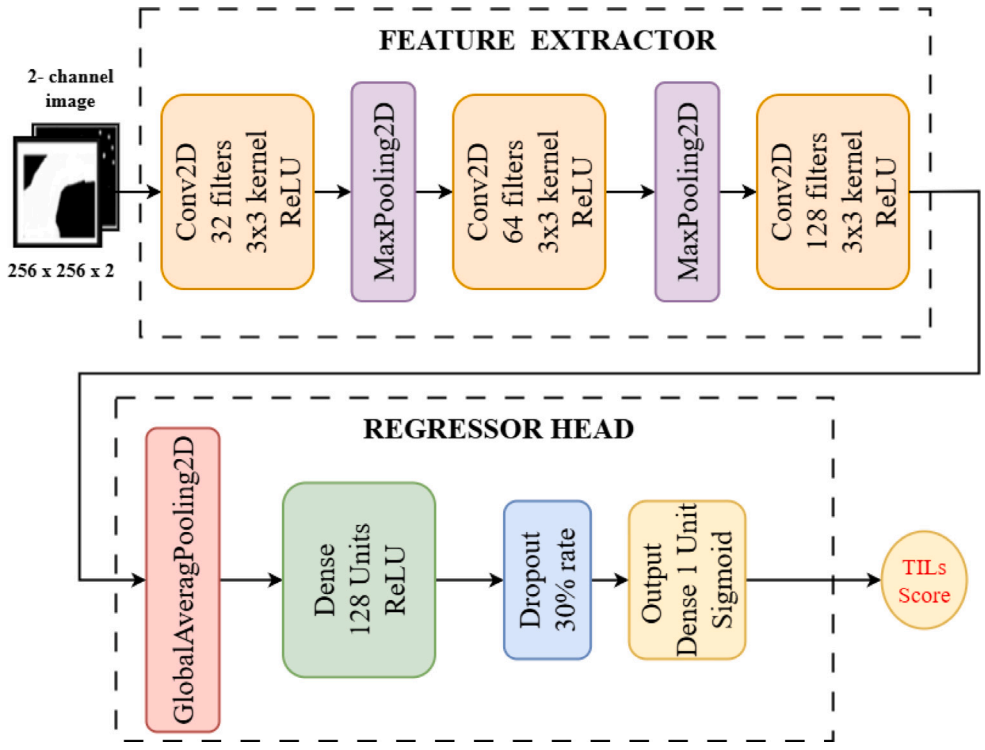


Fig. 5. Lightweight CNN architecture for continuous TILs score prediction. The model processes two-channel inputs representing TILs-inside-stroma and stroma masks.

Intuition Behind the Aggregation Strategy: Patch-level TILs scores within an ROI typically follow a highly skewed distribution, where most stromal patches exhibit very low infiltration and only a minority contain moderate or densely infiltrated regions. As a result, simple averaging approaches are dominated by low-valued patches and systematically suppress the contribution of clinically meaningful immune hotspots. This leads to global scores that underestimate the true infiltration level and artificially reduce score variance, weakening agreement metrics such as ICC and CCC even when rank-based correlations remain high. The proposed adaptive strategy explicitly addresses this imbalance: when infiltration is sparse, a stromal-weighted mean yields a stable estimate, while in ROIs where moderate or high infiltration is more prevalent, a percentile-based statistic amplifies the impact of densely infiltrated patches in a manner that more closely mirrors expert visual scoring.

Threshold values of $T_r = 0.15$, 0.20, and 0.25 were empirically evaluated, along with percentile values P_{90}, P_{95}, and P_{99}. The resulting global scores were compared to pathologist-derived ROI-level TILs scores using the correlation and agreement metrics described below.

Validation Metrics: To evaluate the agreement between the automated global TILs scores and the reference scores provided by expert pathologists, we used a set of well-established statistical measures. Pearson and Spearman correlations were computed to assess linear and rank-based relationships between predicted and reference values. For absolute agreement, we used the Intraclass Correlation Coefficient (ICC), following the ICC(2,1) formulation described by Shrout and Fleiss [38], and the Concordance Correlation Coefficient (CCC) introduced by Lin [39]. In addition, Cohen's Kappa [40] was calculated after mapping continuous scores into low, moderate, and high categories according to the criteria of Salgado et al. [6]. These metrics were applied across all threshold and percentile configurations to identify the settings that provided the most reliable alignment with expert assessments.

3.7. Ablation study design

To assess the contribution of each component of the proposed hybrid aggregation scheme, we conducted an ablation study comparing four aggregation strategies:

1. **Global average**: simple arithmetic mean of local scores, serving as a baseline following prior work [24].
2. **Stroma-weighted average** — local scores weighted solely by stromal area, emphasizing tissue composition without adaptive logic.
3. **Percentile-based aggregation**: the xth percentile of patch scores, highlighting highly infiltrated regions.
4. **Proposed hybrid method**: adaptive switching between stroma-weighted averaging and percentile-based aggregation depending on infiltration density T.

Each method was evaluated on the same ROI datasets using identical continuous and categorical agreement metrics (Pearson, Spearman, ICC, CCC, and Cohen's Kappa). This controlled comparison isolates the effect of the adaptive switching criterion and demonstrates how the hybrid design yields progressively improved alignment with expert assessments relative to simpler, non-adaptive baselines.

3.8. Experimental setup

All experiments were implemented in **Python 3.9.18** using the **TensorFlow 2.10.1** deep learning framework with the **Keras 2.10.0** API. The same software environment was used for all modules of the proposed pipeline to ensure experimental consistency and reproducibility.

The experiments were conducted on a workstation equipped with an **NVIDIA GeForce RTX 4080 GPU** with 16 GB of memory, an **Intel 13th Gen Core i5-13600KF CPU** (3.50 GHz), and **32 GB of system RAM**. This hardware configuration was used throughout the study for model training, inference, and evaluation.

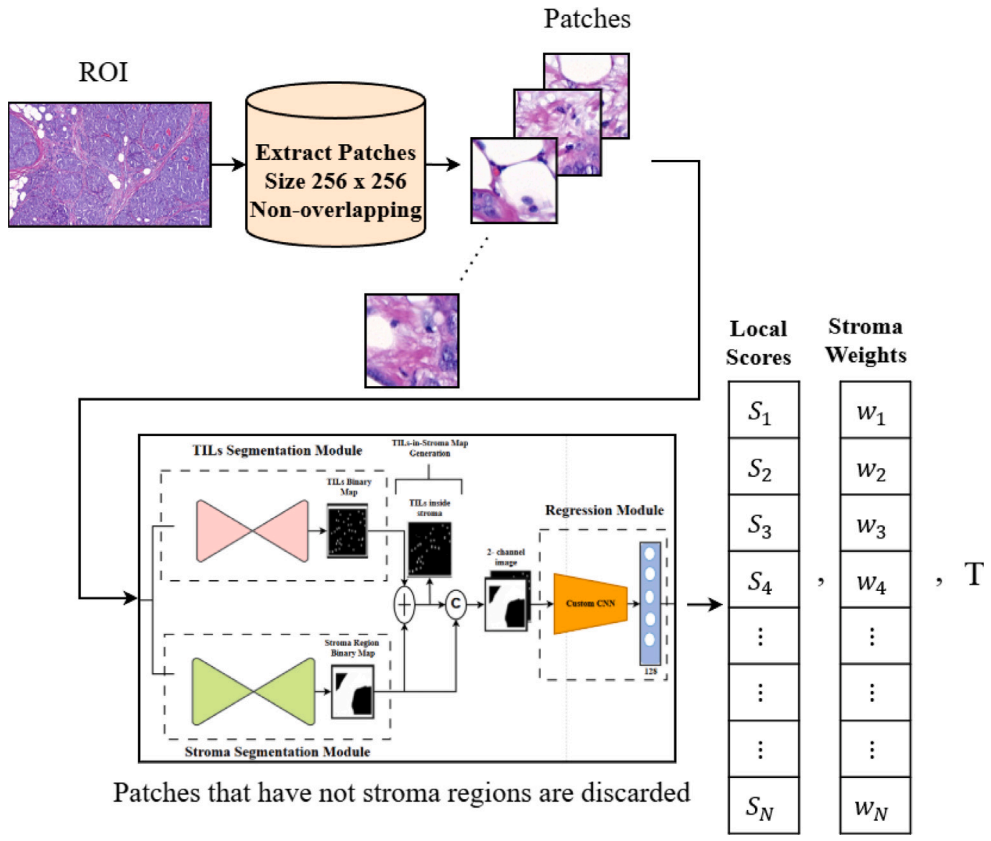


Fig. 6. Overview of the ROI-level evaluation pipeline, illustrating patch extraction, local score prediction, infiltration-density estimation, and adaptive global score computation.

3.8.1. TILs segmentation module

TILs segmentation was treated as a binary semantic segmentation task, using image patches from expert-annotated regions of interest (ROIs). As previously mentioned, to prevent data leakage, the dataset was initially split at the ROI level into training, validation, and testing sets. Then, image tiles of size 256 × 256 pixels were extracted from each dataset split with a 25% overlap relative to the tile size.

Three deep learning architectures were evaluated for this task: **U-Net**, **Attention U-Net**, and **TransUNet**. All models were initialized with ImageNet-pretrained weights and trained under the same configuration to ensure a fair comparison of learning behavior and computational requirements. During training, the full network — including both encoder and decoder components — was fine-tuned.

