Biomedical Signal Processing and Control 120 (2026) 110003

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

AI-based framework pipeline for stromal tumor infiltrating lymphocytes quantification in H&E stained breast cancer histopathological images

Shrief Abdelazeez [a,b,*], Alessio Fiorin [d,e,f], Vincenzo Della Mea [c], Laia Adalid [d,e], Krzysztof Siemion [a,g,h], Carlos Lopez [d,e,f], Marylene Lejeune [d,e,f], Anna Korzynska [a]

[a] Laboratory of Analysis and Processing of Microscopic Image, Nalecz 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 Mathematics, Computer Science and Physics, University of Udine, Udine, Italy
[d] Oncological Pathology and Bioinformatics Research Group, Institut d'Investigació Sanitària Pere Virgili, Tortosa, Spain
[e] Department of Pathology, Hospital de Tortosa Verge de la Cinta, Institut Català de la Salut, Tortosa, Spain
[f] Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, 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:
Breast cancer
Tumor-infiltrating lymphocytes
Artificial intelligence
Lymphocytes segmentation
Lymphocytes scoring
TILs quantification
Digital pathology images

ABSTRACT

The tumor microenvironment (TME) plays a pivotal role in immuno-oncology, influencing cancer progression and treatment outcomes. Among its components, stromal tumor-infiltrating lymphocytes (sTILs) represent important prognostic and predictive biomarkers in breast cancer. However, visual assessment of sTILs is complex, time-consuming, and subject to inter-observer variability. To address these challenges, we propose an AI-based framework for sTILs quantification in hematoxylin and eosin (H&E)-stained breast cancer histopathological images across different molecular subtypes, in accordance with the recommendations of the International Immuno-Oncology Biomarker Working Group (IIOBWG). The framework integrates two parallel and explainable segmentation models trained on multi-center data: an attention-based model for TILs detection and a stroma segmentation model designed to identify both tumoral and infiltrated stromal regions as a unified clinically relevant compartment. The resulting segmentation outputs are combined through a scanner resolution-aware quantification algorithm, enabling robust sTILs scoring across data acquired from different scanners and institutions. The proposed framework was externally validated on two independent private datasets comprising 60 and 112 whole-slide images with expert-annotated Region Of Interests (ROIs). Strong agreement with expert pathologist scoring was achieved (Pearson correlation up to 0.73; Concordance Correlation Coefficient (CCC) up to 0.72), and the framework consistently outperformed state-of-the-art approaches when evaluated on the same data. These results demonstrate that the proposed framework provides a reliable, efficient, and scalable solution for standardized sTILs quantification, supporting its potential utility in both clinical practice and translational research.

1. Introduction

Breast cancer (BC) remains a leading cause of cancer-related morbidity and mortality worldwide, accounting for a significant proportion of neoplastic diagnoses and deaths annually [1–4]. This disease is characterized by its complexity and heterogeneity, presenting unique challenges in diagnosis, treatment, and prognosis. Beyond the intrinsic properties of tumor cells, the tumor microenvironment (TME) plays a pivotal role in shaping disease progression, metastasis, and response to therapy [5,6]. The TME consists of a diverse range of non-cancerous components, including immune cells, stromal cells, blood vessels, and the extracellular matrix. Together, these elements create a dynamic ecosystem that interacts with tumor cells, influencing their behavior and modulating the effectiveness of therapeutic interventions [7,8].

Among the various components of the TME, tumor-infiltrating lymphocytes (TILs) have garnered considerable attention due to their critical role in cancer biology. TILs, primarily composed of T cells, represent

* Corresponding author at: Laboratory of Analysis and Processing of Microscopic Image, Nalecz 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.bspc.2026.110003
Received 13 February 2025; Received in revised form 10 February 2026; Accepted 1 March 2026
Available online 6 March 2026
1746-8094/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

the host immune response to malignancies and infiltrate both tumor tissue and the surrounding stroma. Through complex interactions with cancer cells and other components of the TME, TILs significantly influence tumor behavior and disease progression. Their presence and activity are particularly relevant in BC, where they have been shown to affect tumor growth, metastasis, and improve patient outcomes [9].

In BC, TIL infiltration is most strongly associated with two molecular subtypes: triple-negative breast cancer (TNBC) and human epidermal growth factor receptor 2 (HER2)-positive BC. In TNBC, which lacks expression of estrogen, progesterone, and HER2 receptors, high levels of stromal TILs correlate with improved responses to chemotherapy and enhanced overall survival [10]. Similarly, in HER2-positive breast cancer, stromal TILs serve as favorable prognostic biomarkers, particularly in patients receiving anti-HER2 therapies [11]. These findings underscore the clinical relevance of sTILs as biomarkers capable of informing therapeutic decision-making.

Quantification of sTILs provides valuable prognostic and predictive insights, enabling clinicians to stratify patients according to their immune response and tailor treatment strategies accordingly. A high density of TILs within the stromal compartment is generally associated with a more robust anti-tumor immune response, improved tumor control, and a greater likelihood of therapeutic success [12]. Consequently, sTILs assessment has become an integral component of contemporary oncology practice, guiding immunotherapy and chemotherapy decisions. Moreover, sTILs quantification has emerged as an important biomarker in clinical trials, facilitating patient selection and monitoring the efficacy of immune-based therapies [13].

Despite its clinical importance, visual quantification of sTILs by pathologists presents several challenges. This manual process requires examination of H&E-stained whole-slide images (WSIs) within regions strictly defined by international guidelines [11] to assess the density and spatial distribution of lymphocytes. Variability in staining quality, tumor morphology, and tumor biology, which contributes to heterogeneity in the immune response surrounding the tumor, as well as inter- and intra-observer subjectivity among pathologists, can lead to inconsistent assessments by human eye and raise concerns regarding reliability [11]. Furthermore, the labor-intensive nature of this task imposes a substantial workload on pathologists, potentially delaying clinical decision-making and increasing the risk of diagnostic errors [14].

These challenges underscore the need for standardized, objective, and reproducible methods for evaluating sTILs. Recent advances in digital pathology and artificial intelligence (AI) provide promising solutions by enabling automated and quantitative assessment of sTILs. By overcoming the limitations of manual evaluation, AI-based approaches have the potential to enhance the reliability and efficiency of sTILs quantification, paving the way for more precise and personalized breast cancer management.

The main **contributions** of this study can be summarized as follows:

- We propose a comprehensive AI-based framework for automated sTILs quantification in H&E-stained breast cancer whole-slide images, designed to be applicable across all major molecular subtypes.
- The framework integrates two specialized segmentation pathways: an attention-based model for accurate TILs detection and a dedicated stroma segmentation model capable of jointly identifying tumoral and infiltrated stromal regions in accordance with IIOBWG guidelines.
- We introduce a scanner resolution-aware sTILs quantification algorithm that incorporates scanner-specific metadata to provide consistent and comparable sTILs estimates across images acquired with different scanning resolutions and hardware.
- To enhance robustness and scanner invariance, extensive data augmentation strategies — including geometric, color, stain-like, elastic deformation, and resolution-scaling augmentations — are applied during training to account for inter-scanner, inter-institutional, and staining variability.
- The segmentation models are trained using multi-institutional public datasets and are externally validated on independently collected in-house datasets, demonstrating strong generalization beyond the training domain.
- The proposed framework supports flexible deployment at multiple spatial scales, enabling sTILs quantification on ROIs chosen by pathologists within the same whole-slide image.
- Finally, model interpretability is enhanced through the integration of SegGrad-CAM [15], providing visual explanations for both TILs and stroma segmentation outputs and supporting transparent clinical interpretation.

The remainder of this paper is organized as follows. Section 2 presents a review of related work on sTILs quantification. Section 3 describes the materials and methods, including the datasets and the proposed AI-based framework pipeline. The experimental results are presented and discussed in Section 4. Finally, Section 5 summarizes the main limitations of the study, while Section 6 outlines the conclusion and directions for future research.

2. Related work

sTILs have been recognized as prognostic markers, particularly in aggressive subtypes of BC. The International Immuno-Oncology Biomarker Working Group (IIOBWG) on BC established standardized guidelines for evaluating sTILs [11], emphasizing their prognostic value in clinical practice. According to these guidelines, patients are categorized into three groups based on their sTILs scores: low immune infiltrate response, with scores ranging from 0% to less than 10%; moderate immune infiltrate response, with scores ranging from 10% to less than 40%; and high immune infiltrate response, with scores ranging from more 40%.

Several studies have demonstrated that a higher percentage of sTILs is associated with better outcomes in TNBC and HER2-positive cases. For example, Denkert et al. [16] showed that sTILs levels correlate with pathological complete response (pCR) rates following neoadjuvant chemotherapy.

Immunotherapies, such as immune checkpoint inhibitors, have further highlighted the role of sTILs as predictive biomarkers [17]. Patients with high sTILs levels exhibit stronger responses to immune-based therapies [18]. Thus, sTILs quantification is pivotal for stratifying patients and making therapeutic decisions.

Manual sTILs assessment remains the gold standard, but it poses significant drawbacks, including inter- and intra-observer discrepancies, as well as the time-consuming nature of the process. Additionally, variability in slide quality and tumor heterogeneity contribute to inconsistent evaluations [11]. Moreover, visual quantification is time-consuming because it requires pathologists to analyze numerous slides with high precision [19]. These challenges necessitate the development of automated methods to standardize and expedite the assessment of sTILs.

Advancements in digital pathology and artificial intelligence (AI) have paved the way for the automated quantification of sTILs. Bankhead et al. [20] present QuPath, an open-source, extensible software platform for digital pathology and whole-slide image analysis, offering tools for tumor identification, biomarker evaluation, and customizable workflows that can also support TILs assessment. F. Klauschen et al. [19] explored the idea of using AI-based systems to analyze whole slide images (WSIs) to detect and quantify lymphocytic infiltration in tumor and stromal regions.

Automated systems to be clinically viable must align with established guidelines. IIOBWG has emphasized the need for validation of AI tools against visual assessments by expert pathologists [11]. In this context, Thagaard, Broeckx et al. [21] critically reviewed current ML-based approaches for TILs assessment in BC, highlighting both their potential and the pitfalls — such as false-positive detections, technical

slide issues, and validation challenges — that must be addressed before reliable computational TILs reporting can be incorporated into clinical practice.

Research findings of A. Heindl et al. [22] have reported that AI-assisted quantification can enhance the consistency and reproducibility of sTILs scoring, facilitating broader adoption in clinical practice. A study by J. Saltz et al. [23] employed convolutional neural networks (CNNs) to analyze TILs in BC, showing that AI models can achieve high concordance with pathologists' evaluations. This automated system also enables the analysis of larger regions of interest, thereby reducing the risk of missing subtle patterns of immune infiltration.

Makhlouf et al. [24] developed a deep learning framework for quantifying both tumoral TILs (tTILs) & sTILs in luminal BC. The framework utilized the UNet [25] model for tumor and stroma area region segmentation and the HoVer-Net [26] model for immune cell segmentation. The models were trained on their own dataset and evaluated on 300 cases. However, this study was limited to luminal B and lacked validation across other molecular subtypes, restricting its broader clinical applicability.

Additionally, Choi et al. [27] developed a deep learning framework for evaluating sTILs in BC and enhancing therapeutic response prediction. The framework comprised two models: a Faster R-CNN [28] with a ResNet-34 backbone for detecting tumor and lymphocyte cells, and DeepLabv3 [29] with an EfficientNet-B3 backbone for tissue segmentation. The models were trained and evaluated on a dataset of 402 slides. The evaluation have been done by comparing the framework's performance against the average scores of three pathologists. However, the framework's generalizability remains uncertain due to the absence of external validation on independent datasets or diverse clinical settings, which is critical for real-world applicability.

While Yosofvand et al. [30] proposed a two-stage deep learning framework for automated detection and scoring of sTILs in BC histopathology slides. The first stage employed a U-Net [25] model for stroma segmentation, while the second stage utilized Mask R-CNN [31] with a ResNet-101 backbone for TILs detection. The framework was developed and evaluated using a dataset of 63 slides, which were utilized for both training and evaluation. However, the reported high accuracy may reflect overfitting to the specific dataset rather than true generalizability, since the same slides were used for both training and evaluation without independent external validation.

Moreover, Kim et al. [32] developed a self-supervised learning framework for TILs scoring, training SimCLR [33] and SwAV [34] models on large-scale pathological datasets from different organs [26, 35–45] to pre-train encoder weights, instead of relying on ImageNet-pretrained weights. These pre-trained encoders were transferred to downstream models for tissue segmentation using DeepLabv3 [29] and lymphocyte detection using U-Net [25], with the encoder weights frozen during training. The framework was trained and evaluated using the publicly available TIGER challenge datasets, and TILs scores were calculated based on feature maps generated from tissue segmentation and lymphocyte detection outputs. However, this approach has two key limitations: (1) the evaluation using uniformly resized WSIs (256×256 pixels) leads to significant information loss, potentially making the reported scores misleading compared to the original WSI-level ground truth; and (2) the exclusive reliance on public benchmark data without external validation raises concerns about clinical applicability across diverse real-world datasets.

Recently, Liu et al. [46] proposed MuTILs, a concept bottleneck neural network that jointly performs lymphocyte detection and tumor-associated stroma segmentation in line with IIOBWG guidelines. The model was developed and evaluated on the PanopTILs dataset, a large region- and cell-level annotation resource that includes over 800,000 nuclei from BC patients. MuTILs demonstrated strong performance for both tasks and introduced interpretability through concept bottlenecks, making it clinically relevant for TILs assessment. However, the study considered only HER2-positive and TNBC subtypes, without covering the full range of molecular subtypes, and its validation was limited to public datasets, which may restrict generalizability.

