Article

Not peer-reviewed version

Bridging The Gap: Assessing Interobserver Variability in Tumour-Infiltrating Lymphocyte (TIL) Scoring for Triple-Negative Breast Cancer

Nurkhairul Bariyah Baharun , Mohamed Afiq Hidayat Zailani , Afzan Adam , Qiaoyi XU , Muaatamarulain Mustangin , Reena Rahayu Md Zin *

Posted Date: 10 July 2025

doi: 10.20944/preprints202507.0874.v1

Keywords: tumour-infiltrating lymphocytes (TIL); triple-negative breast cancer (TNBC); manual TILs scoring; interobserver variability; intraclass correlation; diagnostic reproducibility

 doi:10.20944/preprints202507.0874.v1

Article

Bridging The Gap: Assessing Interobserver Variability in Tumour-Infiltrating Lymphocyte (TIL) Scoring for Triple-Negative Breast Cancer

Nurkhairul Bariyah Baharun [1,2], Mohamed Afiq Hidayat Zailani [1], Afzan Adam [3], Qiaoyi Xu [3] and Muaatamarulain Mustangin [1]and Reena Rahayu Md Zin [1,*]

[1] Department of Pathology, Faculty of Medicine, The National University of Malaysia, Jalan Yaacob Latif, Bandar Tun Razak, 56000 Cheras, Wilayah Persekutuan Kuala Lumpur, Malaysia

[2] Department of Medical Diagnostic, Faculty of Health Sciences, Universiti Selangor, Jalan Zirkon A7/7, Seksyen 7, 40000 Shah Alam, Selangor, Malaysia

[3] Centre for Artificial Intelligence Technology (CAIT), Faculty of Information Science & Technology, The National University of Malaysia, 43600 Bangi, Selangor, Malaysia

* Correspondence: reenarahayu@ppukm.ukm.edu.my

Abstract

Background: Tumour-infiltrating lymphocytes (TILs) are emerging as a crucial prognostic biomarker in triple-negative breast cancer (TNBC). However, their clinical utility remains constrained by the subjectivity and interobserver variability of manual scoring, despite standardization efforts by the International TILs Working Group (TIL-WG). This study aimed to evaluate the interobserver agreement among pathologists in scoring stromal and intratumoral TILs from H&E-stained TNBC slides and to identify contributing histological factors. **Methods:** Two consultant pathologists at Hospital Canselor Tuanku Muhriz, Kuala Lumpur independently assessed 64 TNBC cases using TIL-WG guidelines. Interobserver agreement was quantified using intraclass correlation coefficient (ICC) and Cohen's Kappa coefficient. Cases with over 10% scoring discrepancies underwent review by a third pathologist and a consensus discussion was held to explore the underlying confounders. **Results:** Our results showed moderate interobserver agreement for stromal TILs (ICC = 0.58) and strong agreement for intratumoral TILs (ICC = 0.71). Significant variability was attributed to three main confounding variables: heterogeneous TIL distribution, poorly-defined tumor-stroma interface and focal dense lymphoid infiltrates. **Conclusions:** These findings underscore the critical need for standardized TILs assessment criteria advocating for the adoption of AI-based scoring method to overcome interobserver variability and enhance the reproducibility of evaluations.

Keywords: tumour-infiltrating lymphocytes (TIL); triple-negative breast cancer (TNBC); manual TILs scoring; interobserver variability; intraclass correlation; diagnostic reproducibility

1. Introduction

Triple-negative breast cancer (TNBC) is an aggressive immunogenic [1], subtype of breast cancer characterized by the absence of oestrogen, progesterone, and Human Epidermal Growth Factor Receptor 2 (HER-2) [2] presenting significant challenges in prognostication and treatment due to limited targeted therapies. Within the tumour microenvironment, tumour-infiltrating lymphocytes (TILs), have emerged as a valuable prognostic biomarker in TNBC [3–7] with high densities correlating with improved patient outcomes, including overall and disease-free survival. Therefore, accurate TILs assessment is crucial for understanding the immune landscape in TNBC. The International TILs Working Group (TIL-WG) has established standardised guidelines for evaluating stromal TILs (sTILs) by quantifying their proportion within tumour-associated stromal area while

 doi:10.20944/preprints202507.0874.v1

also outlining specific exclusion criteria [8]. The density of sTILs can be calculated using the following formula:

$$\text{sTILs Density} = \left(\frac{\text{Area of sTILs}}{\text{Area of tumor-associated stroma}}\right) \times 100\%$$

In contrast, the density of intratumoural TILs (iTILs) is represented by the percentage of the area occupied by TILs within the area of the tumour epithelium [9]. The following formula describes the calculation for the iTILs density:

$$\text{iTILs Density} = \left(\frac{\text{Area of iTILs}}{\text{Area of tumor epithelium}}\right) \times 100\%$$

Areas of regressive hyalinisation, necrosis, TILs around in situ carcinoma, normal structures, and previous needle biopsy sites should be excluded in the assessment [8].

Despite these standardised criteria and endorsements for clinical use manual visual assessment by pathologists remains inherently subjective [8,10]. This often leads to variability in scoring thresholds, and significant interobserver discrepancies. Such inconsistencies can undermine the reliability of pathology reports, potentially impacting clinical trials outcome and patient therapeutic decisions [11–13]. Consequently, ensuring the reproducibility of TIL scoring is paramount for consistent clinical management. Interobserver agreement, commonly assessed using the intraclass correlation coefficient (ICC) for continuous data and Kohen's cappa for categorical data, is vital for evaluating TIL assessment reproducibility across different observers [14–18].

Previous studies have reported a wide range of ICC values (0.50 -0.933) [11,17,19,20] and Cohen's Kappa values, (0.21 to 0.881) [11,18,19] highlighting considerable variability in agreement levels influenced by factors such as study design, observer experience, and methodology. This variability stems from a complex interplay of biological, technical, and human factors. Biological contributors include intra-tumor heterogeneity, characterised by uneven TIL distributions leading to inconsistent region of interest (ROIs) selection [15,16,19] as well as ambiguous features like abundant tertiary lymphoid structures or poorly defined tumor-stroma interface [20,21]. Technical factors such as suboptimal slide quality inadequate tissue processing or artifacts can obscure cellular details, critical for accurate lymphocytes identification [22]. Furthermore, human elements including pathologists' varying experience levels, personal interpretation styles, and familiarity with TIL scoring guidelines, can lead to scoring discrepancies, particularly in challenging or borderline cases [23,24].