To address class imbalance and improve delineation of small lymphocyte regions, a composite loss function combining **Tversky focal loss** [41] and **IoU loss** was employed. Model optimization was performed using the Adam optimizer with an adaptive learning rate schedule based on validation performance.

Table 2 summarizes the dataset composition and training hyperparameters used for the TILs segmentation models, while Table 3 reports the model size and training time of the evaluated architectures.

3.8.2. Stroma segmentation module

Stroma segmentation was formulated as a binary semantic segmentation task using image tiles extracted from the ROI images. Tiles were generated with 50% overlap to preserve spatial continuity between neighboring regions. A stratified five-fold cross-validation strategy was adopted for model development, resulting in slight variations in the number of training and validation tiles across folds. All backbone networks were initialized with ImageNet-pretrained weights, and the entire network, including both encoder and decoder components, was fine-tuned during training.

Class imbalance between stromal and non-stromal regions was addressed using Dice loss to emphasize overlap between predicted and ground-truth masks. Model optimization was performed using the Adam optimizer with an adaptive learning schedule, and early stopping was applied to prevent overfitting. Segmentation performance was evaluated using Accuracy, Intersection over Union (IoU), and Dice score. The dataset composition and training configuration are summarized in Table 4.

Table 4 summarizes the dataset composition and training configuration used for the stroma segmentation models, while Table 5 reports the model size and training time of the evaluated architectures.

3.8.3. TILs quantification module

TILs quantification was formulated as a regression task that predicts a continuous stromal TIL score from spatial representations of immune infiltration. The model receives two-channel binary maps as input, representing the predicted TILs-within-stroma mask and the corresponding stroma mask for each tile. These maps capture the spatial relationship between lymphocytes and stromal regions, enabling the network to learn discriminative patterns associated with stromal TIL density.

The regression network operates on 256 × 256 tiles with two input channels. Model optimization was performed using the Adam optimizer with an initial learning rate of 1×10^{-4}. Mean Squared Error (MSE) was used as the loss function, while Mean Absolute Error (MAE) served as the primary evaluation metric. Training was conducted with early stopping to prevent overfitting, and the model achieving the lowest validation loss was retained for final evaluation. The dataset composition and training configuration are summarized in Table 6.

Table 6 summarizes the dataset composition and training configuration used for the TILs quantification model, while Table 7 reports the model size and training time.

Table 2
Dataset composition and training hyperparameters for the TILs segmentation models.

Parameter	Value
Input patch size	256 × 256
Training tiles	22,695
Validation tiles	1766
Testing tiles	2087
Batch size	8
Optimizer	Adam
Initial learning rate	1×10^{-4}
Learning rate scheduler	ReduceLROnPlateau
Maximum epochs	100
Early stopping patience	10 epochs
Loss function	Tversky focal loss + IoU loss
Tversky parameters	$\alpha = 0.8,\ \beta = 0.2$
Focal parameter	$\gamma = \frac{4}{3}$
Evaluation metrics	Accuracy, Dice, Tversky, Precision, Recall, F1-score

Table 3
Model parameters and training time of the evaluated TILs segmentation architectures.

Model	Parameters (M)	Training time (h)
U-Net	37.67	3.66
Attention U-Net	34.88	3.85
TransUNet	42.72	9.20

Table 4
Dataset composition and training configuration for the stroma segmentation models. Training and validation tile counts are reported as mean ± standard deviation across the five cross-validation folds, while the test set is fixed.

Parameter	Value
Input tile size	256 × 256
Training tiles (per fold)	12902 ± 372
Validation tiles (per fold)	3225 ± 372
Test tiles (fixed)	5125
Batch size	32
Optimizer	Adam
Initial learning rate	1×10^{-4}
Maximum epochs	50
Early stopping patience	10 epochs
Loss function	Dice loss
Backbone initialization	ImageNet pretrained
Evaluation metrics	Accuracy, IoU, Dice score

Table 5
Model complexity and training time for the evaluated stroma segmentation architectures. Training time per fold is reported as mean ± standard deviation across the five folds.

Model	Parameters (M)	Train time/Fold (h)	Total training time (h)
ResNet101-DeepLabV3+	30.96	0.59 ± 0.09	2.96
ResNeSt-DeepLabV3+	63.25	1.18 ± 0.30	5.89
MobileViT-DeepLabV3+	7.70	1.04 ± 0.33	5.20
EfficientViT-DeepLabV3+	19.52	0.58 ± 0.17	2.92

Table 6
Dataset composition and training configuration for the TILs quantification model.

Parameter	Value
Input tile size	256 × 256
Input channels	2 (TILs within stroma mask, Stroma mask)
Training samples	1570
Validation samples	393
Batch size	16
Optimizer	Adam
Initial learning rate	1×10^{-4}
Maximum epochs	100
Early stopping patience	20 epochs
Loss function	Mean Squared Error (MSE)
Evaluation metric	Mean Absolute Error (MAE)

Table 7
Model complexity and training time for the TILs quantification network.

Model	Parameters (M)	Total training time (h)
CNN Regression model	0.11	0.64

Table 8
Baseline performance of segmentation models before active learning. Best values in **bold**.

Model	Segmentation metrics		Cell detection metrics		
	Dice	Tversky	Precision	Recall	F1 Score
U-Net	0.667	**0.718**	0.708	**0.724**	0.716
Attention U-Net	**0.668**	0.703	**0.733**	0.708	**0.720**
TransUNet	0.665	0.687	0.692	0.705	0.698

Table 9
False positives and false negatives before and after active learning refinement.

Error type	Before	After
False Positives	2732	1494 (−**45.3%**)
False Negatives	1769	1228 (−**30.6%**)

4. Results & discussion

This section presents the experimental results of the proposed modular pipeline for automated TILs scoring in breast cancer H&E slides. Results for each module—TILs segmentation, stroma segmentation, and regression-based TILs quantification—are reported independently to ensure interpretability and transparency. Quantitative and qualitative evaluations highlight each component's contribution to the overall system. Finally, the complete pipeline is assessed at the ROI level by comparing global predicted scores with expert annotations using the validation metrics.

4.1. TILs segmentation module results

This section summarizes the performance of the three evaluated architectures—U-Net, Attention U-Net, and TransUNet—before and after the active learning refinement step. Both pixel-wise segmentation metrics and cell-level detection metrics are reported to provide a comprehensive assessment of model behavior.

4.1.1. Baseline results before active learning

Before applying active learning, all models were trained on the initial annotations and evaluated on the held-out test set. The pixel-wise segmentation results in Table 8 show that U-Net and Attention U-Net achieved very similar Dice Scores (0.667–0.668). U-Net provided the highest Tversky Score (0.718), suggesting slightly better balance between false positives and false negatives. TransUNet obtained slightly lower scores, which may reflect its greater dependence on high-quality supervision when trained on limited data.

Cell-level detection results (Table 8) indicate that Attention U-Net achieved the strongest overall performance, reaching the highest F1 Score (0.720) and strongest precision. U-Net showed comparable performance, whereas TransUNet tended to produce more false positives at this early stage.

These results suggest that all models were able to learn meaningful representations, but that TransUNet — being the most complex architecture — would likely benefit the most from improved annotations. This motivated the use of expert-guided active learning.

4.1.2. Active learning findings

Active learning was performed on 10 ROIs using the best baseline model (Attention U-Net). The expert corrected systematic false positives and false negatives, producing a more reliable training set. Table 9 summarizes the reduction in errors.

The expert corrections mainly addressed over-segmentation of densely packed lymphocytes and missed detections of small or faint cells. These refinements were particularly valuable for TransUNet, which responds more strongly to improved supervision.

4.1.3. Results after active learning

After retraining with the refined annotations, all three models showed measurable improvements in pixel-wise segmentation (Table 10). Dice Scores increased and became more consistent across models. TransUNet achieved the highest Dice Score (0.719), while U-Net achieved the highest Tversky Score (0.726).

Cell-level performance improved for all three models (Table 10). TransUNet achieved the highest post–active learning precision (0.821), reflecting a notable reduction in false positives. All architectures reached similar F1 Scores around 0.776–0.780.

A before/after comparison is provided in Table 11. All models benefited from the refined annotations, but TransUNet showed the largest relative gains across precision, recall, and F1 Score. This suggests that the expert corrections were particularly valuable for architectures that rely heavily on fine-grained supervision.

Beyond segmentation accuracy, we also evaluated the computational efficiency of the evaluated architectures to better understand their practical deployment characteristics.

4.1.4. Inference efficiency

Inference time was measured during GPU forward passes using 256 × 256 input tiles. This metric reflects the computational cost of the segmentation models when processing individual image tiles during inference.

Table 12 reports the average inference time per tile for the evaluated architectures.

In summary, active learning played a key role in improving segmentation quality, particularly for the TransUNet model. The expert-guided corrections helped address systematic errors and led to increased precision and more stable detection performance across all architectures. Among the evaluated models, TransUNet achieved the highest precision in TILs detection after active learning and was therefore selected for subsequent stages of the pipeline. In terms of computational efficiency, all architectures demonstrated comparable tile-level inference times (approximately 4.6–5.0 ms per tile). While U-Net and Attention U-Net exhibited slightly faster inference, the difference remained small, indicating that the additional transformer components in TransUNet introduce only a minor computational overhead while providing improved detection precision.