To address the limitations identified in previous studies — including restricted molecular subtype coverage, limited external validation, disregarding to the characteristics of various scanners, and the lack of transparent interpretability — we propose a robust AI-based framework for comprehensive sTILs quantification in breast cancer. Our approach combines precise TILs and stoma segmentation, scanner resolution-aware quantification, extensive augmentation for scanner and stain invariance, and systematic external validation on independent multi-center datasets. By integrating explainability through SegGrad-CAM and enabling multi-scale analysis across ROIs and whole-slide images, the proposed framework aims to provide a standardized, reproducible, and clinically relevant solution that overcomes key shortcomings of existing automated sTILs assessment methods.

3. Materials and methods

3.1. Data description

In this study, we employed two datasets: a publicly available dataset for model training and a private dataset for external validation. A detailed description of both datasets is provided in the following subsections.

3.1.1. Public data

As the initial step in model development, we employed the publicly available PanopTILs dataset [47] for training.[1] PanopTILs represents a unique large-scale resource that integrates and extends two existing datasets, BCSS [48] and NuCLS [49], to support in-depth analysis of the TME in WSIs of H&E-stained BC tissue. The dataset provides comprehensive annotations, including histologic region boundaries, nuclear locations, nuclear classification labels, and segmentation boundaries, thereby enabling panoptic segmentation of tissue and nuclei. Such an integrated representation is particularly valuable for the study of TILs, as it aligns with established clinical guidelines and allows the development of deep learning models that produce biologically meaningful and clinically relevant predictions.

The PanopTILs dataset encompasses the Cancer Genome Atlas (TCGA) invasive BC cases collected from 151 patients across **multiple hospitals**, covering a total of 1709 annotated ROIs and 814,886 nuclei. Each ROI is provided at a size of 1024×1024 pixels with 0.25 μm per pixel, corresponding to a 40× magnification in standard digital pathology practice. For each ROI, a paired RGB image and a three-channel segmentation mask are provided. The first channel encodes histologic region segmentation, including categories such as cancerous epithelium, tumoral stroma, TILs dense regions, and other tissue structures. The second channel contains nuclear-level annotations, covering a broad range of cell types, including cancer, stromal, lymphocyte, and plasma cell nuclei. The third channel provides a binary map highlighting nuclear boundaries, which supports the accurate delineation of individual nuclei. This comprehensive annotation structure ensures that all relevant elements within the tumor microenvironment are exhaustively represented.

PanopTILs combines both region-level and nucleus-level annotations from the same areas of tissue, which makes it a consistent and reliable dataset for studying the tumor microenvironment. Its large size and detailed annotation make PanopTILs especially valuable for training deep learning models. In this study, models were trained for two main tasks: detecting and segmenting TILs, and segmenting stromal regions.

[1] https://sites.google.com/view/panoptils/home

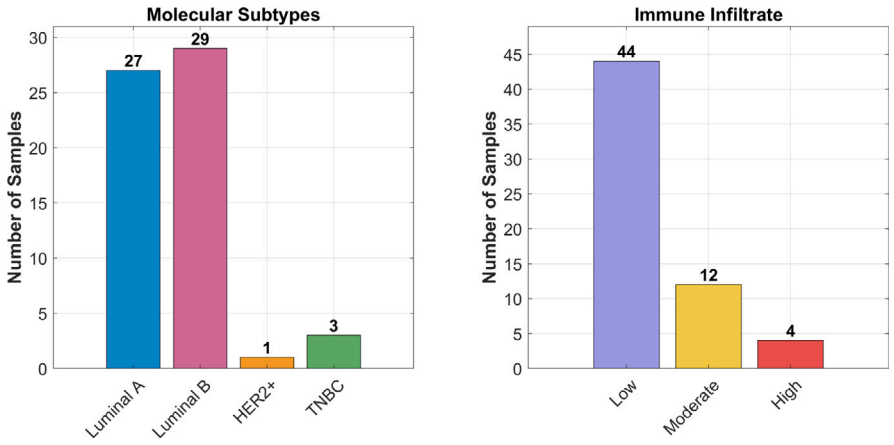


Fig. 1. Distribution of patient samples in the first private validation subset according to molecular subtype and sTILs score category.

3.1.2. Private data

The private dataset was prepared collaboratively by the Molecular Biology and Research Section and the Pathology Department of Hospital de Tortosa Verge de la Cinta, IISPV, URV (Tortosa, Spain), following the methodology described in [50]. These data were used exclusively for external validation of the proposed framework and were not involved in model training.

The private dataset comprises three independent subsets, each designed to evaluate a specific aspect of the framework, including generalizability across molecular subtypes and robustness to scanner variability.

Subset 1: The first subset consists of 60 H&E-stained whole-slide images (WSIs) from BC patient samples covering different molecular subtypes, namely Luminal A, Luminal B, HER2-positive, and triple-negative breast cancer (TNBC). These data were used for model evaluation and external validation of the complete framework. All slides were digitized using a 3DHISTECH Panoramic scanner at a resolution of 0.2433 μm/px to ensure high-quality imaging for quantitative analysis.

For each slide, a team of expert pathologists annotated the tumor-dominant region of interest (ROI). Within each ROI, lymphocytes and plasma cells were annotated separately; both cell types were considered as TILs for the purpose of sTILs quantification. In addition, tumoral and infiltrated stroma regions, tumor areas, and other tissue structures such as glands, fibroblasts, and necrotic regions were carefully delineated. All annotations were performed using QuPath and stored in GeoJSON format for subsequent preprocessing.

The sTILs score for each ROI was independently and blindly assessed by two expert pathologists, and the average of their evaluations was used as the reference standard for validation. Based on these reference scores, ROIs were categorized into low, moderate, and high immune infiltrate groups according to the Salgado criteria [11] described in Related Work section. The distribution of patient samples across molecular subtypes and sTILs score categories for this subset is shown in Fig. 1.

Subset 2: The second subset includes 112 H&E-stained WSIs from BC patient samples representing the same molecular subtypes (Luminal A, Luminal B, HER2-positive, and TNBC). All slides were scanned using a 3DHISTECH Panoramic scanner at a resolution of 0.2433 μm/px.

In this subset, ROIs were annotated using SlideMaster, with annotations stored in XML format. The sTILs score for each ROI was visually assessed by a single expert pathologist and considered as the ground truth for external framework evaluation. The distribution of patient samples by molecular subtype and sTILs score category is presented in Fig. 2.

Subset 3: The third subset consists of 50 H&E-stained WSIs from breast cancer patient samples and was specifically designed to assess the scanner invariance of the proposed framework. Slides in this subset were digitized at two different institutions using whole-slide imaging systems from distinct vendors: a 3DHISTECH Panoramic 250 Flash III scanner (3DHISTECH Ltd., Hungary) with a spatial resolution of 0.2433 μm/px at Hospital de Tortosa Verge de la Cinta, IISPV, URV (Tortosa, Spain), and a Hamamatsu NanoZoomer 2.0 HT scanner (model C9600; Hamamatsu Photonics, Japan) with a spatial resolution of 0.227 μm/px at Université Bourgogne, Dijon, France.

For each case, the same tumor-dominant ROI was identified and consistently annotated across scanners, with annotations stored in GeoJSON format. The corresponding sTILs score was visually assessed by an expert Pathologist and used as the reference standard for evaluation. This subset enables a systematic assessment of the robustness of the proposed framework to variations in scanner hardware, spatial resolution, and institutional acquisition settings, which is a critical requirement for real-world clinical deployment.

3.2. Data preprocessing

3.2.1. Public data

A consistent preprocessing pipeline was applied to the public dataset for both tasks, namely TILs segmentation and stromal binary segmentation. Given that the provided ROI annotations consist of three-channel masks, task-specific binary masks were generated by selecting and merging relevant annotation classes.

For the **TILs segmentation task**, the second channel of the ROI masks, which contains nuclear-level annotations, was processed by extracting lymphocyte and plasma cell classes only. These two cell types were subsequently merged and converted into a single binary mask representing TILs.

For the **stroma segmentation task**, the first channel of the ROI masks, which encodes region-level tissue annotations, was used. Guided by an expert pathologist, the tumoral stroma and TILs-dense region labels were selected and merged into a single stromal category. This merging strategy was adopted to represent the stromal compartment as a whole, since both tumoral and inflamed stroma constitute clinically relevant microenvironmental regions that are difficult to separate reliably in an automated segmentation setting. Binary masks were then generated accordingly for all ROIs.

To address cases of touching or overlapping nuclei in the TILs masks, a watershed-based separation strategy was applied. Specifically, the watershed algorithm was implemented on the distance transform of the inverted binary masks, with minima imposition used to suppress shallow local minima. This procedure effectively separated adjacent or overlapping nuclei, resulting in improved delineation of individual cells. Following watershed refinement, the total number of detected TILs increased from 166,119 to 191,894, corresponding to an additional 25,775 TILs that were previously merged in the raw masks. A quantitative comparison of TILs counts before and after watershed segmentation

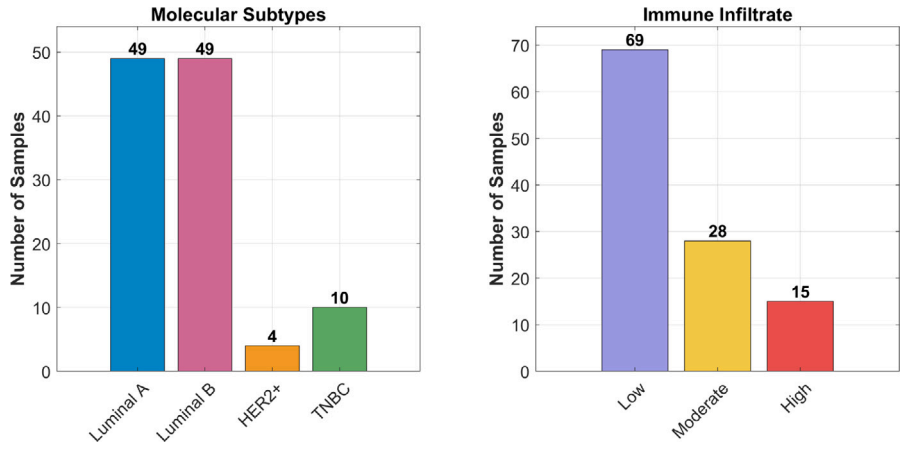


Fig. 2. Distribution of patient samples in the second private validation subset according to molecular subtype and sTILs score category.

Table 1
Number of detected TILs before and after applying watershed-based separation at the ROI level.

	Number of TILs
Before watershed segmentation	166,119
After watershed segmentation	191,894
Difference	25,775

at the ROI level is summarized in Table 1, while a schematic illustration of the watershed procedure is provided in the supplementary material (Figure S1).

For both segmentation tasks, the dataset was split into 80% for training and 20% held out for testing, with the split performed at the hospital level to prevent data leakage. Specifically, all ROIs originating from the same hospital were assigned exclusively to either the training or testing set, ensuring that ROIs from the same patient cohort or acquisition center did not appear in both sets. Within the training set, a five-fold cross-validation strategy was employed following the same hospital-level separation criteria, such that ROIs from a given hospital were contained entirely within a single fold.

During training, each ROI was tiled on-the-fly into patches of size 256 × 256 pixels with an overlap of 128 × 128 pixels, corresponding to 50% overlap between adjacent tiles. Tiles corresponding to empty TIL masks in the TILs segmentation task or empty stromal masks in the stromal segmentation task were discarded to ensure that the models were trained only on informative regions.

To enhance model generalization and mitigate overfitting, a range of data augmentation techniques was applied dynamically during training. These augmentations included: (i) geometric transformations such as random horizontal and vertical flips and rotations of 90°, 180°, or 270°; (ii) brightness adjustments through random intensity scaling to simulate variations in staining and illumination; (iii) color perturbations in the HSV color space, including random hue shifts, saturation scaling, and value scaling to account for inter-slide and inter-scanner variability; (iv) stain-like augmentation via independent per-channel affine transformations of the RGB channels to model stain and color variations across laboratories [51]; (v) elastic deformation using smooth random displacement fields applied consistently to images and masks to mimic realistic tissue distortions [52]; and (vi) resolution scaling (magnification augmentation), where images were down-sampled to simulate lower scanning resolutions and subsequently resized back to the original input size.

These augmentation strategies ensured exposure of the models to diverse visual appearances during training, thereby improving robustness and performance across different ROIs, staining conditions, and scanner settings. Representative examples of the applied augmentations are provided in the supplementary material (Figures S2 & S3).

3.2.2. Private data

To prepare the private datasets for external validation, a set of preprocessing steps was applied, tailored to the annotation format and intended use of each subset.

For the first private subset, regions of interest (ROIs) were extracted from the whole-slide images together with their corresponding binary masks for TILs and stroma regions. These masks were generated from the expert annotations provided in GeoJSON format. The extracted ROIs and masks were subsequently prepared for direct use in the evaluation of both the segmentation models and the complete framework.

For the second private subset, preprocessing was limited to the extraction of ROI images from the whole-slide images based on the XML annotation files. Since this subset was used exclusively for external validation, no additional mask generation or modification was performed beyond preparing the extracted ROIs for inference and evaluation.

For the third private subset, preprocessing followed the same procedure as for the second subset. ROIs were extracted from whole-slide images acquired using each scanner, based on the corresponding GeoJSON annotation files. The extracted ROIs were then prepared for the validation phase, enabling assessment of the framework's robustness and scanner invariance across different scanning devices and resolutions.