The observed discrepancies in agreement underscore a critical need for enhanced standardization and additional training in TILs assessment methodologies to improve consistency and ensure quality control in histopathological evaluations. Therefore, this study aims to evaluate the interobserver agreement among pathologists when scoring both sTILs and iTILs in TNBC cases utilizing a standardized assessment criterion. Secondarily, this study seeks to identify specific histological features that contribute to these scoring discrepancies.

2. Materials and Methods

2.1. Sample Collection and Dataset Description

A prospective study was conducted on 64 TNBC cases diagnosed at Hospital Canselor Tuanku Muhriz (HCTM), Kuala Lumpur between January 2012 and December 2021. Inclusion criteria comprised TNBC cases with no prior history of malignancies within 5 years of diagnosis, verified availability of formalin fixed paraffin- embedded (FFPE) tissue blocks, and sufficient material for H&E and immunohistochemistry (IHC) analysis. Exclusion criteria included inadequate tissue quality, technical artifacts, or failure to meet staining standards which may jeopardize TILs assessment.

TNBC status was confirmed through IHC for oestrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER-2). ER and PR negativity were defined as <1% positive tumour nuclei, per American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) guidelines [25]. HER-2 negative status was defined by IHC scores of 0, 1+,

 doi:10.20944/preprints202507.0874.v1

and 2+, without HER-2 gene amplification confirmed by fluorescence in situ hybridization (FISH) or dual-colour dual hapten in situ hybridization (DDISH) on cases where HER-2 score was equivocal. Selected FFPE tissue blocks were sectioned at 4 μm for H&E and IHC staining.

2.2. Pathologist Selection and Manual TIL Assessment

The selected and quality-checked slides were scanned at 20x magnification using Pannoramic DESK II DW slide scanner. Figure 1 illustrates the general workflow of pathologists for manual TILs assessment and annotation. Two board- certified pathologists (P1 and P2) with comparable clinical experience and expertise in breast pathology independently assessed TILs using digital whole slide images (WSI). Both were trained on the standardized TIL assessment criteria established by the TIL-WG. A calibration session was conducted prior to scoring to ensure adherence to protocol and minimize interpretive variability.

Each pathologist independently annotated regions of interests (ROIs) and scored sTILs and iTILs as continuous variable (0% to 100%). The annotation and scoring process were guided by corresponding IHC, CD4+ and CD8+ markers to enhance the accuracy of lymphocyte identification and to minimize ambiguity in distinguishing TILs within the tumor microenvironment. Stroma TILs were quantified as the percentage of tumor-associated stroma occupied by mononuclear inflammatory cells, excluding areas of necrosis, hyalinization, and in situ components. Intratumoural TILs were assessed as the proportion of lymphocytes within tumor epithelial nests. TILs score was further categorised using five binary cut-off systems: ≤10%, ≤20%, ≤30%, ≤40%, and ≤50%. Low and high TIL classifications were defined relative to each threshold (Table 1). For cases exhibiting more than 10% scoring discrepancy between P1 and P2, a third independent pathologist (P3) adjucated the score. The final consensus score was determined by averaging the two most concordant score among the three reviewers.

Table 1. Different cut-off systems for TILs assessment.

Cut-off system	TILs classification	
	Low TILs	High TILs
Cut off value 1	≤10%	>10%
Cut off value 2	≤20%	>20%
Cut off value 3	≤30%	>30%
Cut off value 4	≤40%	>40%
Cut off value 5	≤50%	>50%