4.1.5. Explainability results

Fig. 7 illustrates example SegGradCAM visualizations for TransUNet after active learning. The heatmaps highlight spatial regions that most strongly influence the model's decisions.

The explainability results confirm that TransUNet attends to appropriate histological structures, supporting model trustworthiness and

Table 10
Baseline performance after active learning. Best values in **bold**.

Model	Segmentation metrics		Cell detection metrics		
	Dice	Tversky	Precision	Recall	F1 Score
U-Net	0.718	**0.726**	0.768	**0.792**	**0.780**
Attention U-Net	0.715	0.712	0.776	0.779	0.778
TransUNet	**0.719**	0.703	**0.821**	0.735	0.776

Table 11
Comparison of cell-level detection metrics before and after active learning (updated results). Percentage changes shown in parentheses.

Model	Before			After		
	Precision	Recall	F1	Precision	Recall	F1
U-Net	0.708	0.724	0.716	0.768 (+8.5%)	0.792 (+9.4%)	0.780 (+9.0%)
Attention U-Net	0.733	0.708	0.720	0.776 (+5.9%)	0.779 (+10.0%)	0.778 (+8.1%)
TransUNet	0.692	0.705	0.698	**0.821 (+18.7%)**	0.735 (+4.3%)	0.776 (+11.2%)

Table 12
Tile-level inference time of the evaluated TILs segmentation models measured on 256 × 256 input tiles. Values are reported as mean ± standard deviation in milliseconds.

Model	Inference/Tile (ms)
U-Net	4.73 ± 0.14
Attention U-Net	4.58 ± 0.08
TransUNet	4.97 ± 0.07

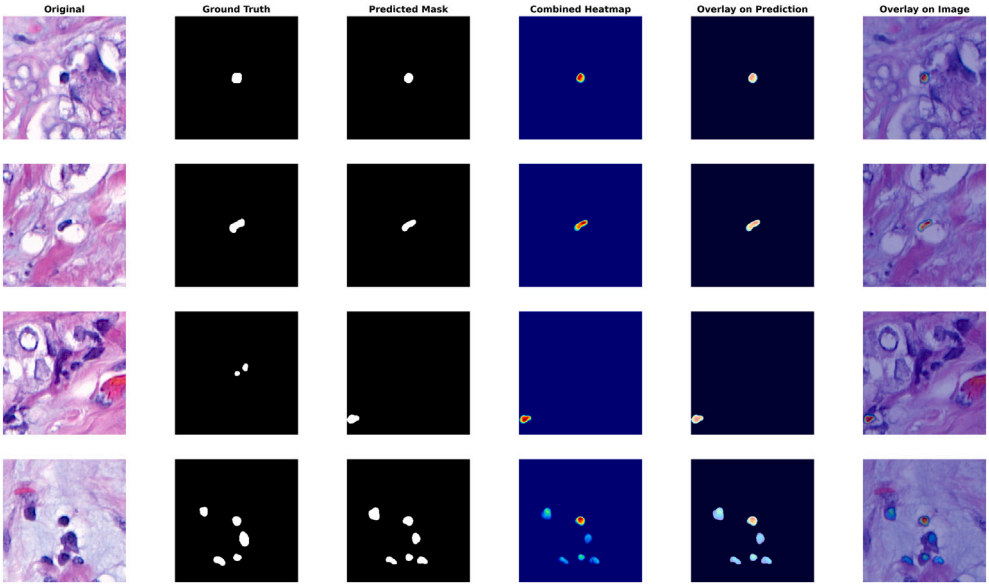


Fig. 7. SegGradCAM explainability. From left: input patch, ground truth, prediction, heatmap, overlay on prediction, and overlay on input. The model's attention focuses on morphologically relevant lymphocyte regions.

clinical interpretability. Instance-wise heatmaps enable localized inspection of TIL detections, which is particularly important in dense or ambiguous tissue regions.

4.2. Stroma segmentation module results

This section presents the quantitative and qualitative results of the proposed stroma segmentation framework. Four SE-enhanced DeepLabV3+ variants — using ResNet101, ResNeSt101, MobileViT, and EfficientViT encoders — were trained following the methodology described in Section 3.4. Performance was assessed on (a) the held-out test subset of the public dataset to evaluate in-domain segmentation quality, and (b) the private dataset to measure cross-domain generalization.

4.2.1. Performance on the public dataset

Table 13 summarizes segmentation performance on the public dataset. All four backbones achieved similarly high accuracy, IoU, and Dice scores, with mean Dice values above 0.83 and IoU values above 0.74. The small performance differences indicate that the SE-enhanced DeepLabV3+ decoder provides strong and stable feature extraction regardless of encoder choice in this well-standardized dataset.

Among the backbones, **EfficientViT** achieved the highest mean IoU (0.753) and Dice score (0.845), followed closely by **ResNeSt101** and **MobileViT**. This slight advantage is likely due to EfficientViT's ability to capture broader contextual information in homogeneous, high-quality images, giving it a small edge when domain variability is minimal.

Table 13
Quantitative performance of different backbones on the public dataset (test subset). Metrics represent mean ± standard deviation across five folds. Best values in **bold**.

Backbone	Accuracy	IoU	Dice
ResNet101	0.855 ± 0.008	0.742 ± 0.015	0.836 ± 0.012
ResNeSt101	0.862 ± 0.007	0.749 ± 0.013	0.842 ± 0.009
MobileViT	0.863 ± 0.009	0.751 ± 0.007	0.841 ± 0.005
EfficientViT	0.859 ± 0.024	**0.753 ± 0.024**	**0.845 ± 0.018**

Table 14
Cross-domain evaluation results on the private dataset. Metrics represent mean ± standard deviation across five folds. Best values in **bold**.

Backbone	Accuracy	IoU	Dice
ResNet101	0.689 ± 0.031	0.589 ± 0.072	0.715 ± 0.068
ResNeSt101	0.707 ± 0.055	0.595 ± 0.088	0.725 ± 0.077
MobileViT	**0.736 ± 0.029**	**0.646 ± 0.046**	**0.769 ± 0.039**
EfficientViT	0.693 ± 0.094	0.581 ± 0.159	0.702 ± 0.156

Table 15
Tile-level inference time of the evaluated stroma segmentation backbones measured on 256 × 256 input tiles. Values are reported as mean ± standard deviation in milliseconds.

Backbone	Inference/Tile (ms)
ResNet101	2.16 ± 0.67
ResNeSt101	3.16 ± 0.85
MobileViT	2.46 ± 1.00
EfficientViT	2.29 ± 1.05

4.2.2. Cross-domain evaluation on the private dataset

To assess robustness under real clinical variability, the trained models were evaluated on 60 private ROIs from a different institution without any fine-tuning. As shown in Table 14, performance decreased for all models due to domain shift resulting from differences in staining, scanners, and tissue preparation. However, the **MobileViT-based model** achieved the highest cross-domain IoU (0.646) and Dice score (0.769), with the lowest variance across folds.

In contrast, EfficientViT — despite achieving the best in-domain results — experienced the largest degradation on private data. This suggests that EfficientViT's strong reliance on attention-driven global context makes it more sensitive to color and staining disparities across institutions. MobileViT, which combines convolutional inductive biases with lightweight transformer blocks, appears more robust to such variability and better able to preserve discriminative local texture information.

These results show that although all backbones perform similarly in-domain, **MobileViT is the only architecture that maintains strong performance under distributional shift**. Its hybrid CNN–Transformer design allows it to retain robust local feature extraction while still capturing enough global structure to generalize across imaging conditions. This makes it the most reliable choice for multi-center deployment.

4.2.3. Inference efficiency

In addition to segmentation accuracy, the computational efficiency of the evaluated backbones was analyzed to assess their practical applicability. Inference time was also measured during GPU forward passes using 256 × 256 input tiles. This metric reflects the computational cost of each architecture when processing individual tiles during inference.

Table 15 summarizes the average inference time per tile for the evaluated stroma segmentation backbones.

4.2.4. Backbone selection

Considering both segmentation performance and computational efficiency, the **MobileViT-based DeepLabV3+ model** was selected as the final stroma segmentation module for the end-to-end TILs scoring pipeline. It achieved the highest cross-domain Dice and IoU with low variance across folds while maintaining efficient tile-level inference, providing a strong balance between robustness and computational cost.

4.2.5. Explainability results

Explainability analysis using SegGradCAM [32] was performed on the MobileViT-based model. Representative heatmaps in Fig. 8 show that the model consistently focuses on stromal structures such as fibrous extracellular matrix and elongated fibroblasts. Activation overlays align well with ground truth boundaries, indicating that the model relies on biologically meaningful features rather than artifacts.

These visualizations confirm the interpretability and reliability of the selected backbone for integration into the TILs scoring pipeline.

4.3. TILs quantification module results

This section reports the behavior and performance of the regression module responsible for predicting continuous TILs scores from the stroma and TILs-inside-stroma binary masks.

4.3.1. Training and convergence behavior

The model was trained for up to 100 epochs using the configuration described in Section 3.5. Training and validation curves for MAE and MSE showed smooth and parallel convergence, indicating stable learning without overfitting. Detailed curves are included in the Supplementary Material (Figures S4 & S5).