3.3. Framework architecture

The proposed framework adopts a dual-pathway approach for analyzing ROIs in H&E-stained BC WSI acquired from digital scanners. In the first pathway, a dedicated segmentation model detects and delineates TILs within the ROI, irrespective of their spatial distribution. In parallel, the second path focuses on segmenting stromal tissue regions within the same ROI. The outputs from both pathways are then combined with scanner resolution metadata through a tailored quantification algorithm, which estimates the overall sTILs score. This design aims to provide a systematic and adaptable method for sTILs assessment, while acknowledging that further refinements may enhance its robustness across diverse datasets. Fig. 3 illustrates the complete framework pipeline. The subsequent sections provide a detailed description of each component within the framework.

3.3.1. TILs segmentation path

For the task of TILs segmentation and detection, three state-of-the-art deep learning architectures were investigated: UNet [25], Attention-UNet [53], and TransUNet [54]. UNet was selected as a baseline model due to its widespread adoption in medical image segmentation and its proven effectiveness in producing accurate pixel-level predictions through an encoder–decoder structure with skip connections. Attention

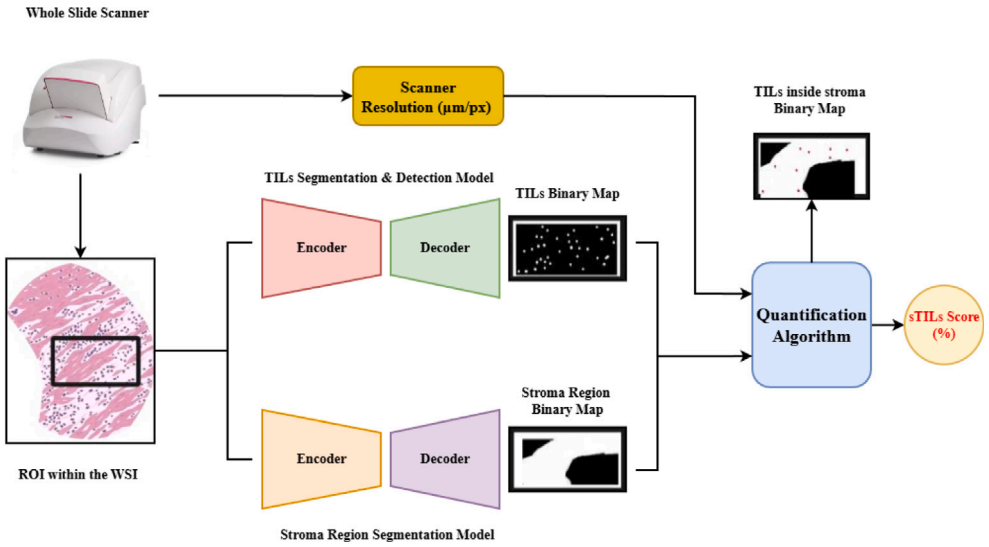


Fig. 3. Overall framework pipeline architecture.

Table 2
Training hyperparameter configurations used during five-fold cross-validation for TILs segmentation.

Config.	Optimizer	Learning rate	Batch size	Momentum	Nesterov	ρ (RMSProp)
1	Adam	5×10^{-4}	8	–	–	–
2	SGD	1×10^{-4}	16	0.9	True	–
3	RMSProp	1×10^{-4}	8	–	–	0.9
4	Adagrad	3×10^{-4}	16	–	–	–
5	Adamax	2×10^{-4}	8	–	–	–

U-Net was included as an extension of this baseline, incorporating attention mechanisms that enable the model to emphasize informative regions while suppressing irrelevant background. This function is particularly beneficial in histopathological images where TILs often appear within complex tissue environments. TransUNet was evaluated as a more recent architecture that integrates convolutional layers with transformer-based encoders, allowing the model to capture both local texture information and long-range contextual dependencies. This capability is especially relevant in breast cancer pathology, where the spatial distribution of TILs within the stromal compartment can influence clinical interpretation.

By evaluating these three representative architectures — spanning conventional convolutional networks, attention-enhanced models, and transformer-based designs — the study aimed to identify the most effective approach for TILs segmentation and detection. The best-performing model was subsequently selected for integration into the proposed framework.

To ensure a fair comparison between models and to reduce bias associated with a single train–validation split, a five-fold cross-validation strategy was adopted. This approach allowed each subset of the dataset to contribute to both training and validation across different folds, thereby improving the robustness and generalizability of the evaluation.

For each fold, the candidate models were trained on the preprocessed ROIs from the public dataset using distinct, randomly assigned hyperparameter configurations. Employing diverse optimization strategies aimed to enhance generalization and reduce sensitivity to specific hyperparameter choices. All models were initialized with ImageNet-pretrained weights, and both encoder and decoder parameters were optimized during training. The number of training epochs was set to 100, and early stopping was applied with a patience of 10 epochs based on validation loss. For each fold, the model weights corresponding to the minimum validation loss were retained. The hyperparameter configurations used during five-fold cross-validation are summarized in Table 2.

To address the pronounced class imbalance between background and foreground pixels in the TILs segmentation task, a hybrid loss function defined as the sum of Tversky Focal Loss (TFL) [55] and Jaccard Loss (JL) was consistently employed during training. TFL introduces three parameters: α and β, which control the relative penalties for false negatives and false positives (with $\alpha+\beta = 1$), and γ, which increases the emphasis on harder-to-classify examples. In this study, γ was fixed to $4/3$ as suggested in [55], while the optimal value of α was empirically determined to be 0.8, as illustrated in Fig. 4.

Model performance was evaluated on both the validation folds and the held-out test set of the public dataset using complementary segmentation and detection metrics. Segmentation quality was assessed using pixel-wise accuracy, the Dice coefficient, and the Tversky index. Detection performance was evaluated by counting true positive, false positive, and false negative TILs, from which precision, recall, and F1-score were derived. This evaluation strategy ensured a comprehensive assessment of both the segmentation quality and the ability of the models to accurately detect individual TILs.

Finally, to gain insight into the discriminative features underlying model predictions, SegGrad-CAM [15] was applied to the best-performing TILs segmentation model to visualize the regions that contributed most strongly to the detection decisions.

3.3.2. Stroma segmentation path

For the task of stroma region segmentation, three state-of-the-art semantic segmentation architectures were investigated: U-Net [25], LinkNet [56], and PSPNet [57]. Each architecture was evaluated using different backbone encoders, namely VGG16, InceptionResNetV2, and DenseNet169, to assess their ability to capture both low-level spatial details and high-level contextual features that are critical for accurate stromal segmentation.

U-Net has become a widely adopted benchmark in medical image segmentation due to its symmetric encoder–decoder architecture and skip connections, which facilitate effective propagation of fine-grained spatial information from early to deeper layers. This design

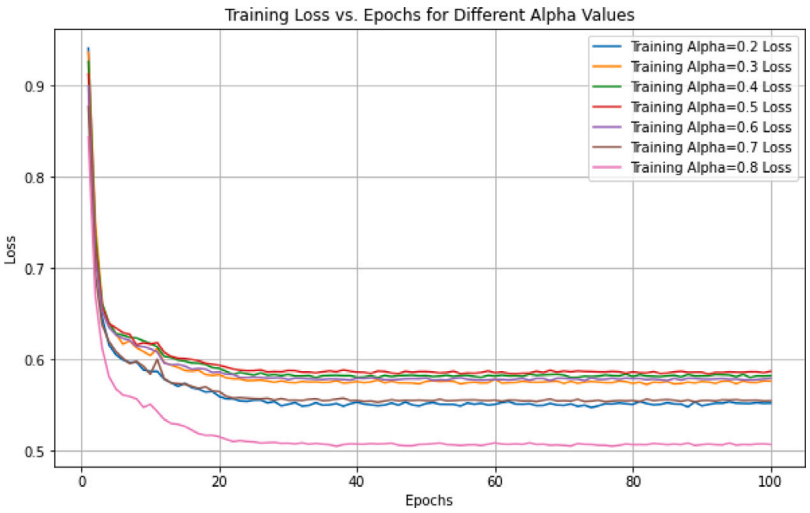


Fig. 4. Empirical analysis of the Tversky Focal Loss parameter α for TILs segmentation.

is particularly well suited for histopathology images, where precise delineation of tissue boundaries is required. LinkNet also follows an encoder–decoder paradigm but incorporates residual connections in both components, enabling efficient feature reuse while reducing computational complexity. Its lightweight architecture makes it an attractive choice when balancing segmentation accuracy and computational efficiency for large-scale image analysis. In contrast, PSPNet emphasizes multi-scale contextual representation through pyramid pooling, allowing the integration of both local and global contextual information. This characteristic is especially beneficial for stromal region segmentation, where accurate identification of tissue structures requires understanding both fine boundaries and broader tissue organization.

The evaluation was further strengthened by the use of diverse backbone encoders. VGG16 provides a simple yet effective baseline for hierarchical feature extraction, InceptionResNetV2 combines residual learning with inception modules to enrich contextual representations, and DenseNet169 leverages dense connectivity to promote feature reuse and alleviate gradient vanishing. Together, these architecture–backbone combinations enabled a comprehensive comparison of different strategies for stromal segmentation.

Consistent with the TILs segmentation task, the best-performing model and backbone combination was selected based on quantitative segmentation metrics and subsequently adopted for stromal segmentation within the proposed framework.

The training strategy for stromal region segmentation closely followed that used for TILs segmentation (see Section 3.3.1). A five-fold cross-validation scheme was employed, with each fold trained on the preprocessed public dataset using distinct hyperparameter configurations, as summarized in Table 2. This strategy was adopted to improve model robustness and generalizability.

A pronounced class imbalance was observed in the stroma segmentation task, with stromal regions (foreground) being substantially less represented than non-stromal regions (background). To address this challenge, a hybrid loss function defined as the sum of Dice Loss (DL) and Binary Focal Loss (BFL, $\gamma = 2$) was employed. Dice Loss encourages accurate spatial overlap between predicted and reference regions, while Binary Focal Loss emphasizes harder-to-classify pixels and reduces the influence of the dominant background class. Their combined use helps balance gradient contributions during training and improves segmentation performance on underrepresented stromal regions.

For model initialization, encoder weights were initialized using ImageNet-pretrained parameters, and both encoder and decoder weights were optimized during training. Model performance was evaluated on both the validation folds and the held-out test set using three complementary metrics: pixel-wise accuracy, intersection over union (IoU), and Dice score. This evaluation protocol provided a comprehensive assessment of each model's ability to accurately segment stromal regions.

Finally, to enhance interpretability and better understand the features driving model predictions, the same explainability approach based on SegGrad-CAM [15] was applied to the selected stroma segmentation model.

3.3.3. sTILs quantification algorithm

The proposed sTILs quantification algorithm is designed to estimate the sTILs score by jointly analyzing the outputs of the TILs segmentation model and the stroma region segmentation model. Specifically, the algorithm operates on two binary maps: a TILs mask and a stroma mask, both generated at the ROI level.

To identify TILs located within stromal regions, a pixel-wise logical AND operation is applied between the two binary masks. This operation retains only those TIL pixels that overlap with stromal regions, while excluding TILs located outside the stroma from further analysis. This strategy is consistent with the recommendations of the IIOBWG, which define sTILs as lymphocytes present within the stromal compartment.

The sTILs score is defined as the percentage of stromal area occupied by TILs. To estimate the area covered by individual lymphocytes, the average lymphocyte diameter in breast cancer tissue — typically reported to range between 6 and 8 μm [58] — is converted from physical units to pixel units based on the scanner resolution. Let d denote the assumed lymphocyte diameter in micrometers and t the scanner resolution in micrometers per pixel (μm/px). The lymphocyte diameter in pixel units, D, is then given by:

$$D = \frac{d}{t} \quad \text{(pixels)}. \tag{1}$$

The area of a single TIL, denoted as X, is approximated using the area of a square bounding box that fully encloses the lymphocyte, with both width and height equal to the diameter in pixels:

$$X = D^2. \tag{2}$$

Bounding box-based area estimation was adopted instead of contour-based measurements due to the observed systematic underestimation of TIL surface area by the segmentation model during inference, as reflected by a Tversky score of approximately 0.69 ± 0.01 (see Section 4.1). This behavior is partly attributable to the intrinsic imaging characteristics of histopathology: lymphocytes are three-dimensional, opaque cellular structures that are visualized in H&E slides as two-dimensional projections under transmitted light

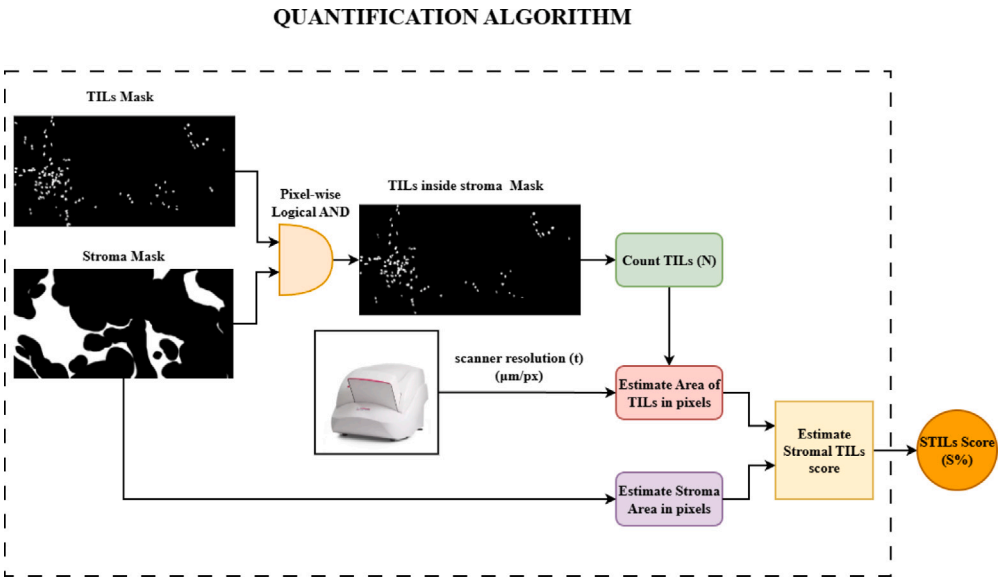


Fig. 5. Overview of the proposed sTILs quantification algorithm. The workflow illustrates the combination of TILs and stromal segmentation outputs, identification of stromal TILs, resolution-aware area estimation, and computation of the final sTILs score.

microscopy. As a result, optical effects such as halo formation, edge blurring, and gradual intensity transitions can introduce ambiguity in defining precise cell boundaries in the imaging plane, even for expert annotators. These factors make contour-based area estimation particularly sensitive to small segmentation inaccuracies.