 doi:10.20944/preprints202507.0874.v1

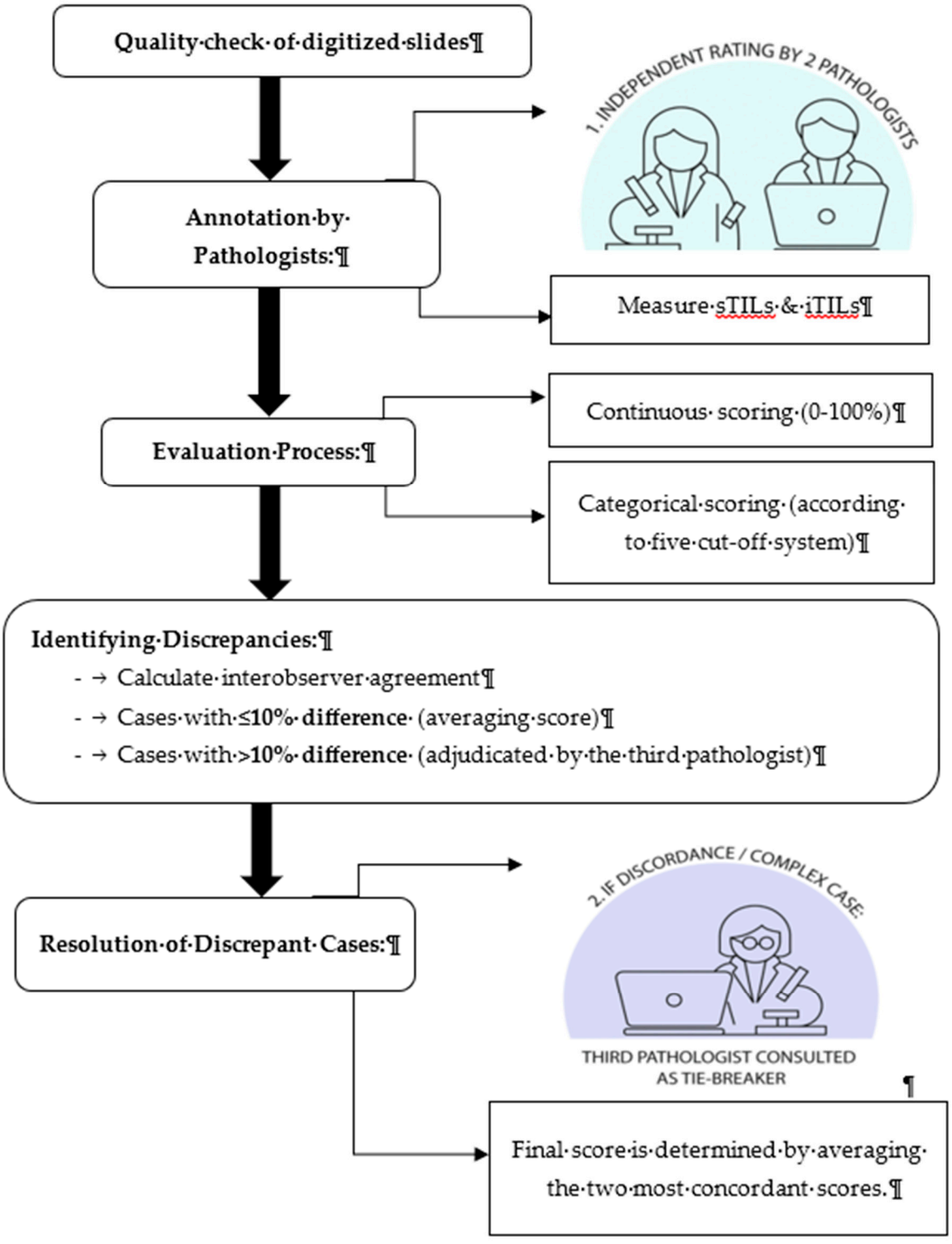

Figure 1. Workflow of manual TILs assessment and annotation by the pathologist.

2.3. Interobserver Agreement and Consensus Review.

Interobserver reliability was analysed for both categorical and continuous TILs scores. Cohen's kappa coefficient was applied to assess agreement for categorical TIL classifications across the five cut-off thresholds with interpretation based on Landis and Koch criteria [26]. Intraclass correlation coefficients (ICCs) using two-way random effects models assessed agreement and consistency for continuous TIL data, with interpretation based on the guidelines of Koo and Li [27]. Bland–Altman plots were generated to visualize agreement and identify any systematic biases. For discrepant cases

 doi:10.20944/preprints202507.0874.v1

reviewed by P3, pairwise Kappa and ICC values were recalculated to evaluate whether consensus review improved interobserver agreement.

3. Results

3.1. Interobserver Agreement Using Continuous Scores

Assessment of interobserver agreement for continuous TIL scoring between pathologists P1 and P2 showed moderate consistency for sTILs with ICC values of 0.57 (agreement) and 0.58 (consistency) respectively (95% CI, $p < 0.001$), In contrast, iTILs exhibited stronger agreement, with ICC of 0.70 (agreement) and 0.75 (consistency) indicating good reliability (Table 2).

Table 2. Intraclass correlation coefficients for consistency in sTILs and iTILs assessment.

Reliability measure	sTILs ICC (95% CI)	iTILs ICC (95% CI)
ICC (agreement)	0.57	0.70
ICC (consistency)	0.58	0.75

Abbreviations: sTILs, stromal tumour-infiltrating lymphocytes; iTILs, intratumoral tumour-infiltrating lymphocytes; ICC, intraclass correlation coefficient.

3.2. Interobserver Agreement Using Categorical Scores

Cohen's kappa (κ) coefficient was calculated for five TIL cut-off thresholds (10%, 20%, 30%, 40%, and 50%) to assess categorical agreement between the two pathologists. For sTILs, kappa values ranged from 0.13 to 0.40, with the highest agreement observed at the 10% (κ = 0.40), indicating a moderate level of interobserver agreement. Similarly, iTILs showed kappa values ranging from 0.25 to 0.48, with the 40% cut-off demonstrating the highest level of agreement (κ =0.48). Agreement declined with increasing cut-off thresholds, suggesting greater subjectivity in identifying higher TIL densities (Table 3).

Table 3. Cohen's kappa (κ) coefficient for interobserver agreement in sTILs and iTILs at various cut-off thresholds between P1 and P2.

Cut-off (%)	sTILs (κ)	iTILs (κ)
10	0.40*	0.43*
20	0.29	0.31
30	0.13	0.25
40	0.16	0.48*
50	0.34	0.38

Note: Asterisks (*) indicate the highest agreement values for each category.

To visualize the discordance among pathologists based on the sTIL and iTILs density scores, Bland-Altman plots were drawn (Figure 2 (A) and (B). Bland–Altman plots were constructed to visualize scoring variability. For sTILs, the mean difference between observers was -4.26, with wide limits of agreement (-52.89 to +44.35), indicating moderate to high variability, particularly at higher densities. For iTILs, the mean difference was -8.22 with narrower limits of agreement (-37.98 to +20), reflecting relatively better consistency (Figure 2).