4.3.2. Quantitative evaluation on validation data

Performance on the held-out validation set was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Results are provided in Table 16.

The low MSE (0.003) and MAE (0.033) indicate that predicted scores closely match expert annotations. The R^2 value of 0.8977 shows that the model explains nearly 90% of the variance in true scores. Fig. 9 further confirms strong agreement through tight clustering around the identity line.

4.3.3. Inference efficiency

In addition to predictive accuracy, the computational efficiency of the regression model was evaluated to assess its suitability for integration within the proposed pipeline. Inference time was measured

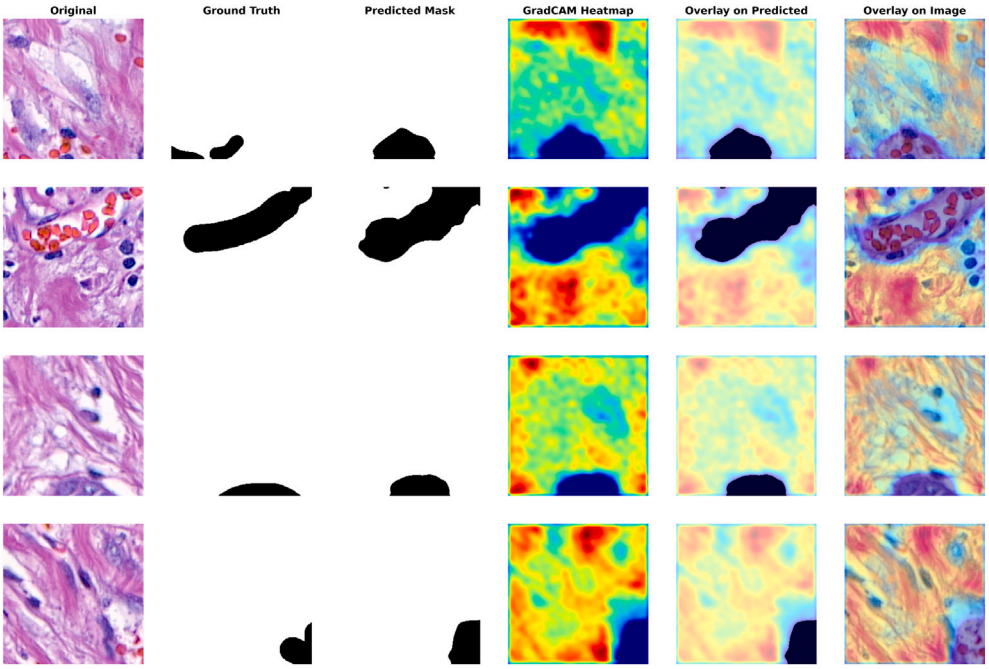


Fig. 8. SegGradCAM visualization of the MobileViT-based DeepLabV3+ model. From left to right: original histopathological image, ground truth mask, predicted mask, GradCAM heatmap, overlay on predicted mask, and overlay on the original image. The highlighted regions correspond to areas that most strongly influenced the model's prediction, showing that the model attends to stromal structures consistent with histopathological features.

Table 16
Performance of the TILs regression model on the validation dataset.

Metric	MSE	MAE	R^2 Score
Value	0.003	0.033	0.8977

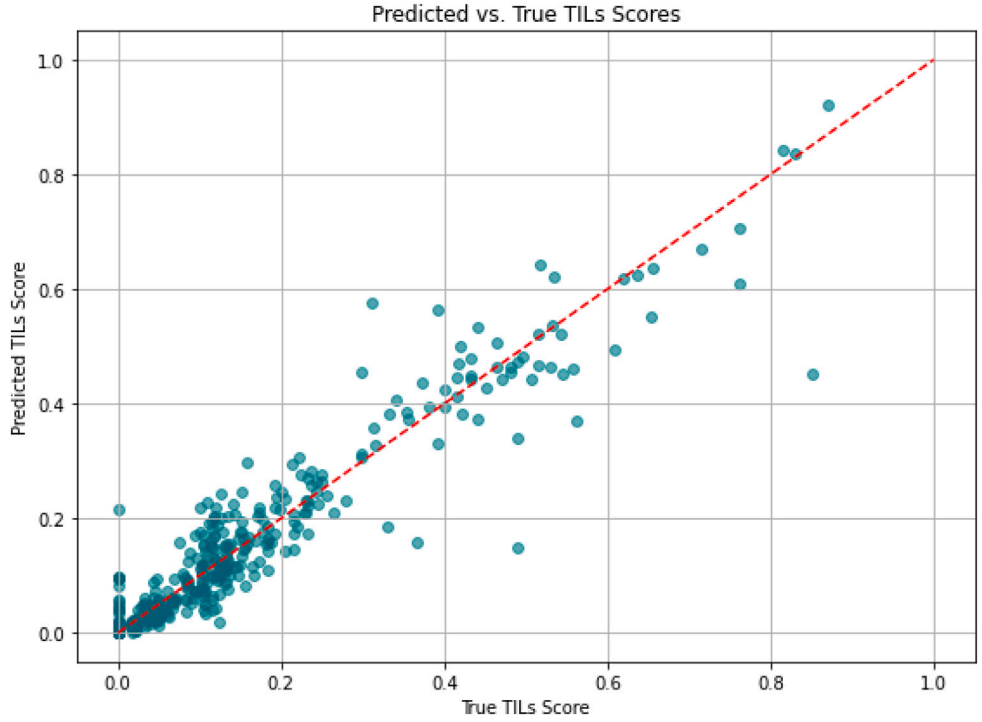


Fig. 9. Predicted versus true TILs scores. Each point corresponds to a predicted patch-level score and its ground truth value. The red dashed line represents the ideal $y = x$ relation.

Table 17
Agreement analysis of the automated TILs scoring pipeline on Subset 1 (60 WSIs) for the best configuration ($T_r = 0.15, P_{95}$). Continuous agreement metrics (Pearson, Spearman, ICC, CCC) and categorical agreement (Cohen's Kappa) are reported. Best values in **bold**.

Reference	Pearson	Spearman	ICC	CCC	Kappa
Pathologist 1	0.785	0.725	0.714	0.711	**0.768**
Pathologist 2	0.691	0.652	0.663	0.659	0.738
Mean (1 & 2)	**0.784**	0.738	**0.743**	**0.740**	0.665

during GPU forward passes using 256 × 256 input tiles containing the two-channel binary masks.

The regression network demonstrated highly efficient inference, requiring an average of 1.32 ± 0.75 ms **per tile**. This low computational cost is expected given the lightweight architecture of the network (0.11M parameters) and confirms that the regression stage adds minimal overhead to the overall pipeline.

These results indicate that the TILs quantification module can efficiently process large numbers of tiles while maintaining accurate score prediction, supporting its applicability in practical digital pathology workflows.

4.3.4. Interpretation

Despite using only binary inputs, the regression model achieves strong predictive performance because the masks encode all clinically relevant information for stromal TILs assessment. The stroma mask restricts prediction to the appropriate anatomical compartment, while the TILs-inside-stroma mask captures both the density and spatial distribution of infiltrating lymphocytes, the primary determinants of TILs scoring according to IIOBWG guidelines. Moreover, binary masks suppress irrelevant appearance variations (e.g., staining differences or illumination changes), reducing noise and improving generalization.

These findings validate the use of a lightweight CNN for regression, demonstrating that the model effectively learns the quantitative relationship between stromal immune infiltration patterns and continuous TILs scores.

4.4. Evaluation results

This section reports the performance of the proposed ROI-level TILs scoring framework on the two independent datasets described in Section 3.1.1. The evaluation includes correlation-based metrics (Pearson, Spearman, ICC, CCC) and categorical agreement (Cohen's Kappa).

4.4.1. Subset 1 (60 WSIs, dual expert assessment)

Subset 1 contains 60 ROIs independently evaluated by two expert pathologists. The best-performing configuration was obtained at $T_r = 0.15$ and P_{95}. Although the agreement metrics remained broadly consistent across the evaluated threshold and percentile combinations (Supplementary Tables S3–S5), this configuration yielded the highest overall agreement with the expert reference and was therefore selected for reporting. As shown in Table 17, the model achieved strong correlations with both readers, with Pearson values of 0.785 (Pathologist 1) and 0.691 (Pathologist 2). Agreement with the averaged reference was highest (Pearson 0.784, ICC 0.743, CCC 0.740), indicating that the automated scores closely match consensus expert grading.

Substantial categorical agreement was also obtained, with Kappa values of 0.768 (Pathologist 1), 0.738 (Pathologist 2), and 0.665 for the averaged assessment. These results confirm that the model is consistent across both continuous and categorical evaluation perspectives. In addition, the average end-to-end processing time of the proposed pipeline was **12.92 ± 2.14 s per ROI** on Subset 1 when executed on a single GPU, demonstrating that the framework can perform automated ROI-level assessment within a few seconds. Further reductions in processing time could be achieved through parallel execution or multi-GPU deployment when analyzing large batches of slides.

Overall, Subset 1 results demonstrate strong agreement with both individual and averaged expert assessments, showing that the framework is robust to inter-observer variability and accurately reproduces expert scoring trends.