In contrast, approximating lymphocyte area using fixed-size bounding boxes provides a more stable and robust estimate that is less affected by boundary uncertainty, local staining variability, and scanner-dependent optical differences. This choice improves consistency of TIL area estimation across ROIs and scanners, thereby enhancing the robustness of the downstream sTILs quantification algorithm.

The total TIL area within stromal regions is computed by multiplying the number of detected TILs located within the stroma, N, by the estimated area of a single TIL, X. The stromal area, denoted as Y, is calculated by summing the number of foreground (white) pixels in the stromal binary mask. Using these quantities, the sTILs score S is computed as:

$$S(\%) = \frac{N \times X}{Y} \times 100. \tag{3}$$

A schematic overview of the complete quantification workflow, illustrating the sequential steps from binary segmentation masks to the final sTILs score estimation, is provided in Fig. 5.

3.4. Validation phase

The validation of the proposed framework was conducted in a stepwise manner to assess both the performance of the individual segmentation models and the overall accuracy of the sTILs quantification pipeline. First, the TILs segmentation model and the stromal segmentation model were independently validated using the first private external dataset (Subset 1). The best-performing models identified during development were directly evaluated on this dataset in a zero-shot manner, without any fine-tuning, using the evaluation metrics previously defined for each task.

Subsequently, the complete framework was validated against visual sTILs scores provided by expert pathologists from the IISPV, Universitat Rovira i Virgili (URV), and the Medical University of Bialystok. The ROIs prepared as described in Section 3.2.2 were analyzed individually. Each ROI image was processed through the proposed framework to generate a predicted sTILs score.

The framework validation process began by generating binary segmentation maps for TILs and stromal regions. The TILs segmentation model and the stromal segmentation model with the best reported performance (described in the following section) were used for inference. Since both models operate on fixed-size image patches rather than full-resolution ROIs, each ROI was divided into overlapping tiles prior to inference. Predictions were generated for each tile independently and subsequently recombined to reconstruct full-resolution binary maps for TILs and stromal regions. Overlapping regions were resolved using a maximum voting strategy to ensure consistent and seamless reconstruction.

The resulting binary maps were then passed to the sTILs quantification algorithm to compute the final sTILs score for each ROI. Assuming a uniform lymphocyte diameter of $d = 8$ µm in breast cancer tissue and using the scanner-specific sampling resolution t for each dataset, as reported in Section 3.1.2, the sTILs score was calculated according to Eq. (3).

The predicted sTILs scores for all ROIs were subsequently recorded and compared with the visual scores assigned by expert pathologists. A schematic overview of the complete framework validation workflow is presented in Fig. 6.

To quantitatively evaluate the performance of the proposed framework for sTILs scoring, several statistical analyses were performed. Linear regression analysis was first applied to assess the goodness of fit between the framework-predicted scores and the pathologist-provided scores. In addition, multiple correlation metrics were computed to examine the strength and consistency of the relationship between the two scoring approaches. These included Pearson correlation, Spearman rank correlation, the Intraclass Correlation Coefficient (ICC) [59], and the Concordance Correlation Coefficient (CCC) [60]. Finally, the Cohen's kappa index [61] was calculated to evaluate the level of agreement between the framework-generated sTILs scores and the pathologists' visual assessments when scores were categorized into low, moderate, and high infiltrate groups according to the IIOBWG guidelines.

3.5. Computational setup

All experiments were conducted using the TensorFlow deep learning framework. The models developed for TILs segmentation were trained on an NVIDIA GeForce RTX 4080 graphics processing unit (GPU) equipped with 16 GB of memory. Due to the higher computational requirements of stromal region segmentation, these models were trained

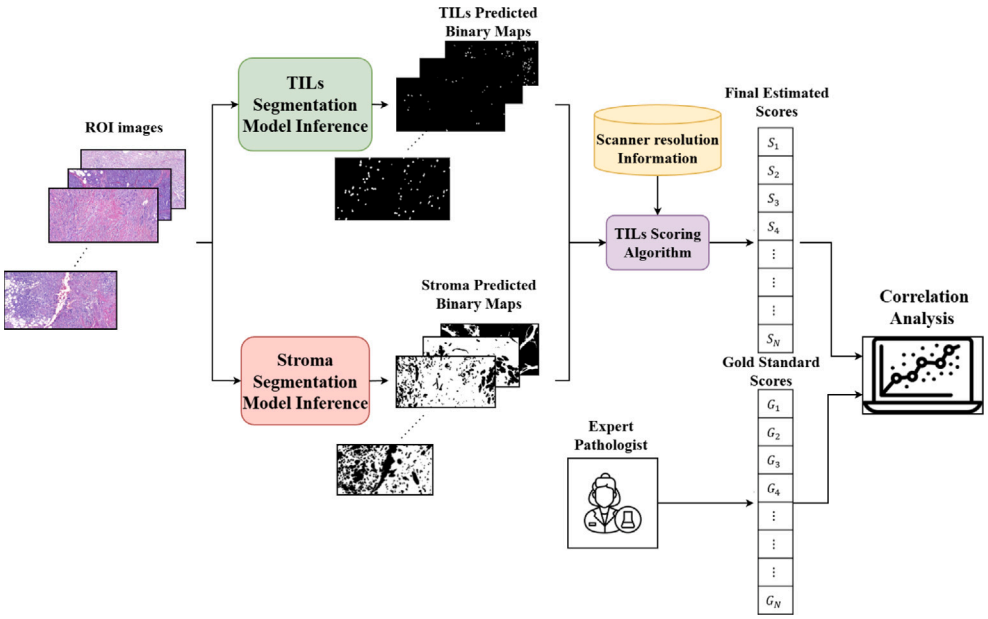


Fig. 6. Overview of the validation workflow for the proposed sTILs quantification framework, illustrating binary map generations, resolution-aware quantification, and comparison with pathologists' visual scores.

Table 3
Pixel-wise TILs segmentation performance on the validation data (mean ± std across five folds).

Model	Accuracy	Dice	Tversky
UNet	0.94 ± 0.01	0.59 ± 0.04	0.67 ± 0.03
Attention-UNet	**0.95 ± 0.01**	**0.60 ± 0.03**	0.66 ± 0.03
TransUNet	0.94 ± 0.01	0.59 ± 0.02	0.67 ± 0.03

Table 4
Pixel-wise TILs segmentation performance on the held-out test set (mean ± std).

Model	Accuracy	Dice	Tversky
UNet	0.95 ± 0.01	0.62 ± 0.02	0.70 ± 0.02
Attention-UNet	**0.95 ± 0.003**	**0.62 ± 0.01**	0.69 ± 0.01
TransUNet	0.94 ± 0.004	0.62 ± 0.02	0.70 ± 0.01

Table 5
Cell-level TILs detection performance on the validation data (mean ± std across five folds).

Model	Precision	Recall	F1-score
UNet	0.58 ± 0.08	0.53 ± 0.07	0.55 ± 0.07
Attention-UNet	**0.60 ± 0.03**	0.53 ± 0.06	**0.56 ± 0.04**
TransUNet	0.58 ± 0.05	0.54 ± 0.06	0.56 ± 0.04

Table 6
Cell-level TILs detection performance on the held-out test set (mean ± std).

Model	Precision	Recall	F1-score
UNet	0.62 ± 0.04	0.59 ± 0.07	0.60 ± 0.05
Attention-UNet	**0.63 ± 0.02**	0.59 ± 0.03	**0.61 ± 0.02**
TransUNet	0.61 ± 0.02	0.60 ± 0.03	0.61 ± 0.03

on an NVIDIA RTX A5000 GPU with 24 GB of memory. This hardware configuration enabled efficient training and evaluation of the proposed framework.

4. Results & discussion

4.1. TILs segmentation results

The performance of the three TILs segmentation models — UNet, Attention-UNet, and TransUNet — was evaluated using both pixel-wise segmentation metrics and cell-level detection metrics. Results are reported for the validation data aggregated across five cross-validation folds, as well as for the held-out test set, which was fixed and identical for all models. This evaluation strategy enables assessment of both model robustness during training and generalization to unseen data.

As shown in Tables 3 and 4, all three architectures achieved comparable pixel-wise segmentation performance on both the validation and held-out test datasets. High accuracy values were consistently observed across models; however, this metric should be interpreted with caution due to the pronounced class imbalance in TILs segmentation, where background pixels substantially outnumber lymphocyte pixels. In such settings, high accuracy largely reflects correct classification of background regions rather than precise delineation of the relatively sparse TIL regions.

Therefore, overlap-based metrics such as the Dice score and Tversky index provide a more informative assessment of segmentation quality, as they explicitly quantify agreement between predicted and reference foreground regions. The comparable Dice and Tversky values across models indicate similar capability in capturing TIL regions, with Attention-UNet exhibiting slightly lower standard deviations for Dice score across both datasets, suggesting more stable and consistent foreground segmentation across folds. Overall, these results indicate that while pixel-wise accuracy remains high due to class imbalance, Dice and Tversky metrics more reliably reflect the true segmentation performance relevant for downstream sTILs quantification.

The close agreement between validation and test performance for all pixel-wise metrics further indicates good generalization and suggests that none of the models suffered from substantial overfitting during training. However, given the minimal differences observed at the pixel level, pixel-wise metrics alone were insufficient to clearly distinguish the most suitable architecture for reliable TILs detection.

To further differentiate model performance in a clinically meaningful manner, cell-level detection metrics were therefore evaluated, as reported in Tables 5 and 6. At this level, clearer differences emerged among the models. Attention U-Net consistently achieved the highest

precision and F1-score on both the validation and test datasets, indicating a more reliable identification of true lymphocyte instances while reducing false-positive detections. Recall values were comparable across all models, suggesting a similar ability to capture the majority of TIL instances.

As with the pixel-wise analysis, the consistency between validation and test results at the cell-detection level supports the robustness and generalizability of the evaluated models. Based on its consistently strong and stable performance across both segmentation-level and cell-detection-level metrics, Attention-UNet was selected as the TILs segmentation component of the proposed framework.

Detailed training curves and fold-wise performance metrics for the selected model are provided in the Supplementary Material (Fig. S4 and Tables S1–S4).

To further interpret the segmentation and detection performance of the selected Attention-UNet model for TILs, SegGrad-CAM was employed to visualize the image regions that most strongly influenced the model's predictions. To improve robustness and reduce fold-specific variability, the final TILs predictions and corresponding attention maps were generated using an ensemble strategy across all five cross-validation folds. Specifically, pixel-wise probability maps from each fold were aggregated by taking their maximum, thereby emphasizing regions of consistent and confident agreement among the models.

Fig. 7 illustrates representative examples from four image samples selected from the held-out test set. For each sample, the original H&E image, the ground truth mask, the predicted TILs mask, the SegGrad-CAM heatmap, and overlays of the heatmap on both the predicted mask and the original image are shown. The visualizations indicate that the model predominantly focuses on lymphocyte-rich regions when generating predictions, which is consistent with the intended TILs segmentation task. Some uncertainty is observed in regions with low contrast, dense cellular packing, or ambiguous nuclear morphology, reflecting the inherent difficulty of lymphocyte detection in challenging histopathological contexts. In addition, a false-positive detections can be observed. These may originate from two primary sources: (i) annotation variability or ambiguity in the public training data, where small or borderline lymphocytic structures may be inconsistently labeled, and (ii) intrinsic model prediction errors in visually complex regions. Despite these cases, the attention maps provide meaningful insight into the model's decision-making process and support its overall reliability for TILs detection.

Overall, the combination of ensemble-based predictions and SegGrad-CAM visualization provides complementary qualitative evidence supporting the reliability of the proposed TILs segmentation approach, while also offering increased transparency into the model's decision-making process.

4.2. Stroma segmentation results

The performance of the stroma segmentation models was evaluated on both the validation and held-out test sets using a five-fold cross-validation strategy to assess robustness and generalizability. Accuracy, intersection over union (IoU), and Dice score were used as evaluation metrics, with quantitative results summarized in Tables 7 and 8.

Across all evaluated architectures and backbone combinations, the U-Net model with a VGG16 backbone consistently achieved the best overall performance on both datasets, yielding the highest accuracy, IoU, and Dice score, along with relatively low variability across folds. The close agreement between validation and test results indicates good generalization and limited overfitting. This performance can likely be attributed to two reasons. The first it is the balanced representational capacity of the VGG16 encoder, which appears to generalize more effectively on histopathological data than deeper backbones. The second it is the skip connections of the U-Net architecture that facilitate precise delineation of heterogeneous stromal regions. Importantly, the selected model was able to accurately segment both tumoral and infiltrated stromal regions, which is essential for reliable downstream sTILs quantification.

Detailed training curves and fold-wise performance metrics for both the validation and test sets of the best performing model are provided in the Supplementary Material (Fig. S5 and Tables S5 and S6).

To further investigate the behavior of the best-performing stroma segmentation model, an explainability analysis was conducted using SegGrad-CAM. To enhance robustness and reduce fold-specific variability, an ensemble strategy was employed across the five cross-validation folds by computing the pixel-wise maximum of the predicted probability maps. This approach emphasizes regions of consistent agreement among the models.