 doi:10.20944/preprints202507.0874.v1

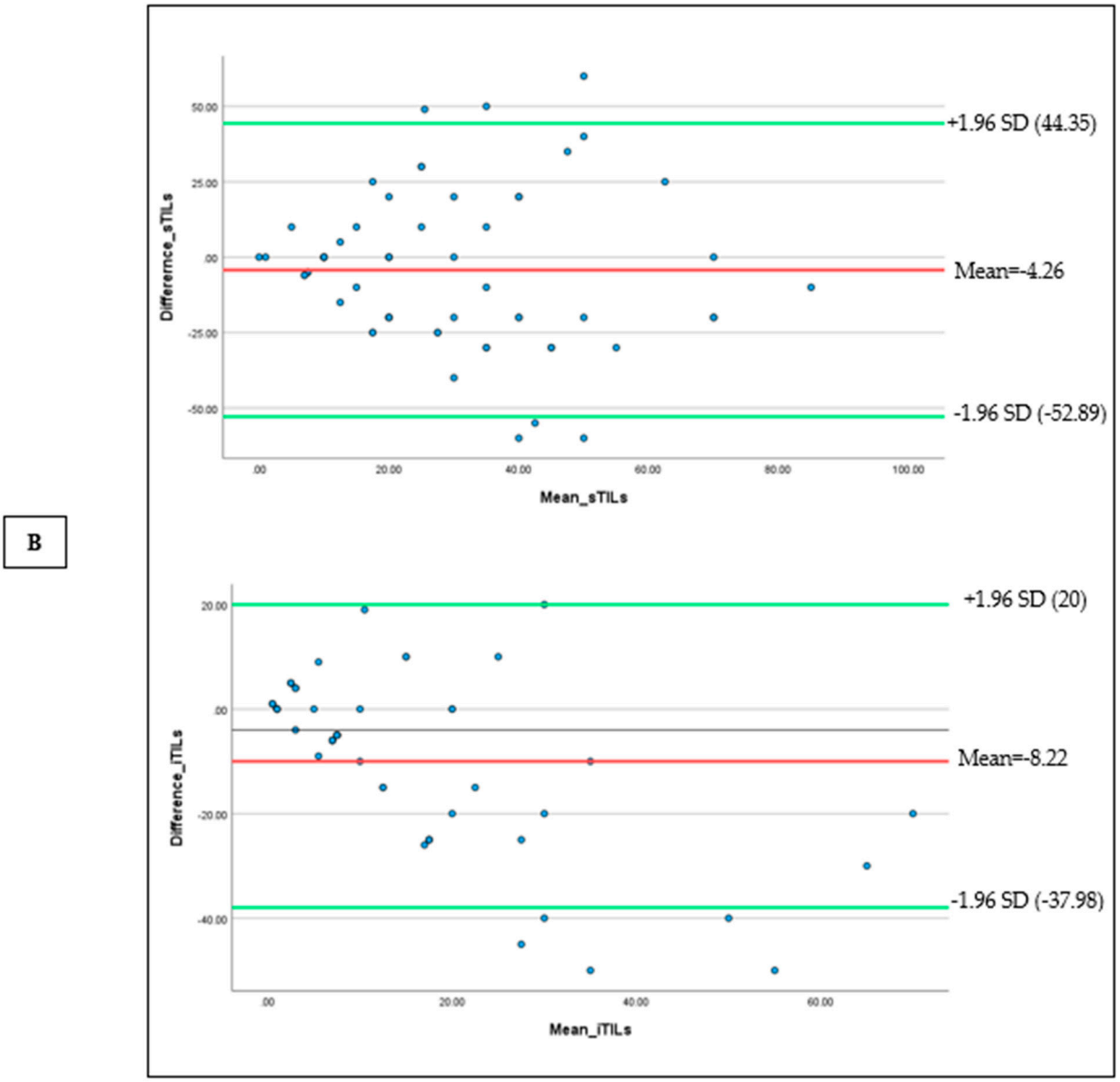

Figure 2. Bland–Altman plots were used to compare the sTIL (A) and iTILs (B) values of two observers (P1 and P2) to the median of their scores. The Y-axis represents the difference between the observer and the X-axis shows their mean values. The red line denotes the mean difference, and the green lines show the upper and lower limits of agreement.

A heat map comparing TIL density scores across all cases further illustrated the variability, particularly in cases with intermediate and high TIL density. Notable discrepancies were evident through colour shifts, especially in the yellow-to-red spectrum, denoting differences in lymphocyte density estimates between observers (Figure 3).

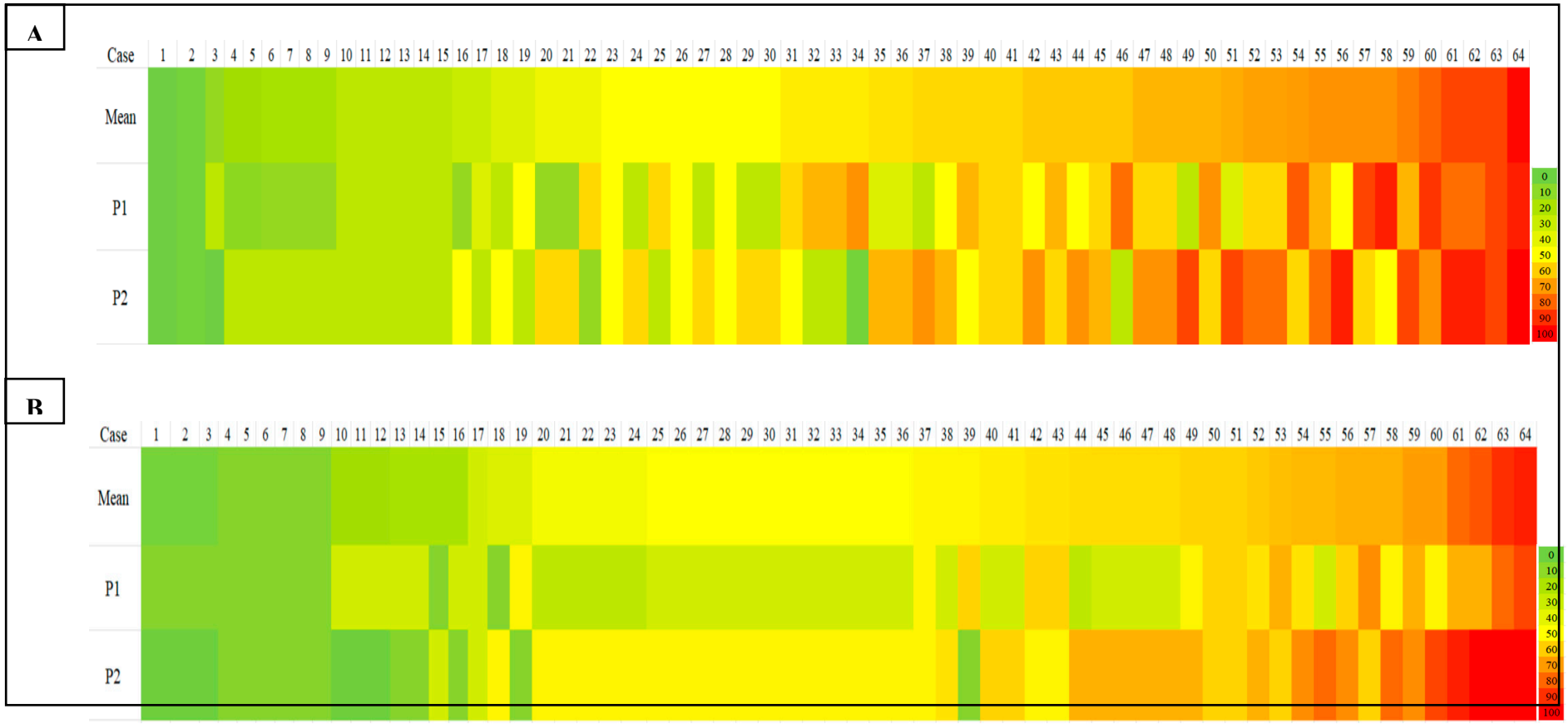

Figure 3. Heat map for scoring of sTILs (A) and iTILs (B) by two pathologists (P1 and P2). The first row shows the mean value for each case scored by both pathologists in an ascending manner from left to right; the second and third rows represent two observers' sTIL and iTILs records for the similar case. Color-coding indicates the percentage density of TILs quantified by each pathologist, from green (0%) to yellow (50%) and red (100%).

 doi:10.20944/preprints202507.0874.v1

Out of the 64 TNBC cases, 36 exhibited scoring discrepancies greater than 10% between P1 and P2. These cases were re-assessed by a third pathologist (P3). Consensus scoring significantly improved agreement metrics. For sTILs, ICC values increased to 0.70 (agreement) and 0.81 (consistency), while iTILs showed even higher values of 0.81 and 0.84, respectively (Table 4).