Molecular Subtype Analysis: To further investigate the robustness of the proposed framework across different biological contexts, we performed a correlation analysis stratified by molecular subtype using the averaged reference scores of the two pathologists. The results are summarized in Table 18.

For the Luminal A subtype ($N = 27$), the automated scores showed strong agreement with the expert reference, achieving a Pearson correlation of 0.834 and an ICC of 0.835, indicating a high level of consistency between automated and expert assessments.

For Luminal B (HER2±) tumors ($N = 29$), moderate correlations were observed, with Pearson and Spearman coefficients of 0.632 and 0.640, respectively. Although agreement metrics were lower compared to Luminal A, the results still demonstrate a meaningful relationship between automated and expert scoring in this subtype.

For the Triple-Negative subtype ($N = 3$), very high correlations were obtained (Pearson = 0.972, Spearman = 1.000). However, these values should be interpreted with caution due to the limited number of available samples in this subset.

Only a single HER2-enriched case was present in Subset 1; therefore, a reliable statistical correlation analysis could not be performed for this subtype.

Overall, these findings suggest that the proposed framework maintains strong agreement with expert assessments across the major molecular subtypes represented in the dataset. However, the distribution of molecular subtypes within Subset 1 is not uniform, with Luminal A and Luminal B tumors representing the majority of samples ($N = 27$ and $N = 29$, respectively), while Triple-Negative and HER2-enriched subtypes are underrepresented. Consequently, the strong correlations observed for Luminal A tumors provide more reliable evidence of the model's performance in this subtype due to the relatively larger sample size.

In contrast, the very high correlation observed for the Triple-Negative subgroup should be interpreted cautiously, as it is based on only three samples, which may artificially inflate correlation metrics. Similarly, the presence of only a single HER2-enriched case prevents meaningful statistical evaluation for this clinically important subtype. These limitations highlight the challenges associated with evaluating automated TIL scoring systems on subtype-stratified cohorts, where certain molecular subtypes may be sparsely represented.

Nevertheless, the consistent agreement observed in the well-represented Luminal subtypes suggests that the proposed framework captures biologically relevant patterns of stromal lymphocyte infiltration. Future studies incorporating larger and more balanced subtype distributions will be important to further validate the generalizability of the model, particularly for subtypes such as HER2-enriched and Triple-Negative breast cancer, where TILs assessment has strong prognostic and therapeutic relevance.

4.4.2. Subset 2 (112 WSIs, single expert assessment)

Subset 2 provides a larger independent cohort for assessing generalization. The best-performing configuration was $T_r = 0.15$ and P_{99}. Although correlation metrics remained relatively stable across the evaluated threshold and percentile combinations (Supplementary Table S6),

Table 18
Correlation analysis between automated TIL scores and the averaged expert reference across molecular subtypes in Subset 1.

Subtype	N	Pearson	Spearman	ICC	CCC
Luminal A	27	0.834	0.735	0.835	0.829
Luminal B (HER2+/-)	29	0.632	0.640	0.492	0.517
Triple Negative	3	0.972	1.000	0.847	0.786

Table 19
Agreement analysis of the automated TILs scoring pipeline on Subset 2 (112 WSIs) for the best configuration ($T_r = 0.15, P_{99}$). Continuous agreement metrics (Pearson, Spearman, ICC, CCC) and categorical agreement (Cohen's Kappa) are reported.

Configuration	Pearson	Spearman	ICC	CCC	Kappa
(0.15, 99)	0.814	0.800	0.808	0.807	0.749

Table 20
Correlation analysis between automated TIL scores and the expert reference across molecular subtypes in Subset 2.

Subtype	N	Pearson	Spearman	ICC	CCC
Luminal A	49	0.704	0.636	0.699	0.694
Luminal B (HER2±)	49	0.811	0.772	0.803	0.801
HER2-enriched	4	0.666	0.949	0.591	0.536
Triple Negative	10	0.770	0.579	0.789	0.768

this configuration achieved the highest agreement with the expert reference and was therefore selected for reporting. As shown in Table 19, the model achieved high correlations with the expert reference (Pearson 0.814, Spearman 0.800, ICC 0.808, CCC 0.807). These values are slightly higher than those observed in Subset 1, indicating good generalization to unseen tissue and staining variations.

The corresponding categorical agreement is also substantial (Kappa 0.749), confirming that the best configuration not only maximizes numerical correlation but also preserves interpretive consistency in category assignments. In addition, the average end-to-end processing time of the proposed pipeline on Subset 2 was **15.91 ± 4.75 s per ROI** when executed on a single GPU. Similar to Subset 1, the runtime remains within a few seconds per ROI, and further reductions could be achieved through parallel or multi-GPU execution.

In summary, Subset 2 results highlight the model's strong generalization ability, with consistently high correlation and agreement values across a larger and heterogeneous dataset.

Molecular Subtype Analysis: Correlation analysis stratified by molecular subtype was performed to examine the behavior of the proposed framework across different tumor subgroups. The results are summarized in Table 20.

For Luminal A tumors ($N = 49$), the automated scores showed moderate to strong agreement with the expert reference, with Pearson and ICC values of 0.704 and 0.699, respectively, indicating consistent estimation of stromal TIL levels within this subtype.

Luminal B (HER2±) tumors ($N = 49$) exhibited the highest agreement among the evaluated subtypes, achieving Pearson and Spearman correlations of 0.811 and 0.772, respectively, together with strong ICC (0.803) and CCC (0.801) values. These results suggest particularly reliable performance of the pipeline in this subgroup.

For HER2-enriched tumors ($N = 4$), moderate correlations were observed (Pearson 0.666, ICC 0.591). However, these values should be interpreted cautiously due to the limited number of samples available for this subtype.

For Triple-Negative breast cancer (TNBC) ($N = 10$), strong Pearson (0.770) and ICC (0.789) values were obtained, although the Spearman correlation was comparatively lower (0.579). Overall, the automated scores remained in good agreement with the expert reference within this subtype.

Overall, the subtype-stratified analysis in Subset 2 provides a more comprehensive evaluation due to the larger cohort size compared with Subset 1. In particular, the Luminal A and Luminal B subtypes are well represented ($N = 49$ each), allowing a more reliable assessment of model behavior within these groups. The strong correlations observed for Luminal B tumors, together with the consistent performance in Luminal A cases, suggest that the proposed framework effectively captures variations in stromal lymphocyte infiltration across these commonly occurring breast cancer subtypes.

The increased number of Triple-Negative cases in this subset ($N = 10$) also provides a more informative evaluation compared with Subset 1. The observed Pearson and ICC values indicate good agreement with expert assessment, supporting the applicability of the framework to this biologically distinct subtype. Nevertheless, the sample size remains relatively modest, and further validation on larger TNBC cohorts would be beneficial to confirm the robustness of these findings.

HER2-enriched tumors remain underrepresented ($N = 4$), which limits the statistical confidence of the corresponding correlation metrics. As a result, the performance of the proposed framework in this subtype should be interpreted with caution.

Taken together, these results suggest that the proposed pipeline demonstrates consistent performance across the major molecular subtypes represented in the dataset, while also highlighting the importance of larger and more balanced subtype distributions to further strengthen conclusions regarding subtype-specific generalizability.

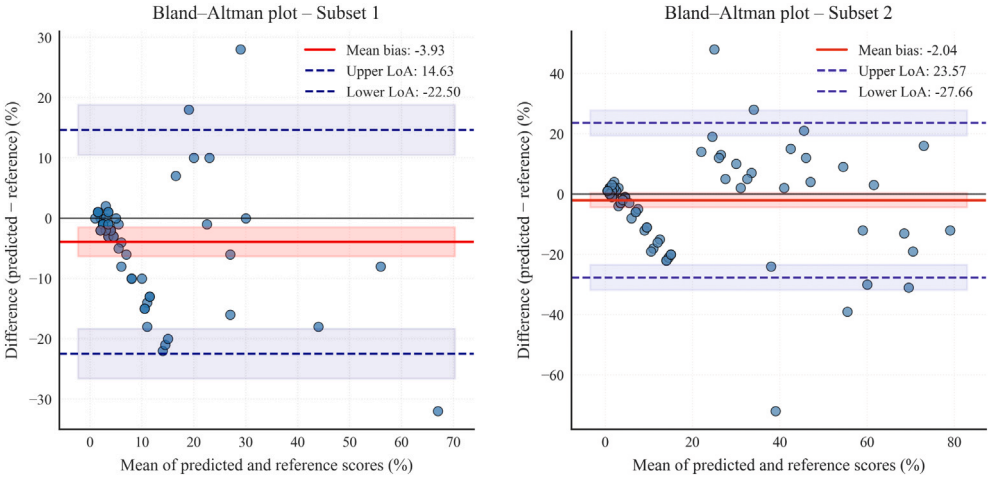
4.4.3. Agreement and calibration analysis

A quantitative summary of the numerical agreement and calibration results for both evaluation subsets is presented in Table 21. Across the two datasets, the model shows small mean biases, acceptable limits of agreement, and stable linear calibration behavior. These findings complement the previously reported correlation and categorical agreement results and indicate that the system produces reliable and interpretable continuous TILs estimates.

To illustrate these findings, the Bland–Altman plots in Fig. 10 show the distribution of prediction differences across the range of reference scores. Both subsets exhibit modest bias and consistent limits of agreement. The calibration curves in Fig. 11 provide a visual assessment of numerical alignment with expert scores, demonstrating a clear linear relationship and slightly improved calibration on the larger second subset.