Representative qualitative results from the held-out test set are presented in Fig. 8, including the input H&E image, ground truth stroma mask, model prediction, SegGrad-CAM heatmap, and overlays of the heatmap on both the predicted mask and the original image. The visualizations show that the model predominantly attends to stromal regions, with particular emphasis on areas infiltrated by lymphocytes, which is consistent with the intended segmentation objective and supports its suitability for downstream sTILs quantification. In the presented masks, white regions correspond to stroma, while black regions represent other tissue structures.

Some over- and under-segmentation can be observed at stromal boundaries. These discrepancies are partly attributable to the intrinsic difficulty of precisely delineating stromal borders, even for expert pathologists, due to gradual transitions and morphological heterogeneity within the tissue. In addition, some errors reflect inherent model limitations. Importantly, these localized inaccuracies did not substantially affect the overall sTILs scoring performance.

4.3. Validation results

The proposed framework was externally validated using three private datasets collected at a single institution under different annotation and acquisition settings (see Section 3.1.2). Each subset provides a complementary validation perspective based on the availability of reference annotations and the intended evaluation objective. The first subset enables direct assessment of both segmentation performance and sTILs scoring against expert annotations, while the second and third subsets support validation of the sTILs scoring component alone, including an evaluation of robustness to scanner-related variability.

4.3.1. External validation on Subset 1: Segmentation and sTILs scoring

The first private validation subset enables direct external evaluation of the segmentation components and the complete sTILs quantification framework against expert annotations. Zero-shot evaluation was performed using the best-performing models identified during development, namely Attention U-Net for TILs segmentation and VGG16-UNet for stroma segmentation.

TILs segmentation performance: Table 9 reports the pixel-wise performance of the Attention-UNet model under zero-shot evaluation on Subset 1. As expected in TILs segmentation, accuracy remains very high due to the strong imbalance between background and lymphocyte pixels, and therefore mainly reflects correct background classification. More informative overlap-based metrics, including Dice and Tversky, indicate that the model maintains reliable delineation of TIL regions when applied to unseen private data. These results support the generalization of the segmentation component beyond the public training cohort.

At the cell-detection level, the model demonstrated balanced performance, as summarized in Table 10. Precision and recall values indicate a stable ability to identify lymphocytes while limiting false-positive detections, supporting the suitability of the model for downstream sTILs quantification.

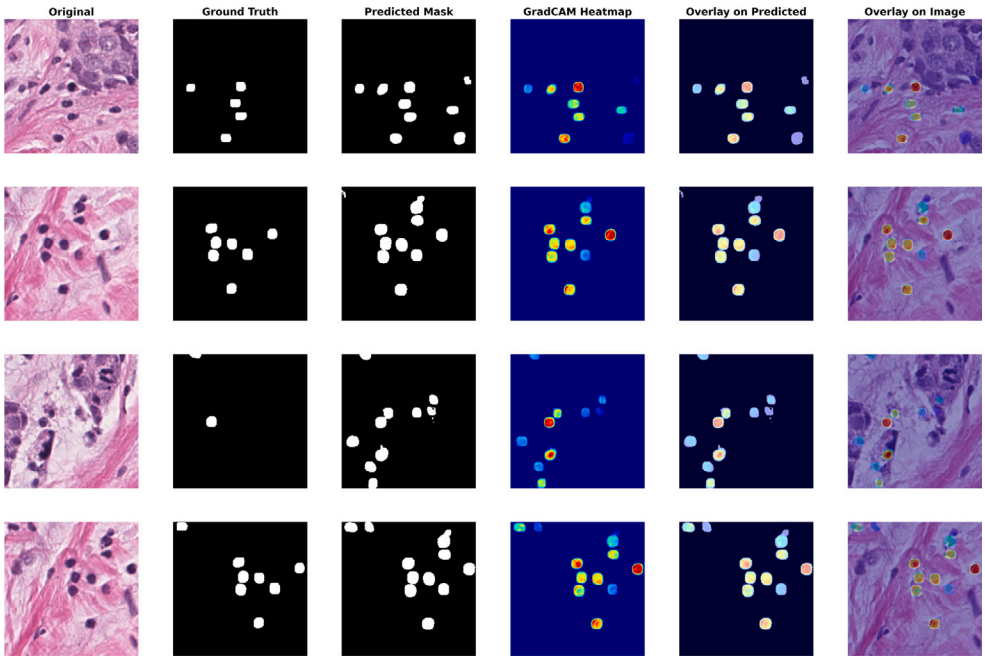


Fig. 7. Qualitative visualization of TILs segmentation and explainability using SegGrad-CAM on representative samples from the held-out test set. For each example, the original H&E image, ground truth mask, predicted TILs mask, SegGrad-CAM heatmap, and overlays on both the predicted mask and the original image are shown. The heatmaps indicate the relative contribution of image regions to the model's predictions, with lower relevance shown in cooler colors (light blue) and higher relevance highlighted in warmer colors (red).

Table 7
Stroma segmentation performance on the validation dataset (mean ± standard deviation).

Model	Backbone	Accuracy	IoU	Dice
UNet	VGG16	**0.83 ± 0.03**	**0.54 ± 0.05**	**0.67 ± 0.04**
	InceptionResNetV2	0.81 ± 0.03	0.48 ± 0.04	0.62 ± 0.04
	DenseNet169	0.82 ± 0.03	0.49 ± 0.06	0.63 ± 0.05
LinkNet	VGG16	0.83 ± 0.03	0.52 ± 0.04	0.65 ± 0.04
	InceptionResNetV2	0.81 ± 0.04	0.48 ± 0.06	0.62 ± 0.06
	DenseNet169	0.82 ± 0.03	0.50 ± 0.06	0.63 ± 0.05
PSPNet	VGG16	0.77 ± 0.15	0.51 ± 0.11	0.64 ± 0.10
	InceptionResNetV2	0.78 ± 0.04	0.48 ± 0.08	0.61 ± 0.08
	DenseNet169	0.82 ± 0.03	0.52 ± 0.07	0.65 ± 0.06

Table 8
Stroma segmentation performance on the held-out test dataset (mean ± standard deviation).

Model	Backbone	Accuracy	IoU	Dice
UNet	VGG16	**0.84 ± 0.01**	**0.57 ± 0.02**	**0.69 ± 0.01**
	InceptionResNetV2	0.82 ± 0.01	0.52 ± 0.02	0.65 ± 0.02
	DenseNet169	0.83 ± 0.01	0.53 ± 0.04	0.66 ± 0.03
LinkNet	VGG16	0.84 ± 0.01	0.55 ± 0.01	0.68 ± 0.01
	InceptionResNetV2	0.82 ± 0.02	0.51 ± 0.04	0.64 ± 0.03
	DenseNet169	0.83 ± 0.01	0.53 ± 0.03	0.66 ± 0.02
PSPNet	VGG16	0.78 ± 0.14	0.54 ± 0.11	0.66 ± 0.10
	InceptionResNetV2	0.80 ± 0.03	0.51 ± 0.08	0.64 ± 0.06
	DenseNet169	0.82 ± 0.02	0.55 ± 0.03	0.67 ± 0.03

Table 9
Pixel-wise TILs segmentation performance on Subset 1 using Attention U-Net (zero-shot evaluation).

Metric	Mean ± Std
Accuracy	0.98 ± 0.003
Dice	0.67 ± 0.06
Tversky	0.70 ± 0.10

Table 10
Cell-level TILs detection performance on Subset 1 using Attention U-Net (zero-shot evaluation).

Metric	Mean ± Std
Precision	0.69 ± 0.10
Recall	0.77 ± 0.18
F1-score	0.71 ± 0.07

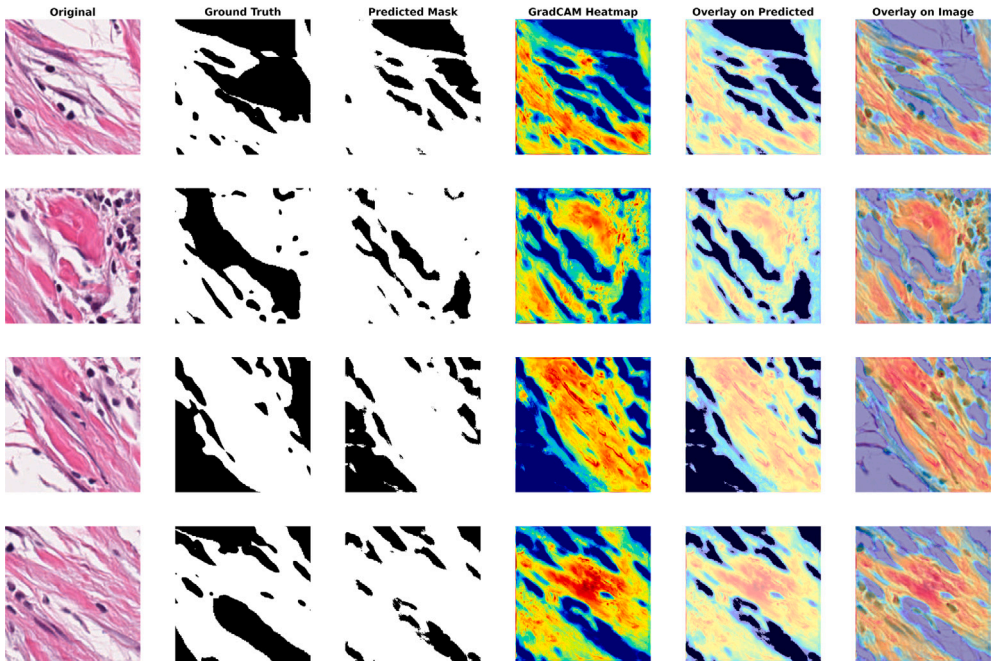


Fig. 8. Qualitative visualization of stroma segmentation and explainability using SegGrad-CAM. For each example, the input H&E image, ground truth stroma mask, predicted mask, SegGrad-CAM heatmap, and overlays on both the prediction and the original image are shown. The heatmaps indicate the relative contribution of image regions to the model's predictions, with warmer colors (red/yellow) corresponding to higher relevance and cooler colors (blue) indicating lower relevance. White regions indicate stroma, while black regions denote other tissue structures.

Table 11
Pixel-wise stroma segmentation performance on Subset 1 using VGG16-U-Net (zero-shot evaluation).

Metric	Mean ± Std
Accuracy	0.74 ± 0.03
IoU	0.55 ± 0.04
Dice	0.70 ± 0.03

Table 12
Average processing time of the full framework per ROI on Subset 1.

Number of ROIs	Processing time (min)
60	0.52 ± 0.09

Stroma segmentation performance: The zero-shot performance of the stroma segmentation model (VGG16-UNet) on Subset 1 is summarized in Table 11. The obtained Dice and IoU values indicate robust identification of stromal regions in externally collected data, supporting reliable downstream sTILs computation.

Detailed fold-wise segmentation results for both TILs and stroma models are provided in the Supplementary Material (Tables S7–S9).

sTILs scoring performance: Following segmentation validation, the complete framework was evaluated for sTILs scoring by comparing predicted scores against independent visual assessments provided by two expert pathologists, as well as their averaged reference score. In addition, the computational efficiency of the full framework was assessed by measuring the processing time per ROI. On average, the framework required approximately half a minute per ROI, demonstrating its practical feasibility for routine analysis (see Table 12).

Fig. 9 presents scatter plots comparing framework-predicted sTILs scores with scores assigned by Pathologist 1, Pathologist 2, and their average. The corresponding linear regression analyses are shown in Fig. 10, demonstrating clear positive associations across all comparisons. Stronger agreement was consistently observed when benchmarking against the averaged reference, indicating closer alignment with consensus expert judgment.

Table 13 summarizes the correlation and agreement analyses between the framework-predicted sTILs scores and expert visual assessment on Subset 1. When compared with the averaged reference of both pathologists, the framework demonstrated consistent agreement across correlation and reproducibility metrics, indicating stable alignment with expert consensus.

Higher agreement was observed with Pathologist 1 than with Pathologist 2, which is consistent with the known variability inherent to manual sTILs evaluation. In addition to continuous score agreement, categorical consistency was evaluated using the kappa statistic, which measures agreement between the framework and pathologists when classifying cases into low, moderate, and high immune infiltrate categories. The resulting kappa values indicate meaningful categorical agreement, supporting the framework's potential for clinically relevant patient stratification.

Subtype-specific correlation analysis against the averaged reference is reported in Table 14. Strong agreement was observed for Luminal A cases, while moderate correlations were obtained for Luminal B tumors. Although high correlation values were observed for TNBC cases, these results should be interpreted with caution due to the very limited number of samples in this subgroup. HER2+ cases were excluded from this analysis because of insufficient sample size.

In addition to subtype-specific analysis, we investigated the sensitivity of the sTILs scoring framework to variations in the assumed lymphocyte diameter used in the quantification algorithm. Although lymphocyte diameters in breast cancer are commonly reported to range between approximately 6 and 8 μm [58], inter-cellular variability and the presence of larger activated lymphocytes motivated the evaluation of a broader, yet biologically plausible, diameter range. Specifically, correlation and agreement metrics were computed using three representative diameters (6, 8, and 10 μm), corresponding to a nominal diameter of 8 μm with a ±2 μm margin to assess the robustness of the scoring procedure.