Table 4. Intraclass correlation coefficients for interobserver agreement in discrepant cases (P1, P2, and P3).

Reliability measure	sTILs ICC (95% CI)	iTILs ICC (95% CI)
ICC (agreement)	0.70	0.81
ICC (consistency)	0.81	0.84

Abbreviations: sTILs, stromal tumour-infiltrating lymphocytes; iTILs, intratumoral tumour-infiltrating lymphocytes; ICC, intraclass correlation coefficient; CI, confidence interval.

Pairwise Cohen's kappa (κ) coefficients analysis among all three pathologists revealed fair to substantial agreement (Table 5). For sTILs, the highest agreement occurred at the 50% cut-off (average κ = 0.67). For iTILs, the agreement was most robust at the 50% threshold (average κ = 0.74), indicating substantial agreement.

Table 5. Pairwise Cohen's kappa (κ) coefficient values for sTILs and iTILs at different cut-off thresholds among P1, P2, and P3.

Cut-off (%)	sTILs			iTILs			Interpretation
	P1 vs P3 (κ)	P2 vs P3 (κ)	Average (κ)	P1 vs P3 (κ)	P2 vs P3 (κ)	Average (κ)	
10	0.50	0.36	0.43	0.48	0.50	0.49	Fair agreement
20	0.44	0.35	0.40	0.62	0.27	0.45	Fair agreement
30	0.50	0.25	0.40	0.64	0.47	0.56	Fair agreement
40	0.50	0.50	0.50	0.65	0.64	0.65	Fair agreement
50	0.65	0.68	0.67*	1.0	0.48	0.74*	Substantial agreement

Abbreviations: κ, Cohen's kappa coefficient. Note: *Asterisks* (*) indicate the highest average agreement values for each category.

4. Discussion

4.1. Interobserver Agreement and Consensus Review for Manual TILs Assessment

Accurate and reproducible assessment of TILs is essential for their use as prognostic biomarkers in TNBC. In this study, we evaluated interobserver variability among pathologists in scoring sTILs and iTILs using both continuous and categorical metrics. Our findings demonstrated moderate agreement for sTILs (ICC = 0.57-0.58, $p < 0.001$) and good agreement for iTILs (ICC = 0.70-0.75, $p < 0.001$), aligning with previous studies that report higher reproducibility for iTILs due to their more well-defined localisation within tumour nests [28,29].

In contrast, the higher agreement seen for iTILs suggests that identifying lymphocytes within tumor cell nests is more straightforward and consistent across observers [18]. The reduced reproducibility in sTIL scoring likely reflects the interpretive challenges associated with assessing lymphocytes dispersed across variable stromal regions. This heterogeneity may lead to subjective variation in identifying representative fields, especially in areas with ambiguous tumour-stroma boundaries or dense focal infiltrates.

4.2. Agreement Across Categorical Cutoffs

When assessed using a categorical threshold, agreement levels decreased with increasing cut-off values. The highest kappa coefficients for sTILs and iTILs were observed at the 10% threshold,

 doi:10.20944/preprints202507.0874.v1

suggesting that pathologists can more reliably identify cases with low TIL density. Agreement declined notably at higher thresholds (≥30%), reflecting increased subjectivity in interpreting moderate to dense lymphocyte infiltration. These findings underscore the importance of standardizing cut-off selection, particularly when using TILs in clinical decision-making or trial stratification.

4.3. Assessment of Scoring Agreement Using Bland–Altman Analysis

The Bland–Altman plot for sTILs demonstrates a relatively wide spread between the upper and lower limits of agreement, indicating moderate to high variability in pathologist scoring, particularly in cases with high TIL density. This finding is consistent with previous studies that have reported greater subjectivity when evaluating densely packed or heterogeneously distributed lymphocytic infiltrates [21]. In contrast, the Bland–Altman plot for iTILs displays narrower limits of agreement, suggesting better interobserver consistency. Nonetheless, several data points still fall outside these limits, especially in the mid-to-high TIL range. This may be attributed to difficulties in delineating tumour borders or accurately identifying lymphocytes within epithelial nests, which can be less distinct than stromal regions. Overall, both plots reinforce that while the level of agreement is generally acceptable, notable variability persists in cases with intermediate to high TIL infiltration, underscoring the inherent challenges of manual TIL assessment in complex histological contexts.

4.4. Heatmap Visualization of Scoring Discrepancies Across TIL Density Levels

Our heatmap analysis revealed notable variability in TIL scoring, particularly in cases with intermediate to high TILs, for both sTILs and iTILs. Pronounced colour shifts between yellow, orange, and red across similar cases reflected inconsistencies in the density estimate between observers, highlighting the subjectivity inherent in manual evaluation. To address this, one promising strategy is the implementation of artificial intelligence (AI)- driven models for automated TIL (aTILs) quantification. A recent systematic review identified 27 studies employing such approaches in breast cancer, with the majority utilizing deep learning architectures, such as convolutional neural networks (CNNs) and fully convolutional networks (FCNs), for tasks like image segmentation and lymphocyte detection. These models were generally trained using pathologist-annotated ground truth datasets, with 58% of the studies reporting moderate to strong correlation (R = 0.6–0.98) between AI-generated outputs and manual TILs scores [30], underscoring their potential to enhance reproducibility and diagnostic precision.

Collectively, evidence from prior studies suggests that concordance between manual TIL (mTIL) assessment and automated (aTIL) scoring ranges from weak to strong, primarily influenced by the differences in algorithm design, training data, and study methodology. [4,5,31–38]. While these findings reinforce the growing potential of AI-based tools in pathology, they also highlight significant variability in performance. Notably, the automated models may struggle to replicate the nuanced judgment applied by experienced pathologists, particularly in histologically complex cases. This underscores the importance of ongoing validation, refinement, and clinical benchmarking of AI algorithms before their routine integration into diagnostic workflows can be justified.