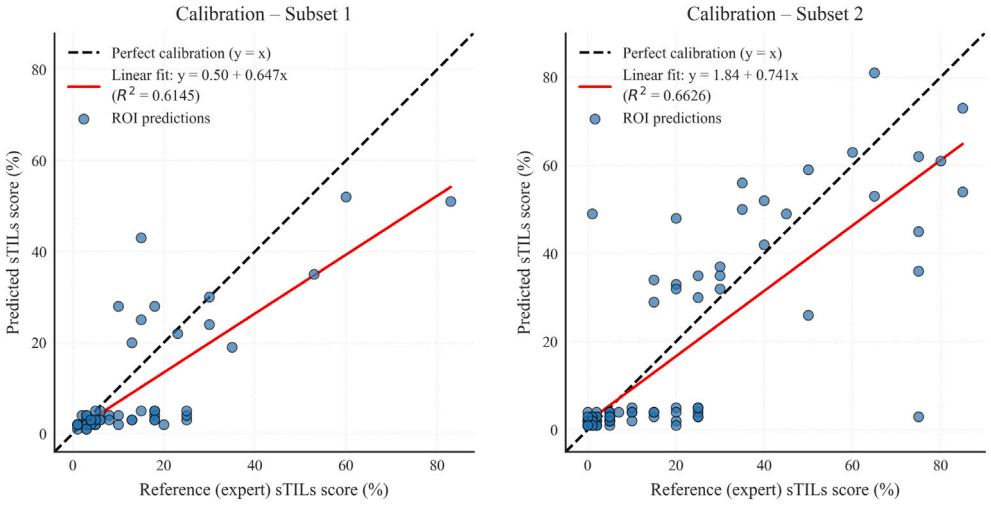
Table 21
Summary of agreement and calibration metrics for both subsets.

Metric	Subset 1 (60 WSIs)	Subset 2 (112 WSIs)
Bland–Altman Analysis		
Mean Bias (%)	−3.93	−2.04
95% CI Mean Bias (%)	[−6.33, −1.54]	[−4.47, 0.38]
SD of Differences (%)	9.47	13.07
Upper LoA (%)	14.63	23.57
Lower LoA (%)	−22.50	−27.66
Calibration Analysis		
Linear Slope	0.647	0.741
Linear Intercept	0.498	1.843
R^2 (Linear Fit)	0.615	0.663
p-value	< 0.001	< 0.001



(a) Bland–Altman plot for Subset 1 (b) Bland–Altman plot for Subset 2

Fig. 10. Bland–Altman analysis of ROI-level TILs predictions for both subsets. Mean biases and limits of agreement (LoA) are shown.



(a) Calibration curve for Subset 1 (b) Calibration curve for Subset 2

Fig. 11. Linear calibration analysis comparing predicted and expert TILs scores. The dashed line indicates perfect calibration; the solid line shows the fitted regression.

4.4.4. Ablation study: Hybrid method validation

To assess the contribution of each aggregation component and to validate the proposed adaptive selection mechanism, we compared four global TILs scoring strategies: (1) global averaging [24], (2) stroma-weighted averaging, (3) fixed percentile aggregation, and (4) the proposed hybrid method, which switches between weighted and percentile-based scoring depending on infiltration density.

Patch-Level Score Distribution: To empirically validate the intuition described in the methodology — that patch-level TILs scores within an ROI are typically highly skewed — we first examined the

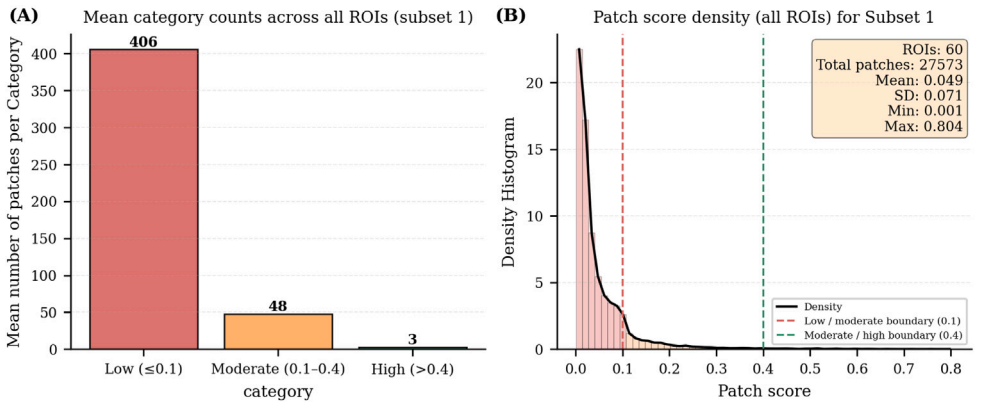


Fig. 12. Patch-level TILs score distribution across all ROIs in Subset 1. The strong skew toward low infiltration highlights why mean-based aggregation suppresses clinically meaningful regions and motivates adaptive percentile-aware strategies.

Table 22
Comparative performance of global TILs scoring methods on Subset 1 (60 WSIs, averaged across two experts).

Method	Pearson	Spearman	ICC	CCC	Kappa
Global Average [24]	0.782	0.627	0.295	0.293	0.609
Stroma-Weighted Average	0.790	0.735	0.282	0.280	0.609
Percentile (P_{95})	0.789	0.640	0.717	0.714	0.542
Hybrid (Proposed)	**0.784**	**0.738**	**0.743**	**0.740**	**0.665**

distribution of predicted patch scores across all ROIs in Subset 1. As shown in Fig. 12, more than 85%–90% of stromal patches fall within the low-infiltration range (≤ 0.1), while only a small fraction exhibit moderate or high infiltration. This confirms that the majority of patches contribute minimal immune signal, causing mean-based aggregation methods to be dominated by low scores. As anticipated, this leads to compressed global-score variance and weak absolute agreement metrics (ICC, CCC), even when Pearson or Spearman correlations remain relatively high. The observed skewness therefore provides direct evidence supporting the need for percentile-aware or hybrid aggregation strategies that give appropriate weight to infiltrated regions. A similar score distribution pattern is observed in Subset 2, and the corresponding plot is provided in the Supplementary Material (Figure S6).

Subset 1 (60 WSIs) Ablation Results: Table 22 summarizes the performance of all four aggregation strategies. Both global averaging and stroma-weighted averaging achieved reasonable Pearson correlation (0.782–0.790) but very poor ICC/CCC (≤0.30), confirming that mean-based methods fail to capture clinically meaningful agreement. The P_{95} percentile method substantially improved ICC (0.717) and CCC (0.714) but reduced Spearman correlation (0.640).

The hybrid method achieved the most balanced performance across all metrics (Pearson 0.784, ICC 0.743, CCC 0.740, Kappa 0.665), outperforming all baselines and offering a 43% ICC improvement over global averaging and a further gain over the percentile method. These results demonstrate that adaptively emphasizing infiltrated patches yields stronger agreement with expert assessment than any fixed aggregation rule.

Subset 2 (112 WSIs) Ablation Results: Results on the larger independent dataset (Table 23) closely mirrored the findings on Subset 1. Both averaging methods again showed good correlation (0.771–0.786) but extremely weak ICC/CCC (≤0.26). The percentile method (P_{99}) substantially improved agreement (ICC 0.704), but categorical agreement (Kappa 0.264) remained unstable.

The hybrid method once more provided the strongest and most stable performance across all metrics (Pearson 0.814, ICC 0.808, CCC 0.807, Kappa 0.749). Notably, it achieved a 227% improvement in ICC over global averaging and a 218% improvement over stroma-weighted averaging, while also outperforming the percentile method.

Comprehensive Interpretation: Across both datasets, mean-based methods consistently failed to achieve acceptable absolute agreement, confirming that correlation alone is insufficient for clinically reliable scoring. Percentile aggregation improved ICC/CCC but showed instability across datasets. Only the hybrid method achieved consistently high correlation, strong absolute agreement, and robust categorical alignment. These findings validate that adaptive weighting — guided by the distribution of local infiltration — is essential for reliable ROI-level TILs quantification. The Subset 2 patch-distribution plot supporting this conclusion is provided in the Supplementary Material.

4.5. Comparison with state-of-the-art method

To further evaluate the effectiveness of the proposed framework, we compared its performance with the method introduced by Kim et al. [24]. For a consistent and fair comparison, the method described in [24] was reimplemented and applied to our in-house datasets, using the same ROI inputs and evaluation metrics as those used for the proposed framework.

Performance was assessed using correlation analysis for continuous TILs-score predictions and Cohen's Kappa coefficient for discretized infiltration categories on Subsets 1 and 2. Subset 1 contains independent visual TILs assessment from two expert raters, whereas Subset 2 includes visual scores from a single expert pathologist.

For the proposed framework, global TILs scores were computed using the parameter configurations identified as optimal in the previous sections: $T_r = 0.15, P_{95}$ for Subset 1 and $T_r = 0.15, P_{99}$ for Subset 2. The comparative results are summarized in Table 24.