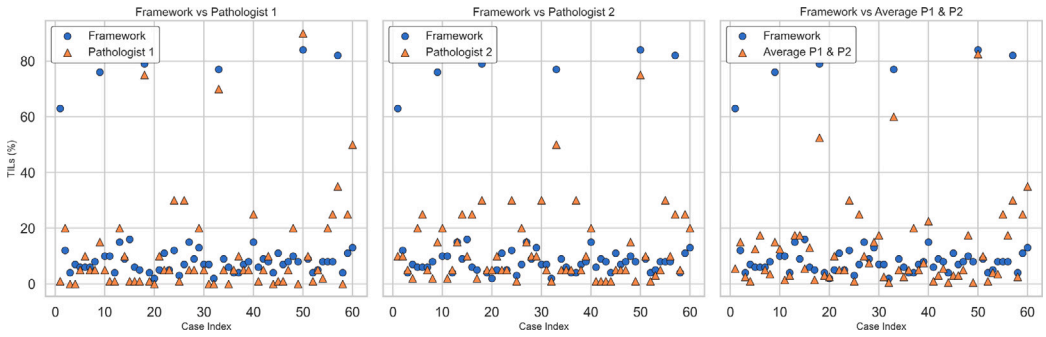


Fig. 9. Comparison between framework-predicted sTILs scores and expert visual assessment on Subset 1. Scatter plots show agreement with Pathologist 1 (left), Pathologist 2 (middle), and the averaged reference of both pathologists (right).

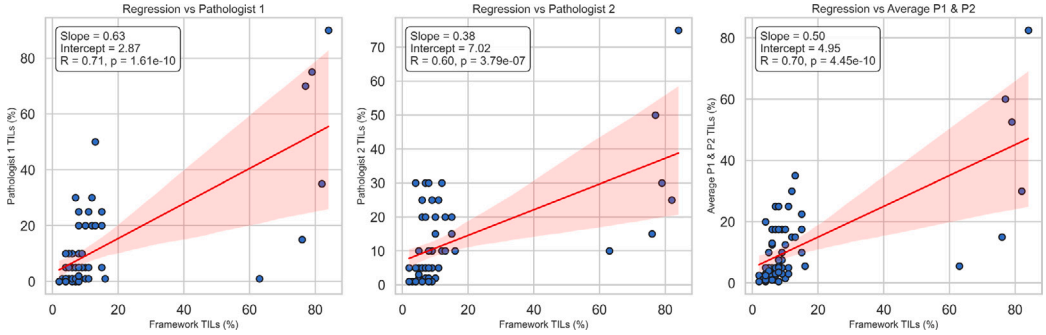


Fig. 10. Linear regression analysis between framework-predicted sTILs scores and expert visual assessment on Subset 1. Regression plots correspond to Pathologist 1 (left), Pathologist 2 (middle), and the averaged reference of both pathologists (right). Shaded regions indicate 95% confidence intervals.

Table 13
Correlation and agreement analysis between framework-predicted sTILs scores and expert visual assessment on Subset 1. The kappa statistic evaluates agreement in categorical sTILs classification (low, moderate, high immune infiltrate).

Comparison	Pearson	Spearman	ICC	CCC	Kappa
Framework vs Pathologist 1	0.71	0.61	0.70	0.70	0.74
Framework vs Pathologist 2	0.60	0.51	0.54	0.54	0.59
Framework vs Average P1 & P2	0.70	0.62	0.66	0.66	0.61

Table 14
Subtype-specific correlation analysis between framework-predicted sTILs scores and averaged expert reference on Subset 1.

Subtype	N	Pearson	Spearman	ICC	CCC
Luminal A	27	0.72	0.59	0.65	0.64
Luminal B	29	0.50	0.51	0.49	0.48
TNBC	3	0.78	1.00	0.74	0.66

The results are summarized in Table 15. Using a diameter of 8 μm yielded the most balanced and consistent performance across correlation and agreement metrics, supporting its selection as the default parameter in this study. Reducing the assumed diameter to 6 μm resulted in higher Pearson correlation but was associated with a noticeable decrease in agreement metrics (ICC and CCC), indicating reduced consistency with expert scoring. Conversely, increasing the diameter to 10 μm led to a general decline in both correlation and agreement measures. Overall, these findings suggest that while sTILs scores are moderately sensitive to lymphocyte size assumptions, the selected diameter of 8 μm provides a robust compromise between correlation strength and agreement with pathologist assessments.

4.3.2. External validation on Subset 2: sTILs scoring

The second private validation subset was used to externally evaluate the sTILs scoring component of the proposed framework. This subset includes ROI-level sTILs scores provided by a single expert pathologist, without pixel-level segmentation annotations. Accordingly, validation focused exclusively on score-level agreement between the framework predictions and the visual reference.

In addition to accuracy, the computational efficiency of the framework was assessed on this subset. The average processing time per ROI was measured to evaluate practical applicability, indicating that the framework maintains stable runtime performance when applied to a larger number of samples (see Table 16).

Fig. 11 presents scatter plots and corresponding regression analysis comparing the framework-predicted sTILs scores with the pathologist-provided scores. A clear positive association is observed across the dataset, indicating consistent alignment between automated and visual assessments.

Quantitative correlation and agreement metrics are summarized in Table 17. The framework achieved strong Pearson and Spearman correlations, alongside good agreement as reflected by the ICC and CCC, indicating reliable and reproducible continuous sTILs score estimation on independently collected data.

In addition, categorical agreement was assessed using the kappa statistic, which evaluates the consistency between the framework and

Table 15
Sensitivity analysis of sTILs scoring with respect to assumed lymphocyte diameter (Subset 1, averaged reference).

TIL Diameter (μm)	Pearson	Spearman	ICC	CCC
6	0.83	0.62	0.54	0.54
8	0.70	0.62	0.66	0.66
10	0.60	0.61	0.46	0.46

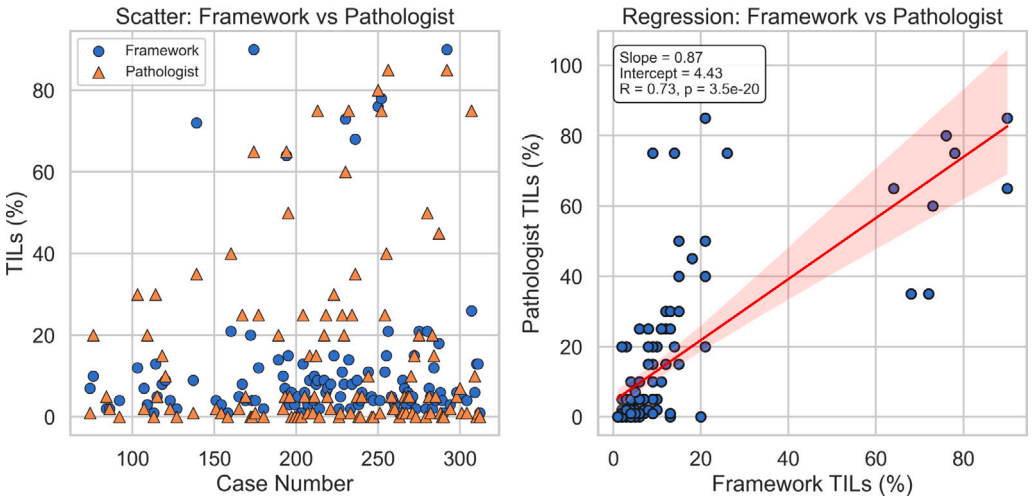


Fig. 11. External validation of sTILs scoring on Subset 2. Left: scatter plot comparing framework-predicted sTILs scores with the pathologist reference. Right: corresponding linear regression analysis with 95% confidence interval.

Table 16
Average processing time of the full framework per ROI on Subset 2.

Number of ROIs	Processing time (min)
112	0.58 ± 0.20

the pathologist in classifying cases into low, moderate, and high immune infiltrate categories based on established sTILs scoring thresholds. The obtained kappa value indicates substantial agreement at the categorical level, further supporting the clinical relevance of the proposed framework.

To further assess robustness across breast cancer molecular subtypes, a subtype-specific correlation analysis was performed, as summarized in Table 18. Strong agreement was observed for Luminal B cases, while moderate to good correlations were obtained for Luminal A tumors. For HER2-positive and TNBC subtypes, correlation estimates should be interpreted with caution due to the limited number of available cases; nevertheless, the observed trends suggest consistent framework behavior across subtypes.

Again, to study the sensitivity of the sTILs scoring framework to variations in the assumed lymphocyte diameter was evaluated on Subset 2. Correlation and agreement metrics were computed using three representative diameters (6, 8, and 10 μm), reflecting biologically plausible lymphocyte size variability.

The results are reported in Table 19. Consistent with observations from Subset 1, assuming a uniform lymphocyte diameter of 8 μm yielded the most balanced and stable performance across both correlation and agreement metrics. While slightly higher Pearson correlation values were obtained when increasing the diameter to 10 μm, this was accompanied by a reduction in agreement measures (ICC and CCC). Conversely, reducing the diameter to 6 μm led to a marked decrease in agreement despite moderate correlation values. Overall, these findings further support the selection of 8 μm as a robust compromise that ensures consistency between framework predictions and expert visual assessment.

Overall, the results obtained on Subset 2 demonstrate that the proposed framework maintains good agreement with expert visual scoring in an independent cohort. Despite limited sample sizes for certain molecular subtypes, the framework shows consistent and robust sTILs quantification performance across breast cancer subgroups.

4.3.3. External validation on Subset 3: Scanner invariance analysis

The third private validation subset was designed to evaluate the robustness of the proposed framework to scanner-related variability, with a particular focus on differences in imaging hardware and spatial resolution. This subset includes WSIs digitized using two whole-slide imaging systems from distinct vendors and institutions, enabling an assessment of the framework's invariance to scanner-specific acquisition characteristics in sTILs scoring.

Table 20 summarizes the correlation and agreement metrics obtained when comparing framework-predicted sTILs scores with the expert visual reference for each scanner. Consistent levels of agreement were observed across both scanning systems, as reflected by comparable Pearson and Spearman correlation coefficients, as well as ICC and CCC values.

The results indicate that the framework maintains stable performance across scanners from different vendors and operating at slightly different spatial resolutions. While minor variations in agreement metrics were observed, the overall consistency suggests that the proposed segmentation and resolution-aware quantification strategy effectively mitigates scanner-related variability. These findings support the applicability of the framework in multi-scanner and multi-institutional clinical settings.

4.4. Comparison to the state of the art

To assess the proposed framework in relation to existing approaches, we conducted a comparative evaluation with the recently introduced MuTILs model by Liu et al. [46]. As MuTILs was originally developed for panoptic segmentation with multiple tissue and cell labels, we reimplemented it as a single multi-task model producing two binary

Table 17
Correlation and agreement analysis between framework-predicted sTILs scores and expert visual assessment on Subset 2. The kappa statistic reflects agreement in categorical sTILs classification (low, moderate, high immune infiltrate).

Comparison	Pearson	Spearman	ICC	CCC	Kappa
Framework vs Pathologist	0.73	0.75	0.72	0.72	0.73

Table 18
Subtype-specific correlation analysis between framework-predicted sTILs scores and expert reference on Subset 2.

Subtype	N	Pearson	Spearman	ICC	CCC
Luminal A	49	0.63	0.65	0.49	0.48
Luminal B	49	0.75	0.66	0.69	0.69
HER2+	4	0.69	0.95	0.70	0.63
TNBC	10	0.49	0.52	0.51	0.47

Table 19
Sensitivity analysis of sTILs scoring with respect to assumed lymphocyte diameter on Subset 2.

TIL diameter (μm)	Pearson	Spearman	ICC	CCC
6	0.67	0.75	0.42	0.42
8	0.73	0.75	0.72	0.72
10	0.74	0.75	0.69	0.69

Table 20
Scanner-specific correlation and agreement analysis for sTILs scoring on Subset 3.

Vendor	Scanner model	Resolution (μm/px)	Pearson	Spearman	ICC	CCC
Hamamatsu Photonics	NanoZoomer 2.0 HT (C9600)	0.2270	0.76	0.75	0.62	0.62
3DHISTECH Ltd.	Panoramic 250 Flash III	0.2433	0.73	0.67	0.72	0.72

Table 21
Comparison of TILs segmentation performance (Dice score, mean ± std) between MuTILs and the proposed Attention-UNet.

Model	Public test data	Private data
MuTILs [46]	0.47 ± 0.01	0.21 ± 0.04
Proposed Attention-UNet	$\mathbf{0.62 \pm 0.01}$	$\mathbf{0.67 \pm 0.06}$

Table 22
Comparison of stroma segmentation performance (Dice score, mean ± std) between MuTILs and the proposed VGG16-UNet.

Model	Public test data	Private data
MuTILs [46]	0.63 ± 0.01	0.56 ± 0.04
Proposed VGG16-UNet	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.70 \pm 0.03}$

outputs, corresponding to TILs and stroma segmentation, to enable a fair and consistent comparison. Both approaches were trained using the same public training dataset under a five-fold cross-validation scheme and evaluated on the same held-out public test set. In addition, zero-shot evaluation was performed on the private datasets to assess generalization to unseen data. Segmentation performance was compared using the Dice score.

TILs segmentation comparison: Table 21 reports the Dice scores obtained for TILs segmentation on both public test data and private data. On the public test set, the proposed Attention-UNet outperformed the adapted MuTILs model, achieving higher Dice scores with lower variability. This performance difference became more evident on the private data, where the MuTILs model exhibited a notable decrease in segmentation accuracy, while the proposed model maintained robust performance. These findings suggest improved generalization of the proposed approach to external data with different characteristics. Detailed fold-wise results for the MuTILs model on both the public test set and private data are provided in Supplementary Table S10.

Stroma segmentation comparison: A similar trend was observed for stroma segmentation, as summarized in Table 22. While MuTILs achieved reasonable performance on the public test set, the proposed VGG16-UNet consistently yielded higher Dice scores on both public and private datasets. Notably, the proposed model demonstrated improved robustness on private data, suggesting better adaptability to variations in staining, tissue morphology, and acquisition conditions. Detailed fold-wise Dice scores for the MuTILs stroma segmentation results on both datasets are provided in Supplementary Table S11.