4.5. Impact of Consensus Review on Scoring Consistency

Of the 64 cases assessed, 36 showed scoring discrepancies exceeding 10% between the two primary observers. Following adjudication by a third pathologist, interobserver agreement improved markedly, particularly for sTILs (ICC = 0.70-0.81) and iTILs (ICC = 0.81–0.84). This highlights the utility of structured consensus review in enhancing scoring reliability and reducing variability. Pairwise kappa analysis confirmed these improvements, with substantial agreement achieved in iTIL scoring at the 50% cut-off (κ = 1.00 between P1 and P3). Consensus review methodologies have proven valuable in diagnostic settings involving subjective assessments and may serve as a quality control measure in pathology workflows, especially in multicentre trials or AI training datasets [39].

 doi:10.20944/preprints202507.0874.v1

These improvements underscore the importance of incorporating consensus-based scoring into clinical and research workflows, particularly when addressing cases that exhibit significant interobserver variability.

4.6. Contributing Factors to Interobserver Discrepancies

To better understand the factors causing disagreements in scoring, the discrepant cases were reviewed in detail. Discussions with the participating pathologist were undertaken to gather expert perspectives on the potential histological factors underlying the lack of agreement. Discrepant cases highlight the challenges of interpreting histological features, including heterogeneous TIL distribution, a poorly defined tumor-stroma interface, and focal dense lymphoid infiltrates. Additionally, the presence of necrosis and immune cell mimics, such as apoptotic bodies or reactive stromal cells, may further interfere with accurate TIL identification and contribute to interobserver variability. Reactive plasma cells sometimes closely resembled tumor cells, leading to possible misinterpretation during assessment. Cases with extensive tumor necrosis made it challenging to distinguish between viable tumor tissue and the surrounding stroma, which in turn obscured the identification of infiltrating lymphocytes. These sources of error are inherently subjective and can mislead even experienced pathologists. In several cases, TILs were densely packed into small, focal areas, making it difficult to determine whether those regions accurately represented the overall immune response. This led to different observers choosing different regions for evaluation, which naturally contributed to scoring differences. These findings are consistent with previous reports indicating that tissue complexity, biological heterogeneity, and technical quality significantly influence TIL interpretation [21,40–42]. Such factors can lead to notable variability, even among experienced pathologists, particularly for ambiguous histological features. Structured training programs focused on these morphologic pitfalls, along with consensus guidelines on region-of-interest (ROI) selection, could help reduce this variability and improve scoring consistency.

In parallel, there is growing support for the integration of artificial intelligence (AI)-based tools, which offer the potential to improve consistency and objectivity in TIL quantification, especially in histologically complex or borderline cases. Despite advances in automated detection and multi-target segmentation, the clinical adoption of AI in pathology remains limited. Barriers include interobserver variability in annotated datasets, complex tissue morphology, inconsistent labelling standards, and the limited transparency of AI decision-making processes [43,44]. These challenges highlight the need for robust, interpretable, and scalable AI models that can be seamlessly embedded into real-world pathology workflows. This is particularly critical in TNBC, where accurate TIL quantification is essential for prognostic classification and treatment planning.

Although this study did not directly evaluate AI, the findings provide a valuable benchmark for future studies that aim to compare manual scoring with automated approaches. Moving forward, combining expert review with AI-driven support systems could be key to improving the reliability of TIL assessment in both clinical and research settings.

4.7. Limitations and Future Directions

This study has several limitations that warrant consideration. First, the sample size was relatively modest (n = 64), which may limit the generalizability of the findings to broader TNBC populations. Second, although the pathologists involved had comparable experience and were trained in standardized scoring protocols, inherent subjectivity in manual TIL assessment could still influence outcomes. Third, while interobserver variability was thoroughly examined, the study did not assess intra-observer consistency, which is also relevant for clinical reproducibility. Additionally, although the potential of AI-assisted TIL scoring was discussed, no automated tools were directly evaluated in this study. This limits our ability to draw empirical conclusions about the comparative performance of manual versus automated approaches. Finally, the absence of clinical outcome correlation (e.g., survival or treatment response) restricts interpretation of the prognostic relevance of the observed scoring discrepancies.

 doi:10.20944/preprints202507.0874.v1

While this study establishes the reliability and limitations of manual TIL scoring, future work should explore the integration of automated digital pathology tools for the quantification of TIL. Given the subjectivity and variability inherent to manual assessments, especially in morphologically complex regions, AI-based models offer an attractive pathway for reproducibility and scalability. Recent studies have demonstrated that deep learning approaches, CNNs and FCNs, can accurately segment lymphocytes and quantify TILs across large histological fields with minimal observer bias [35–38]. Tools such as QuPath, high-throughput analytics for learning and optimization (HALO) AI, and in-house trained pipelines have achieved moderate to strong correlation (r = 0.6–0.98) with pathologist-annotated ground truth. Incorporating such tools into our digitized TNBC slide dataset could provide valuable comparative insights into scoring consistency and highlight cases where AI either resolves or contributes to interobserver disagreement.

A potential analytical pipeline would involve running AI-based TIL quantification on the same set of whole-slide images and evaluating agreement metrics (e.g., ICC, Bland–Altman, and Spearman correlation) against consensus human scores. Additionally, discordant or outlier cases, particularly those resolved through adjudication by a third pathologist, could be re-evaluated to determine if AI models align more closely with consensus outcomes. This would serve to benchmark the practical utility of AI in reducing scoring bias and improving throughput in clinical settings.

Ultimately, validating AI outputs against high-fidelity consensus scores may enable the development of hybrid models, where human oversight is retained for ambiguous regions, and AI handles bulk quantification. Future efforts should focus on building explainable AI frameworks that incorporate histological context, handle variable TIL distributions, and provide confidence scores to guide clinical decision-making.

5. Conclusions

This study underscores the persistent interobserver variability in manual TIL assessment among pathologists, particularly for sTILs in TNBC cases. While moderate to substantial agreement was achieved, especially following consensus review, the findings reveal significant discrepancies in intermediate to high TIL density cases, driven by histological complexity and interpretive subjectivity. Such variability highlights the limitations of manual scoring even when standardized guidelines are applied. The results also emphasize the potential of integrating AI-assisted approaches to enhance the reproducibility and scalability of TIL quantification. As digital pathology continues to evolve, validated AI models can complement human expertise, particularly in challenging cases, to support consistent and high-throughput evaluation of TILs across clinical and research settings.