As shown in Table 24, the proposed framework consistently achieves higher agreement with expert assessments across both datasets. On Subset 1, the proposed method obtained a Pearson correlation of 0.784 compared with 0.534 for the baseline, while the categorical agreement improved from a Kappa value of 0.333 to 0.665. Similarly, on Subset 2 the proposed framework achieved Pearson and Kappa values of 0.814 and 0.749, respectively, compared with 0.358 and 0.191 for the baseline approach.

These improvements can be attributed to several design components of the proposed pipeline. The active-learning–driven refinement of the TILs segmentation module improves the quality of lymphocyte detection and reduces annotation inconsistencies. In addition, the regression-based quantification module enables continuous estimation of stromal TILs density, while the hybrid aggregation strategy

Table 23
Comparative performance of global TILs scoring methods on Subset 2 (112 WSIs with single expert assessment).

Method	Pearson	Spearman	ICC	CCC	Kappa
Global Average [24]	0.786	0.768	0.247	0.246	0.637
Stroma-Weighted Average	0.771	0.797	0.254	0.252	0.637
Percentile (P_{99})	0.810	0.768	0.704	0.703	0.264
Hybrid (Proposed)	**0.814**	**0.800**	**0.808**	**0.807**	**0.749**

Table 24
Comparative performance of Global TIL-score quantification methods.

Dataset	Method	Pearson	Spearman	ICC	CCC	Kappa
Subset 1	Kim et al. [24]	0.534	0.240	0.252	0.249	0.333
	Proposed	**0.784**	**0.738**	**0.743**	**0.740**	**0.665**
Subset 2	Kim et al. [24]	0.358	0.334	0.114	0.113	0.191
	Proposed	**0.814**	**0.800**	**0.808**	**0.807**	**0.749**

used to derive ROI-level scores from patch-level predictions enhances robustness to heterogeneous spatial distributions of immune cells.

4.5.1. Discussion and clinical adaptability

The proposed TILs scoring pipeline consistently demonstrated strong quantitative and categorical agreement with independent assessments from dual experts and single-expert evaluations. The best-performing configurations—($T_r = 0.15, P_{95}$) **for Subset 1** and ($T_r = 0.15, P_{99}$) **for Subset 2**—achieved Pearson correlations above 0.78, ICC values exceeding 0.74, and Kappa scores within the substantial agreement range ($0.65 < \kappa < 0.80$).

Although the optimal percentile parameter differed slightly between the two evaluation subsets, the correlation metrics remained relatively stable across the tested parameter combinations (see Supplementary Tables S3–S6). This observation indicates that the pipeline is not highly sensitive to the exact percentile choice. In practice, the percentile parameter primarily controls how strongly the scoring emphasizes the highest TIL-density regions within a slide. Selecting a higher percentile therefore prioritizes the most inflamed stromal areas, which are often the regions of greatest biological relevance. Consequently, using a high percentile value (e.g., P_{95}–P_{99}) provides a practical and robust choice for deployment.

In contrast, the density threshold parameter T_r plays a more critical role in determining the scoring strategy. This parameter controls the transition between the weighted-average aggregation and the percentile-based aggregation of patch scores. By evaluating the proportion of patches with elevated TIL density, the threshold effectively determines whether a slide exhibits sufficiently dense immune infiltration to justify emphasizing high-intensity regions. As a result, T_r governs how the pipeline highlights biologically meaningful areas with high lymphocyte concentration and therefore represents the key parameter that may require modest dataset-specific tuning when adapting the pipeline to new clinical cohorts.

These results confirm that the proposed pipeline can reproduce expert-level TILs scoring with high reliability across various datasets. The correlation and Kappa consistency between subsets further supports the pipeline's adaptability to varying tissue heterogeneity and staining conditions.

From a clinical perspective, the pipeline offers several advantages:

- **Automation:** The pipeline automates TILs quantification from histopathological images, effectively reducing human subjectivity and inter-observer bias inherent in visual scoring. By replacing manual assessment with objective computational analysis, it provides consistent, reproducible scoring independent of individual pathologist expertise or fatigue.
- **Adaptability:** The inclusion of configurable parameters (T_r, P_x) allows the pipeline to be adapted to different staining conditions, tissue preparation protocols, or dataset characteristics without requiring model retraining. In practical deployment, the density threshold T_r serves as the primary parameter controlling the sensitivity to high TIL-density regions, while the percentile parameter can typically be set to a high value (e.g., P_{95}–P_{99}) to consistently capture the most inflamed stromal areas.
- **Scalability:** While the pipeline currently operates on expert-defined ROIs, its modular architecture enables future extension to slide-level analysis. Once automated ROI selection methods are developed following IIOBWG guidelines (as outlined in Future Work), the patch-based aggregation framework can be applied systematically across entire slides, enabling high-throughput analysis of large histopathological datasets. This would transform TILs quantification from a labor-intensive manual process to a scalable, automated workflow suitable for multicenter studies and clinical trials.

Overall, the strong alignment with expert assessments and stable performance across independent datasets demonstrate that the proposed modular pipeline provides a reproducible, interpretable, and clinically adaptable solution for automated TILs quantification in breast cancer pathology. The pipeline's ability to balance correlation strength with absolute agreement — as demonstrated in the ablation study — positions it as a reliable tool for supporting pathologists in routine practice and facilitating the integration of artificial intelligence into digital pathology workflows.

5. Limitations

Although the proposed pipeline demonstrates strong performance and produces clinically meaningful outputs, several limitations should be acknowledged from both technical and pathological perspectives.

First, the dataset used for training the TILs segmentation module remains relatively modest in size. Detailed cell-level annotation is time-consuming and requires sustained expert involvement. Although the active learning strategy improved annotation quality and reduced false detections, some inaccuracies persist in challenging regions where lymphocytes cluster densely or where tissue structures introduce visual ambiguity. These local segmentation errors can lead to small variations in the derived TILs-in-stroma mask. Representative examples are visible in the explainability analysis (Fig. 7), where certain structures occasionally receive spurious activations. Nevertheless, the regression and aggregation components of the pipeline demonstrated robustness, and the resulting ROI-level scores remained stable. Expanding annotated datasets and incorporating semi-supervised learning or multi-observer annotations could further improve detection robustness.

Second, while the stroma segmentation module generalized well across datasets, occasional over- or under-segmentation occurs at stromal boundaries. Such boundaries can be difficult to delineate even for experienced pathologists due to variability introduced by fibrosis, tumor subtype, or inflammatory processes. These deviations may slightly affect the stroma–TILs intersection mask used during TIL quantification. Despite this, the system consistently captured the dominant spatial patterns within each ROI, and the impact on final scoring was limited. Future improvements could include multi-scale modeling, improved stain normalization, or domain adaptation techniques to further enhance boundary consistency.

Third, the reference scores used for ROI-level evaluation were generated by two experienced breast cancer pathologists. Although their independent assessments provide a strong reference standard, a larger and more diverse panel of experts would allow a more comprehensive characterization of inter-observer variability, which is an inherent challenge in manual TILs scoring. Incorporating consensus-based annotations or pathologist-in-the-loop refinement strategies could further strengthen the reliability of reference labels.

Fourth, the Bland–Altman analysis (Fig. 10) revealed a slight compression of the predicted score range, particularly for highly infiltrated regions. This type of pattern is commonly observed in biological datasets with skewed distributions dominated by low-infiltration samples. Although this effect did not substantially affect the observed agreement and correlation metrics, improved calibration strategies or distribution-aware regression losses may help preserve finer variability in predicted scores.

From a data generalization perspective, the evaluation included an independent hold-out cohort (Subset 2); however, both subsets originated from the same institutional setting. Consequently, the current validation represents internal validation across unseen patient samples rather than true multi-center validation. Differences in staining protocols, scanner characteristics, and patient populations across institutions may influence model performance. Future studies should therefore assess the proposed framework using multi-institutional datasets to further confirm its robustness.

Another limitation concerns the use of expert-defined regions of interest (ROIs) rather than fully automated whole-slide image (WSI) analysis. In this study, ROIs were manually identified according to IIOBWG guidelines to ensure that the analyzed areas contained representative tumor-associated stroma. While this approach reflects current clinical practice for TILs assessment, it does not represent a fully automated end-to-end workflow. A complete WSI-level pipeline would require an additional module to identify informative tumor-associated stromal regions automatically. These automatically detected ROIs could then be processed by the proposed pipeline, followed by aggregation of ROI-level predictions to derive a slide-level TILs score. Variability in ROI detection or automated region selection could influence the final TILs estimates, and integrating robust ROI detection and slide-level aggregation strategies represents an important direction for future work.

Finally, the dataset used in this study originates from a limited number of institutions. Although augmentation strategies were used to simulate staining and acquisition variability, such techniques cannot fully replicate the diversity encountered in real multi-center clinical environments. Broader validation across additional laboratories, staining protocols, and imaging devices will therefore be necessary to further support clinical translation.

In summary, these limitations highlight several opportunities for continued development. Despite these constraints, the proposed pipeline achieved strong agreement with expert scoring, demonstrated robust generalization to independent internal validation data, and provided interpretable outputs at each stage of the analysis. Expanding data diversity, improving segmentation robustness, and conducting larger multi-center validation studies represent important next steps toward clinical deployment.