Overall, the proposed segmentation models demonstrated consistently stronger performance than the adapted MuTILs baseline across both public and private datasets. The improved generalization observed on private data suggests that the proposed framework, supported by targeted architecture selection and robust data augmentation, is better suited for deployment in heterogeneous, real-world clinical settings.

sTILs scoring comparison: In addition to segmentation performance, we compared the full sTILs scoring capability of the proposed framework against the adapted MuTILs approach on the private validation datasets. For this analysis, sTILs scores were computed using the TILs and stroma segmentation outputs generated by each method and compared against expert visual assessments using correlation, agreement, and categorical consistency metrics.

Table 23 summarizes the results obtained on Subset 1 and Subset 2. On both datasets, the proposed framework consistently achieved higher Pearson and Spearman correlations, as well as improved ICC and CCC values, indicating stronger agreement with expert reference scores. Moreover, categorical agreement, assessed using the kappa statistic for low, moderate, and high immune infiltrate classification, was also higher for the proposed framework across both subsets.

These results suggest that improvements in segmentation accuracy and robustness translate into more reliable downstream sTILs quantification. While both approaches show reasonable performance, the proposed framework demonstrates more consistent alignment with expert assessment, particularly in terms of score reproducibility and categorical stratification on independent data.

Table 23
Comparison of sTILs scoring performance between MuTILs and the proposed framework on private validation subsets.

Subset	Method	Pearson	Spearman	ICC	CCC	Kappa
Subset 1	MuTILs [46]	0.66	0.53	0.45	0.44	0.51
	Proposed Framework	**0.70**	**0.62**	**0.66**	**0.66**	**0.61**
Subset 2	MuTILs [46]	0.70	0.62	0.39	0.38	0.56
	Proposed Framework	**0.73**	**0.75**	**0.72**	**0.72**	**0.73**

Table 24
Benchmarking of the proposed sTILs quantification framework against recent state-of-the-art studies. N.A. indicates metrics not reported in the corresponding work.

Method	Evaluation data	#Cases	#Path.	Pearson	Spearman	ICC	CCC	Kappa
Makhlouf et al. [24]	Private	300	1	N.A.	N.A.	0.70	N.A.	N.A.
Choi et al. [27]	Private	402	3	N.A.	N.A.	N.A.	0.73	N.A.
Yosofvand et al. [30]	Private	63	2	0.98	N.A.	N.A.	N.A.	N.A.
Kim et al. [32]	TIGER	82	1	0.72	N.A.	N.A.	N.A.	N.A.
Liu et al. [46]	PanopTILs	305	2	N.A.	0.58	N.A.	N.A.	N.A.
Proposed Framework	Private (Subset 1)	60	2	0.70	0.62	0.66	0.66	0.61
	Private (Subset 2)	112	1	0.73	0.75	0.72	0.72	0.73

4.5. Benchmarking

To position the proposed framework with respect to existing approaches, a comparative benchmarking analysis was conducted against representative state-of-the-art methods for sTILs quantification reported in recent literature (Section 2). The comparison considers correlation- and agreement-based metrics, while explicitly accounting for differences in evaluation datasets, cohort sizes, and the number of expert pathologists involved, in order to provide a fair and transparent assessment.

Table 24 summarizes the reported performance of existing frameworks alongside the results obtained by the proposed method on two independent private cohorts. Prior studies have demonstrated promising correlations with pathologist scoring; however, these results were often derived from limited datasets, lack of explainability, or experimental setups without strict separation between training and evaluation data.

Among the compared methods, Yosofvand et al. [30] reported a notably high Pearson correlation (0.975). Nevertheless, this performance was achieved under an evaluation protocol where the same slides were used for both training the segmentation models and scoring assessment, which may result in optimistic estimates and does not reflect true generalization performance on unseen data.

In comparison, Kim et al. [32] reported a Pearson correlation of 0.718 on the TIGER dataset using whole-slide-level scoring by a single pathologist. The proposed framework demonstrated consistent performance on the 60-sample cohort (Pearson = 0.70) and improved correlation on the larger 112-sample cohort (Pearson = 0.73), despite operating at the ROI level. ROI-based evaluation focuses on tumor-relevant regions explicitly selected by expert pathologists and mitigates dilution effects that arise in whole-slide scoring due to the inclusion of heterogeneous tissue areas. In particular, whole-slide analysis may average over diverse stromal compartments, including non-tumoral stroma, adipose tissue, and benign structures, which can obscure localized immune infiltration patterns and reduce the sensitivity of sTILs quantification.

Additionally, unlike approaches that rely on uniformly resized WSIs, the proposed method preserves spatial resolution and incorporates scanner-aware quantification.

The recently introduced MuTILs-based approach by Liu et al. [46], evaluated on the PanopTILs dataset, reported a Spearman correlation of 0.580. In contrast, the proposed framework achieved higher rank-based correlations of 0.615 and 0.750 on the two private cohorts, indicating a stronger ability to preserve the relative ordering of sTILs scores under external validation conditions. With respect to agreement metrics, the proposed framework achieved ICC and CCC values of 0.72 on the 112-sample cohort, comparable to or exceeding those reported by Makhlouf et al. [24] and Choi et al. [27]. These results indicate not only strong correlation with expert scoring but also reliable reproducibility across a larger evaluation cohort.

Importantly, the proposed framework is the only method among those compared to report Cohen's kappa index based on clinically relevant categorization of sTILs into low, moderate, and high infiltration groups following established guidelines. The achieved kappa values (0.61 and 0.73) indicate substantial categorical agreement with expert pathologists, underscoring the framework's clinical relevance for patient stratification and decision-making.

Overall, the benchmarking results demonstrate that the proposed framework achieves competitive or superior performance relative to existing methods, while uniquely combining strict external validation, scanner-aware quantification, and categorical agreement analysis. These characteristics support its robustness and suitability for clinical trials and for future real-world clinical deployment.

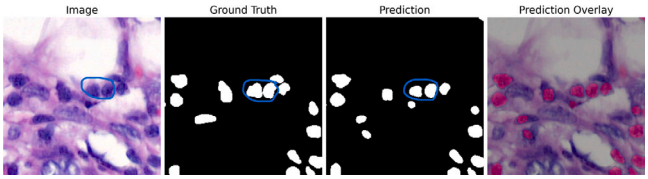
5. Limitations

Although the proposed framework demonstrates strong performance and produces clinically meaningful sTILs estimates, several limitations should be acknowledged from a technical and methodological perspective.

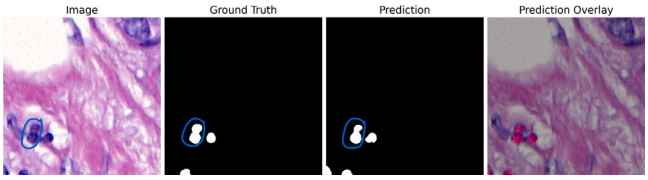
First, with respect to the TILs segmentation model, watershed-based preprocessing was applied prior to training to encourage the separation of touching lymphocytes into individual cell instances. In several cases, the model successfully distinguishes partially connected lymphocytes, as illustrated in Fig. 12(a). However, when lymphocytes are fully overlapping or densely compacted, separation remains challenging and such cells may be detected as a single instance, as shown in Fig. 12(b). This limitation can affect the total number of detected TILs and, consequently, influence the robustness of the sTILs quantification.

In addition, occasional false-positive lymphocyte detections were observed, as highlighted by the explainability visualizations in Fig. 7. These detections may further contribute to variability in the estimated TIL counts. Future work will explore instance-aware segmentation approaches, such as StarDist-based models, to better address fully overlapping lymphocytes and improve the accuracy of TILs instance detection.

Second, although the stroma segmentation model generalized well across both public and private datasets, minor over- and under-segmentation was observed at stromal boundaries in some cases. These boundary regions are inherently challenging to delineate due to the high morphological variability of stroma and the gradual transitions



(a) Partially touching lymphocytes successfully separated by the model.



(b) Fully overlapping lymphocytes where separation remains challenging.

Fig. 12. Illustration of limitations in TILs instance separation. While watershed preprocessing and the segmentation model can resolve partially connected lymphocytes (a), fully overlapping cells may still be detected as a single instance (b), which can affect TILs counting and downstream sTILs quantification.

between tissue compartments, which can also pose difficulties for expert pathologists. While these local inaccuracies did not substantially affect the overall sTILs scoring, further refinement of boundary modeling remains an important direction for improvement. Future work will investigate multi-scale feature modeling and the incorporation of squeeze-and-excitation blocks to enhance contextual awareness and improve boundary consistency in complex stromal regions.

Finally, the reference sTILs scores used for external validation were visually assessed by two expert pathologists. While this provides a strong and reliable baseline, involving a larger and more diverse panel of experts would allow for a more comprehensive characterization of inter-observer variability, which is an inherent challenge in manual TILs assessment. In addition, scanner invariance was evaluated using data acquired from two digital slide scanners. Although encouraging results were obtained, extending validation to a broader range of scanner vendors and acquisition settings would further strengthen the generalizability of the framework.

In summary, these limitations highlight natural directions for continued development rather than fundamental shortcomings. Despite them, the proposed pipeline achieved good agreement with expert visual assessment and demonstrated robust generalization on external data. Future efforts will focus on refining instance-level segmentation, expanding expert involvement, and validating the framework across a wider range of scanners and multi-institutional cohorts to further enhance clinical readiness and scalability.

6. Conclusions & future work

This study presents an AI-based framework for the detection and quantification of stromal tumor-infiltrating lymphocytes (sTILs) in H&E-stained breast cancer histopathology images. The proposed framework was developed in alignment with the recommendations of the IIOBWG for breast cancer. It is important to note that these guidelines are cancer-type specific and cannot be directly generalized to other malignancies, where TILs assessment criteria and clinical relevance may differ.

The proposed solution integrates two complementary segmentation components: an attention-based network for TILs detection, designed to capture fine-grained lymphocytic features through attention mechanisms, and a UNet with VGG16 beckbone-based model for stroma segmentation, optimized to identify both tumoral and infiltrated stroma as a single clinically relevant compartment. To enhance robustness and generalizability, both models were trained using a comprehensive augmentation strategy that includes geometric transformations, stain-like augmentation, elastic deformation, and resolution scaling. These augmentations were explicitly designed to simulate variability in staining protocols and scanner resolutions, thereby improving adaptability across multi-institutional datasets.

To further ensure generalizability and minimize data leakage, data splitting and cross-validation were performed at the hospital level. The framework was evaluated on both a held-out public test set and multiple private external datasets. This validation strategy demonstrates that the proposed system maintains stable performance when applied to unseen data, which is a critical requirement for clinical deployment.

Explainability was incorporated into both segmentation components to support interpretability of model predictions, while a scanner-aware resolution-based quantification algorithm was introduced to enable robust sTILs estimation across data acquired with different scanning devices. Beyond quantitative performance, the framework provides practical benefits for pathologiest and in consequence for their patients.

For pathologists, it offers a standardized, reproducible, and efficient tool for sTILs scoring, reducing inter-observer variability and supporting diagnostic decision-making. The framework follows a pathologist-in-the-loop paradigm, in which expert pathologists define one or multiple relevant ROIs per whole-slide image based on tumor localization and dispersion, after which sTILs are automatically quantified through parallel TILs and stroma segmentation pathways combined in a unified scoring algorithm. The complete computational processing from ROI input to final score requires on average 0.5 ± 0.05 minutes per ROI on a single GPU, with potential for further acceleration in multi-GPU environments due to the parallel architecture. For patients, reliable and reproducible sTILs quantification may contribute to improved prognostication and support personalized treatment strategies, particularly in breast cancer subtypes where immune infiltration has established predictive or prognostic value [62].

Overall, this work lies within the scope of translational research, as it bridges advanced AI-based methodological developments with practical clinical needs, supporting the integration of automated sTILs quantification into real-world breast cancer pathology workflows.

Building on the current framework and addressing its limitations, several directions for future work can be envisaged. First, instance-aware segmentation approaches, such as StarDist-based models, will be investigated to better address cases of fully overlapping lymphocytes that remain challenging for pixel-wise segmentation, thereby improving the accuracy of individual TIL detection. In addition, incorporating active learning strategies may help reduce false-positive detections and further improve precision by iteratively guiding expert review toward the most uncertain or error-prone regions, enhancing model performance while minimizing additional annotation effort.

Second, to further improve stroma segmentation — particularly at challenging stromal boundaries — future work will investigate the

incorporation of multi-scale feature representations and squeeze-and-excitation (SE) blocks. Multi-scale modeling can enhance the capture of both local texture details and broader contextual information, which is essential for accurately delineating heterogeneous stromal regions. SE blocks, by adaptively recalibrating channel-wise feature responses, may further strengthen the model's sensitivity to clinically relevant stromal patterns and improve boundary consistency across varying tissue morphologies.

Third, beyond ImageNet-pretrained backbones, future studies should explore foundation models specifically developed for histopathology, which are expected to provide richer domain-specific representations and further improve both TILs and stroma segmentation accuracy. These directions align with the improvements identified in the limitation analysis.

Finally, expanding validation to include a broader range of scanners and multi-institutional cohorts will be essential to support future clinical translation of the proposed framework. Such validation will allow systematic assessment of robustness across diverse acquisition protocols, institutional practices, and patient populations, which is critical for regulatory acceptance and real-world deployment. Ultimately, integrating the framework into routine digital pathology workflows — while preserving the pathologist-in-the-loop paradigm — will facilitate prospective clinical studies aimed at evaluating its impact on diagnostic efficiency, reproducibility, and clinical decision-making.