Future research should focus on cross-validating AI outputs with expert consensus scores, incorporating clinical outcome data to refine cut-offs, and developing hybrid workflows that leverage both computational precision and pathologist oversight. By bridging manual expertise with automated tools, the field can move closer to establishing TILs as reliable, standardized biomarkers in personalized breast cancer management.

Author Contributions: N.K.B.B., M.M. and M.A.H.Z. contributed to the study design, conducted the literature review, coordinated data collection, performed statistical analyses, and drafted the initial manuscript. R.R.M.Z. and A.A contributed to the literature review, assisted in the interpretation of interobserver variability findings, and critically reviewed the manuscript. M.M. and Q.X. supported manuscript review and contributed to the refinement of the discussion and conclusion. R.R.M.Z. and A.A. provided conceptual guidance and supervision, contributed to methodological development, and critically revised the manuscript for intellectual content. All authors have read and approved the final version of the manuscript.

Funding: This research was funded by the Fundamental grant, Faculty of Medicine, UKM (FF-2021-444).

Institutional Review Board Statement: The study was approved by the Ethics Committee of the Faculty of Medicine, UKM (UKM PPI/111/8/JEP-2021-724 on 14th October 2021).

Informed Consent Statement: Not applicable.

 doi:10.20944/preprints202507.0874.v1

Data Availability Statement: Upon request from the corresponding author on reasonable request.

Acknowledgments: We would like to express our heartfelt gratitude to The National University of Malaysia (UKM) for providing us with essential grant support and valuable opportunities to publish our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TIL	Tumour-Infiltrating Lymphocyte
TNBC	Triple-negative breast cancer
TIL-WG	International TILs Working Group
H&E	Haematoxylin & Eosin
ICC	Intraclass correlation coefficient (ICC)
AI	Artificial Intelligence
sTIL	Stromal TIL
HER-2	Human Epidermal Growth Factor Receptor 2
ROI	Region of interest
iTIL	Intratumoural TIL
HCTM	Hospital Canselor Tuanku Muhriz
FFPE	Formalin fixed paraffin- embedded
IHC	Immunohistochemistry
ER	Oestrogen receptor
PR	Progesterone receptor
FSIH	Fluorescence in situ hybridization
DDISH	Dual-colour dual hapten in situ hybridization
CNN	Convolutional neural network
FCN	Fully convolutional network
mTIL	Manual TIL
aTIL	Automated TIL
HALO	High-throughput analytics for learning and optimization

References

1. Badraoui R, Rebai T, Elkahoui S, Alreshidi M, Veettil VN, Noumi E, et al. Allium subhirsutum L. As a potential source of antioxidant and anticancer bioactive molecules: Hr-lcms phytochemical profiling, in vitro and in vivo pharmacological study. Antioxidants. 2020;9(10):1–22.
2. Yanchuan Zhang, Qinghua Li, Jie Lan, Guojing Xie, Guangjie Zhang, Junhao Cui, Ping Leng YW. Triple-negative breast cancer molecular subtypes and potential detection targets for biological therapy indications. Carcinogenesis. 2025;46(2).
3. Lotfinejad P, Jafarabadi MA, Shadbad MA, Kazemi T, Pashazadeh F, Shotorbani SS, et al. Prognostic role and clinical significance of tumor-infiltrating lymphocyte (TIL) and programmed death ligand 1 (PD-L1) Expression in Triple-Negative Breast Cancer (TNBC): A systematic review and meta-analysis study. Diagnostics. 2020;10(9):1–13.
4. Thagaard J, Stovgaard ES, Vognsen LG, Hauberg S, Dahl A, Ebstrup T, et al. Automated quantification of sTIL density with H&E-based digital image analysis has prognostic potential in triple-negative breast cancers. Cancers (Basel). 2021 Jun 2;13(12).