6. Conclusion & future work

In this study, we addressed the challenge of producing consistent and clinically meaningful stromal TILs scores in breast cancer H&E slides by developing a modular deep learning pipeline aligned with IIOBWG recommendations. The system integrates three complementary components: a TransUNet-based TILs segmentation model, an SE-enhanced DeepLabV3+ stroma segmentation module, and a regression model guided by an adaptive aggregation strategy. The use of unified preprocessing, targeted augmentation, and ROI-level active learning allowed the TILs module to focus on ambiguous or difficult regions, while SegGradCAM visualizations demonstrated that the model consistently attends to morphologically meaningful lymphocyte patterns. Likewise, the stroma module showed stable behavior across both public and private datasets, with explainability analyses confirming that relevant stromal boundaries were reliably identified. To translate these patch-level insights into clinically interpretable ROI-level scores, we introduced a percentile-weighted aggregation approach tailored to the skewed distribution of TILs infiltration, where only a minority of patches carry significant biological signal. Across two independent cohorts, the full pipeline achieved strong correlation, substantial agreement, and minimal numerical bias relative to expert pathologists. The modular design and transparent intermediate outputs offer practical advantages, supporting pathologist review, quality control, and integration into existing diagnostic workflows.

Future work will focus on extending the pipeline toward a fully automated whole-slide workflow, allowing the system to identify tumor-rich regions and select diagnostically relevant stromal areas without manual intervention. Strengthening generalizability will also require larger and more diverse multi-center datasets that capture real variations in staining, scanners, and laboratory workflows. In parallel, expanding the reference panel to include more pathologists would help refine the clinical ground truth and reduce inter-observer uncertainty. Beyond technical improvements, future work will investigate how automated TILs scores relate to treatment response and patient outcomes, paving the way for integration into prognostic or therapy-guidance models.

CRediT authorship contribution statement

Shrief Abdelazeez: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Faisal Ahmed:** Writing – review & editing, Visualization, Validation, Investigation. **Laia Adalid:** Writing – review & editing, Visualization, Validation, Resources, Data curation. **Krzysztof Siemion:** Writing – review & editing, Visualization, Validation, Resources, Data curation. **Carlos Lopez:** Writing – review & editing, Visualization, Resources, Investigation. **Marylene Lejeune:** Writing – review & editing, Visualization, Resources. **Hatem Rashwan:** Writing – review & editing, Visualization, Validation, Conceptualization. **Anna Korzynska:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Investigation, Funding acquisition.

Ethics statement

This study was conducted in compliance with relevant ethical guidelines and regulations. The use of histopathological images and associated data was approved by the Drug Research Ethics Committee (CEIm) at IISPV, under approval reference number 128/2022 (1st July 2022) & 218/2022 (20th December 2022).

Informed consent was obtained from all participants (or their legal representatives) prior to the use of their histopathological images for research purposes. All data were anonymized to ensure the privacy and confidentiality of the participants, in accordance with applicable data protection laws and institutional guidelines.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Anna Korzynska reports financial support was provided by European Research Executive Agency. Anna Korzynska reports administrative support was provided by Nalecz Institute of Biocybernetics and Biomedical Engineering Polish Academy of Sciences. Carlos Lopez reports financial support was provided by SCARLET, a project funded by Proyectos Estratégicos Orientados a la Transición Ecológica y a la Transición Digital, from the 2021 call of the Ministerio de Ciencia e Innovación, with grant number TED2021-130081B-C22 and funding from NextGenerationEU. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was conducted as part of the *BOSOMSHIELD* European Project, funded by the Horizon Europe Research and Innovation Programme under Grant Agreement No. 101073222. We gratefully acknowledge this support, which made the research and development activities presented in this study possible. We would also like to express our sincere gratitude to Dr. Łukasz Roszkowiak for his generous support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2026.109374.

References

[1] R. Salgado, et al., International TILs working group 2014: recommendations for assessing TILs in breast cancer, Ann. Oncol. 26 (2) (2015) 259–271.
[2] J. Saltz, et al., Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images, Cell Rep. 23 (1) (2018) 181–193.
[3] Q. Teng, Z. Liu, Y. Song, et al., A survey on the interpretability of deep learning in medical diagnosis, Multimedia Syst. 28 (6) (2022) 2335–2355, http://dx.doi.org/10.1007/s00530-022-00960-4.
[4] S. Li, Q. Zhu, C. Tian, W. Shao, D. Zhang, Interpretable dynamic brain network analysis with functional and structural priors, IEEE Trans. Med. Imaging 44 (12) (2025) 4878–4889, http://dx.doi.org/10.1109/TMI.2025.3584231.
[5] G. Campanella, et al., Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Med. 25 (2019) 1301–1309.
[6] R. Salgado, C. Denkert, S. Demaria, et al., The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international TILs working group 2014, Ann. Oncol. 26 (2) (2015) 259–271.
[7] N. Naik, A. Madabhushi, et al., A deep learning framework for identifying immunological biomarkers in breast cancer, Front. Oncol. 10 (2020) 132.
[8] N.A. Koohbanani, M. Jahanifar, et al., Nuclear instance segmentation using a novel adaptive attention mechanism in digital pathology images, Med. Image Anal. 70 (2021) 102004.
[9] J. Kwon, et al., AI-based digital assessment of tumor-infiltrating lymphocytes as a biomarker in breast cancer, J. Pathol.: Clin. Res. 8 (3) (2022) 251–263.
[10] E. Marostica, et al., Development of a deep learning model to quantify stromal tumor-infiltrating lymphocytes in breast cancer, Mod. Pathol. 34 (3) (2021) 446–456.
[11] B. Settles, Active Learning Literature Survey, Tech. Rep. 1648, University of Wisconsin–Madison, 2009, URL http://burrsettles.com/pub/settles.activelearning.pdf.
[12] A. Smailagic, P. Costa, G. Mankovich, D. Walawalkar, A. Kogan, H. Xu, H. Noh, R. Klatzky, D. Siewiorek, J. Tan, Medal: Accurate and robust deep active learning for medical image analysis, IEEE J. Biomed. Health Inform. 24 (9) (2020) 2595–2605, http://dx.doi.org/10.1109/JBHI.2020.2972041.
[13] M.Y. Lu, et al., Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (2021) 555–570.
[14] P. Bandi, M. Balkenhol, M.C. van Dijk, et al., A context-aware deep learning approach to TILs quantification in breast cancer from H&E-stained whole-slide images, Med. Image Anal. 74 (2022) 102248.
[15] A. Heindl, et al., Mapping spatial heterogeneity and tumor–immune interactions in breast cancer by computational pathology, J. Pathol. (2018).
[16] H. Makhlouf, et al., Deep learning framework for automated quantification of tumor-infiltrating lymphocytes in luminal breast cancer, Sci. Rep. (2021).
[17] O. Ronneberger, et al., U-net: Convolutional networks for biomedical image segmentation, MICCAI (2015).
[18] S. Graham, et al., Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Med. Image Anal. (2019).
[19] J. Choi, et al., Deeptils: Deep learning-based stromal TILs quantification for breast cancer prognostic assessment, Cancers (2022).
[20] S. Ren, et al., Faster R-CNN: Towards real-time object detection with region proposal networks, Neural Inf. Process. Syst. (2015).
[21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018, arXiv:1802.02611.
[22] M. Yosofvand, et al., Two-stage deep learning framework for automated detection and scoring of stromal TILs in breast cancer histopathology, Diagnostics (2022).
[23] K. He, et al., Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. (2017).
[24] S. Kim, K.R. Hasan, Y. Ando, S. Ko, D. Lee, N.J. Park, J. Cho, Improving tumor-infiltrating lymphocytes score prediction in breast cancer with self-supervised learning, Life 14 (1) (2024) http://dx.doi.org/10.3390/life14010090.
[25] T. Chen, et al., A simple framework for contrastive learning of visual representations, ICML (2020).
[26] M. Caron, et al., Unsupervised learning of visual features by contrasting cluster assignments, NeurIPS (2020).
[27] Y. Liu, et al., Mutils: Concept-bottleneck neural network for interpretable tils quantification, Nat. Commun. (2024).
[28] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, 2015, arXiv:1505.04597.
[29] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-net: Learning where to look for the pancreas, 2018, arXiv:1804.03999.
[30] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, 2021, arXiv:2102.04306.
[31] Y. Shen, Y. Luo, D. Shen, J. Ke, RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, 2022, pp. 212–221, http://dx.doi.org/10.1007/978-3-031-16434-7_21.
[32] K. Vinogradova, A. Dibrov, G. Myers, Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract), Proc. AAAI Conf. Artif. Intell. 34 (10) (2020) 13943–13944, http://dx.doi.org/10.1609/aaai.v34i10.7244.
[33] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, 2019, arXiv:1709.01507.
[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, arXiv:1512.03385.
[35] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, A. Smola, ResNeSt: Split-attention networks, 2020, arXiv:2004.08955.
[36] S. Mehta, M. Rastegari, MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer, 2022, arXiv:2110.02178.
[37] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, Y. Yuan, EfficientViT: Memory efficient vision transformer with cascaded group attention, 2023, arXiv:2305.07027.
[38] P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychol. Bull. 86 (2) (1979) 420–428.
[39] L.I.-K. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1) (1989) 255–268.
[40] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.
[41] N. Abraham, N.M. Khan, A novel focal tversky loss function with improved attention U-net for lesion segmentation, 2018, arXiv:1810.07842.