Overall, these proposed above complementing development directions will contribute to a more robust, scalable, and clinically applicable solution for automated sTILs quantification. Given its efficiency and reproducibility, the proposed framework represents a promising and potentially cost-effective complement to manual microscopic evaluation in routine clinical practice.

CRediT authorship contribution statement

Shrief Abdelazeez: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alessio Fiorin:** Writing – review & editing, Investigation. **Vincenzo Della Mea:** Writing – review & editing, Investigation. **Laia Adalid:** Writing – review & editing, Validation, Investigation, Data curation. **Krzysztof Siemion:** Writing – review & editing, Visualization, Resources. **Carlos Lopez:** Writing – review & editing, Validation, Investigation, Funding acquisition. **Marylene Lejeune:** Writing – review & editing, Resources. **Anna Korzynska:** Writing – review & editing, Validation, Supervision, Resources, Investigation, Formal analysis, Conceptualization.

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 study is part of the **BOSOMSHIELD** European Project, funded under the Horizon Europe Research and Innovation Program (Grant Agreement No. 101073222). Further 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. We also extend my gratitude to **Dr. Łukasz Roszkowiak** for providing us a powerful machine to train my model for stroma region segmentation. We would like to express our sincere appreciation to **Dr. Hatem Rashwan** for his continuous support and guidance throughout this work.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2026.110003.

Data availability

The data that has been used is confidential.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 68 (6) (2018) 394–424, http://dx.doi.org/10.3322/caac.21492.

[2] C.E. DeSantis, J. Ma, M.M. Gaudet, L.A. Newman, K.D. Miller, A. Goding Sauer, A. Jemal, R.L. Siegel, Breast cancer statistics, 2019, CA: Cancer J. Clin. 69 (6) (2019) 438–451, http://dx.doi.org/10.3322/caac.21583.

[3] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, 2023, CA: Cancer J. Clin. 73 (1) (2023) 17–48, http://dx.doi.org/10.3322/caac.21763.

[4] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, K. Ruddy, J. Tsang, F. Cardoso, Breast cancer, Nat. Rev. Dis. Prim. 5 (1) (2019) 66, http://dx.doi.org/10.1038/s41572-019-0111-2.

[5] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: The next generation, Cell 144 (5) (2011) 646–674, http://dx.doi.org/10.1016/j.cell.2011.02.013.

[6] D.F. Quail, J.A. Joyce, Microenvironmental regulation of tumor progression and metastasis, Nature Med. 19 (11) (2013) 1423–1437, http://dx.doi.org/10.1038/nm.3394.

[7] C. López, et al., The immune response in nonmetastatic axillary lymph nodes is associated with the presence of axillary metastasis and breast cancer patient outcome, Am. J. Pathol. 190 (3) (2020) 660–673, http://dx.doi.org/10.1016/j.ajpath.2019.11.002.

[8] C. López, et al., Development of automated quantification methodologies of immunohistochemical markers to determine patterns of immune response in breast cancer: A retrospective cohort study, BMJ Open 4 (8) (2014) http://dx.doi.org/10.1136/bmjopen-2014-005643.

[9] W.H. Fridman, F. Pagès, C. Sautès-Fridman, J. Galon, The immune contexture in human tumours: impact on clinical outcome, Nat. Rev. Cancer 12 (4) (2012) 298–306, http://dx.doi.org/10.1038/nrc3245.

[10] M.V. Dieci, et al., Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ, Sem. Cancer Biol. 52 (2018) 16–25, http://dx.doi.org/10.1016/j.semcancer.2017.10.003.

[11] R. Salgado, 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, http://dx.doi.org/10.1093/annonc/mdu450.

[12] S. Loi, et al., Tumor-infiltrating lymphocytes and prognosis: A pooled individual patient analysis of early-stage triple-negative breast cancers, J. Clin. Oncol. 37 (7) (2019) 559–569, http://dx.doi.org/10.1200/JCO.18.01010.

[13] S. Hendry, et al., Assessing tumor-infiltrating lymphocytes in solid tumors: A practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group, Adv. Anat. Pathol. 24 (5) (2017) 235–251.

[14] M. Veta, P.J. van Diest, S.M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A.B.L. Larsen, J.S. Vestergaard, A.B. Dahl, D. Ciresan, et al., Assessment of algorithms for mitosis detection in breast cancer histopathology images, Med. Image Anal. 20 (1) (2015) 237–248, http://dx.doi.org/10.1016/j.media.2014.11.010.

[15] K. Vinogradova, A. Dibrov, G. Myers, Towards interpretable semantic segmentation via gradient-weighted class activation mapping, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 13943–13944, http://dx.doi.org/10.1609/aaai.v34i10.7244.
[16] C. Denkert, et al., Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy, Lancet Oncol. 19 (1) (2018) 40–50, http://dx.doi.org/10.1016/S1470-2045(17)30904-X.
[17] R. Bai, Z. Lv, D. Xu, J. Cui, Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors, Biomark. Res. 8 (1) (2020) 1–17, http://dx.doi.org/10.1186/s40364-020-00209-0.
[18] L.A. Emens, Breast cancer immunotherapy: Facts and hopes, Clin. Cancer Res. 24 (3) (2018) 511–520, http://dx.doi.org/10.1158/1078-0432.CCR-16-3001.
[19] F. Klauschen, et al., Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning, Sem. Cancer Biol. 52 (2018) 151–157, http://dx.doi.org/10.1016/j.semcancer.2018.07.001.
[20] P. Bankhead, et al., QuPath: Open source software for digital pathology image analysis, Sci. Rep. 7 (1) (2017) 1–7, http://dx.doi.org/10.1038/s41598-017-17204-5.
[21] J. Thagaard, G. Broeckx, et al., Pitfalls in machine learning-based assessment of tumor-infiltrating lymphocytes in breast cancer: A report of the International Immuno-Oncology Biomarker Working Group, J. Pathol. 260 (5) (2023) 498–513, http://dx.doi.org/10.1002/path.6155.
[22] A. Heindl, I. Sestak, K. Naidoo, J. Cuzick, M. Dowsett, Y. Yuan, Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer, J. Natl. Cancer Inst. 110 (2) (2018) 166–175, http://dx.doi.org/10.1093/jnci/djx137.
[23] 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.e7, http://dx.doi.org/10.1016/j.celrep.2018.03.086.
[24] S. Makhlouf, et al., Evaluation of tumour infiltrating lymphocytes in luminal breast cancer using artificial intelligence, Br. J. Cancer 129 (11) (2023) 1747–1758, http://dx.doi.org/10.1038/s41416-023-02451-3.
[25] W. Weng, X. Zhu, INet: Convolutional networks for biomedical image segmentation, IEEE Access 9 (2021) 16591–16603, http://dx.doi.org/10.1109/ACCESS.2021.3053408.
[26] S. Graham, et al., Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Med. Image Anal. 58 (2019) 101563, http://dx.doi.org/10.1016/j.media.2019.101563.
[27] S. Choi, et al., Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer, Npj Breast Cancer 9 (1) (2023) 1–13, http://dx.doi.org/10.1038/s41523-023-00577-4.
[28] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.
[29] S.C. Yurtkulu, Y.H. Şahin, G. Unal, Semantic segmentation with extended DeepLabv3 architecture, in: 2019 27th Signal Processing and Communications Applications Conference, SIU, 2019, pp. 1–4, http://dx.doi.org/10.1109/SIU.2019.8806244.
[30] M. Yosofvand, et al., Automated detection and scoring of tumor-infiltrating lymphocytes in breast cancer histopathology slides, Cancers 15 (14) (2023) http://dx.doi.org/10.3390/cancers15143635.
[31] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 386–397, http://dx.doi.org/10.1109/TPAMI.2018.2844175.
[32] S. Kim, et al., Improving tumor-infiltrating lymphocytes score prediction in breast cancer with self-supervised learning, Life 14 (1) (2024) 1–15, http://dx.doi.org/10.3390/life14010090.
[33] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, ICML, 2020, pp. 1575–1585.
[34] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. (2020).
[35] S. Akbar, M. Peikari, S. Salama, A.Y. Panah, S. Nofech-Mozes, A.L. Martel, Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment, Sci. Rep. 9 (1) (2019) http://dx.doi.org/10.1038/s41598-019-50568-4.
[36] G. Nir, et al., Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts, Med. Image Anal. 50 (2018) 167–180, http://dx.doi.org/10.1016/j.media.2018.09.005.
[37] J.N. Kather, et al., Collection of textures in colorectal cancer histology, 2016, http://dx.doi.org/10.5281/zenodo.53169.
[38] J.N. Kather, N. Halama, A. Marx, 100,000 histological images of human colorectal cancer and healthy tissue, 2018, http://dx.doi.org/10.5281/zenodo.1214456.
[39] A. Gupta, MiMM_SBILab dataset: Microscopic images of multiple myeloma, 2018, http://dx.doi.org/10.7910/DVN/XCX7ST.
[40] N.V. Orlov, et al., Automatic classification of lymphoma images with transform-based global features, IEEE Trans. Inf. Technol. Biomed. 14 (4) (2010) 1003–1013, http://dx.doi.org/10.1109/TITB.2010.2050695.
[41] N. Shahadat, Lung and colon cancer histopathological image classification using 1D convolutional channel-based attention networks, Proc. Fla. Artif. Intell. Res. Soc. Conf. 37 (2024) http://dx.doi.org/10.32473/flairs.37.1.135538.
[42] N. Kumar, et al., A multi-organ nucleus segmentation challenge, IEEE Trans. Med. Imaging 39 (5) (2020) 1380–1391, http://dx.doi.org/10.1109/TMI.2019.2947628.
[43] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, IEEE Trans. Biomed. Eng. 63 (7) (2016) 1455–1462, http://dx.doi.org/10.1109/TBME.2015.2496264.
[44] G. Aresta, et al., BACH: Grand challenge on breast cancer histology images, Med. Image Anal. 56 (2019) 122–139, http://dx.doi.org/10.1016/j.media.2019.05.010.
[45] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: A comprehensive tutorial, J. Pathol. Informatics 7 (1) (2016) 29, http://dx.doi.org/10.4103/2153-3539.186902.
[46] S. Liu, M. Amgad, D. More, M.A. Rathore, R. Salgado, L.A.D. Cooper, A panoptic segmentation dataset and deep-learning approach for explainable scoring of tumor-infiltrating lymphocytes, Npj Breast Cancer 10 (1) (2024) http://dx.doi.org/10.1038/s41523-024-00663-1.
[47] S. Liu, M. Amgad, M.A. Rathore, R. Salgado, L.A.D. Cooper, A panoptic segmentation approach for tumor-infiltrating lymphocyte assessment: Development of the MuTILs model and panoptils dataset, MedRxiv (2023) http://dx.doi.org/10.1101/2022.01.08.22268814.
[48] M. Amgad, et al., Structured crowdsourcing enables convolutional segmentation of histology images, Bioinformatics 35 (February) (2019) 3461–3467, http://dx.doi.org/10.1093/bioinformatics/btz083.
[49] M. Amgad, et al., NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer, GigaScience 11 (2022) 1–12, http://dx.doi.org/10.1093/gigascience/giac037.
[50] A. Korzynska, L. Roszkowiak, J. Zak, K. Siemion, A review of current systems for annotation of cell and tissue images in digital pathology, Biocybern. Biomed. Eng. 41 (4) (2021) 1436–1453, http://dx.doi.org/10.1016/j.bbe.2021.04.012.
[51] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, J. van der Laak, Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology, Med. Image Anal. 58 (2019) 101544, http://dx.doi.org/10.1016/j.media.2019.101544.
[52] P.Y. Simard, D. Steinkraus, J.C. Platt, Best practices for convolutional neural networks applied to visual document analysis, in: Proceedings of the Seventh International Conference on Document Analysis and Recognition, ICDAR, 2003, pp. 958–963, http://dx.doi.org/10.1109/ICDAR.2003.1227801.
[53] O. Oktay, et al., Attention U-net: Learning where to look for the pancreas, 2018, arXiv:1804.03999, URL http://arxiv.org/abs/1804.03999.
[54] J. Chen, et al., TransUNet: Transformers make strong encoders for medical image segmentation, 2021, arXiv:2102.04306 URL http://arxiv.org/abs/2102.04306.
[55] N. Abraham, N.M. Khan, A novel focal tversky loss function with improved attention U-net for lesion segmentation, in: Proceedings of the IEEE International Symposium on Biomedical Imaging, ISBI, 2019, pp. 683–687, http://dx.doi.org/10.1109/ISBI.2019.8759329.
[56] A. Chaurasia, E. Culurciello, LinkNet: Exploiting encoder representations for efficient semantic segmentation, in: Proceedings of the IEEE Visual Communications and Image Processing, VCIP, 2017, pp. 1–4, http://dx.doi.org/10.1109/VCIP.2017.8305148.
[57] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6230–6239, http://dx.doi.org/10.1109/CVPR.2017.660.
[58] Z. Swiderska-Chadaj, et al., Learning to detect lymphocytes in immunohistochemistry with deep learning, Med. Image Anal. 58 (2019) 101547, http://dx.doi.org/10.1016/j.media.2019.101547.
[59] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2) (2016) 155–163, http://dx.doi.org/10.1016/j.jcm.2016.02.012.
[60] L.I.-K. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1) (1989) 255–268, http://dx.doi.org/10.2307/2532051.
[61] M.L. McHugh, Lessons in biostatistics: Interrater reliability—The kappa statistic, Biochem. Medica 22 (3) (2012) 276–282.
[62] A. Ciarka, M. Piątek, R. Pęksa, M. Kunc, E. Senkus, Tumor-infiltrating lymphocytes (TILs) in breast cancer: Prognostic and predictive significance across molecular subtypes, Biomedicines 12 (4) (2024) http://dx.doi.org/10.3390/biomedicines12040763.