 doi:10.20944/preprints202507.0874.v1

5. Krijgsman D, Van Leeuwen MB, Van Der Ven J, Almeida V, Vlutters R, Halter D, et al. Quantitative whole slide assessment of tumor-infiltrating CD8-positive lymphocytes in ER-positive breast cancer in relation to clinical outcome. IEEE J Biomed Health Inform. 2021;25(2):381–92.
6. Bhattarai S, Saini G, Li H, Seth G, Fisher TB, Janssen EAM, et al. Predicting neoadjuvant treatment response in triple-negative breast cancer using machine learning. Diagnostics. 2024;14(1):1–13.
7. Fisher TB, Saini G, Rekha TS, Krishnamurthy J, Bhattarai S, Callagy G, et al. Digital image analysis and machine learning-assisted prediction of neoadjuvant chemotherapy response in triple-negative breast cancer. Breast Cancer Research. 2024;26(1):1–13. Available from: https://doi.org/10.1186/s13058-023-01752-y
8. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The evaluation of tumor-infiltrating lymphocytes (TILS) in breast cancer: Recommendations by an International TILS Working Group 2014. Vol. 26, Annals of Oncology. Oxford University Press; 2015. p. 259–71.
9. Wu R, Oshi M, Asaoka M, Yan L, Benesch MGK, Khoury T, et al. Intratumoral tumor infiltrating lymphocytes (TILs) are associated with cell proliferation and better survival but not always with chemotherapy response in breast cancer. Ann Surg. 2023;278(4):587–97.
10. Balic M, Thomssen C, Würstlein R, Gnant M, Harbeck N. St. Gallen/Vienna 2019: A brief summary of the consensus discussion on the optimal primary breast cancer treatment. Breast Care. 2019;14(2):103–10.
11. Azar, Kazemi., Masoumeh, Gharib., Nema, Mohamadian, Roshan., Shirin, Taraz, Jamshidi., Fabian, Stögbauer., Saeid, Eslami., Peter, J. Schüffler. Assessment of the tumor–stroma ratio and tumor-infiltrating lymphocytes in colorectal cancer: inter-observer agreement evaluation. Diagnostics. 2023;
12. Frances Reznitsky, J. D. Jensen, A. Knoop AL. Inter-observer agreement of tumor infiltrating lymphocytes in primary HER2-positive breast cancer and correlation between tissue microarray and full tumor-sections. Acta Pathologica, Microbiologica et Immunologica Scandinavica (APMIS). 2022;
13. Fatmagül, Kuşku, Çabuk., Fatma, Aktepe., Fatma, Nilgun, Kapucuoglu., Ipek, Coban., Dauren, Sarsenov., Vahit O. Interobserver reproducibility of tumor-infiltrating lymphocyte evaluations in breast cancer. Indian Journal of Pathology & Microbiology. Indian J Pathol Microbiol. 2018;
14. Khoury T, Peng X, Yan L, Wang D, Nagrale V. Tumor-infiltrating lymphocytes in breast cancer: Evaluating interobserver variability, heterogeneity, and fidelity of scoring core biopsies. Am J Clin Pathol. 2018;150(5):441–50.
15. Van Bockstal MR, Cooks M, Nederlof I, Brinkhuis M, Dutman A, Koopmans M, et al. Interobserver agreement of pd-l1/sp142 immunohistochemistry and tumor-infiltrating lymphocytes (Tils) in distant metastases of triple-negative breast cancer: A proof-of-concept study. a report on behalf of the international immuno-oncology biomarker worki. Cancers (Basel). 2021;13(19):1–20.
16. Altinay S, Arnould L, Bletard N, Colpaert C, Dedeurwaerdere F, Dessauvagie B, et al. HHS Public Access. 2021;33(3):354–66.
17. Cserni B, Kilmartin D, O'Loughlin M, Andreu X, Bagó-Horváth Z, Bianchi S, et al. ONEST (Observers Needed to Evaluate Subjective Tests) analysis of stromal tumour-infiltrating lymphocytes (sTILs) in breast cancer and its limitations. Cancers (Basel). 2023;15(4).
18. Capar A, Ekinci DA, Ertano M, Niazi MKK, Balaban EB, Aloglu I, et al. An interpretable framework for inter-observer agreement measurements in TILs scoring on histopathological breast images: A proof-of-principle study. PLoS One [Internet]. 2024;19(12 December):1–21. Available from: http://dx.doi.org/10.1371/journal.pone.0314450
19. Dano H, Altinay S, Arnould L, Bletard N, Colpaert C, Dedeurwaerdere F, et al. Interobserver variability in upfront dichotomous histopathological assessment of ductal carcinoma in situ of the breast: the DCISion study. Modern Pathology. 2020;33(3):354–66.
20. Kilmartin D, O'Loughlin M, Andreu X, Bagó-Horváth Z, Bianchi S, Chmielik E, et al. Intra-tumour heterogeneity is one of the main sources of inter-observer variation in scoring stromal tumour infiltrating lymphocytes in triple negative breast cancer. Cancers (Basel). 2021;13(17):1–16.
21. Choi S, Cho SI, Jung W, Lee T, Choi SJ, Song S, et al. Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer. NPJ Breast Cancer. 2023;9(1):1–13.

 doi:10.20944/preprints202507.0874.v1

22. Jeppe Thagaard 1 2, Glenn Broeckx 3 4, David B Page 5, Chowdhury Arif Jahangir 6, Sara Verbandt 7, Zuzana Kos 8, Rajarsi Gupta 9, Reena Khiroya 10, Khalid Abduljabbar 11, Gabriela Acosta Haab 12, Balazs Acs 13 14, Guray Akturk 15, Jonas S Almeida 16, Isab ESS 22 152. Pitfalls in machine learning-based assessment of tumor-infiltrating lymphocytes in breast cancer: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. J Pathol. 2023;260(5):498–513.
23. Locy H, Verhulst S, Cools W, Waelput W, Brock S, Cras L, et al. Assessing tumor-infiltrating lymphocytes in breast cancer: a proposal for combining immunohistochemistry and gene expression analysis to refine scoring. Front Immunol. 2022;13(February):1–13.
24. Arian Arab, Victor Garcia, Seyed Mostafa Kahaki, Nicholas Petrick, Brandon D. Gallas WC. Assessment of machine learning algorithms for TILs scoring using whole slide images: comparison with pathologists. Digital and Computational Pathology. 2024;
25. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. Journal of Clinical Oncology. 2020;38(12):1346–66.
26. Landis JR, Koch GG. Landis amd Koch1977_agreement of categorical data. Biometrics. 1977;33(1):159–74.
27. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–63. Available from: http://dx.doi.org/10.1016/j.jcm.2016.02.012
28. Fatmagül, Kuşku, Çabuk., Fatma, Aktepe., Fatma, Nilgun, Kapucuoglu., Ipek, Coban., Dauren, Sarsenov., Vahit O. Interobserver reproducibility of tumor-infiltrating lymphocyte evaluations in breast cancer. Indian Journal of Pathology & Microbiology. Indian J Pathol Microbiol. 2018;
29. Swisher SK, Wu Y, Castaneda CA, Lyons GR, Yang F, Tapia C, et al. Interobserver agreement between pathologists assessing tumor-infiltrating lymphocytes (TILs) in breast cancer using methodology proposed by the International TILs Working Group. Ann Surg Oncol. 2016;23(7):2242–8.
30. Baharun NB, Adam A, Zailani MAH, Rajpoot NM, Xu Q, Zin RRM. Automated scoring methods for quantitative interpretation of Tumour infiltrating lymphocytes (TILs) in breast cancer: a systematic review. BMC Cancer [Internet]. 2024;24(1):1202. Available from: https://doi.org/10.1186/s12885-024-12962-8
31. Heindl A, Sestak I, Naidoo K, Cuzick J, Dowsett M, Yuan Y. Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. J Natl Cancer Inst. 2018 Feb 1;110(2):166–75.
32. Amgad M, Sarkar A, Srinivas C, Redman R, Ratra S, Bechert CJ, et al. Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. In SPIE-Intl Soc Optical Eng; 2019. p. 20.
33. Swiderska-Chadaj Z, Pinckaers H, van Rijthoven M, Balkenhol M, Melnikova M, Geessink O, et al. Learning to detect lymphocytes in immunohistochemistry with deep learning. Med Image Anal [Internet]. 2019; 58:101547. Available from: https://doi.org/10.1016/j.media.2019.101547
34. Entenberg D, Oktay MH, D'alfonso T, Ginter PS, Robinson BD, Xue X, et al. Validation of an automated quantitative digital pathology approach for scoring TMEM, a prognostic biomarker for metastasis. Cancers (Basel). 2020;12(4).
35. Yosofvand M, Khan SY, Dhakal R, Nejat A, Moustaid-Moussa N, Rahman RL, et al. Automated detection and scoring of tumor-infiltrating lymphocytes in breast cancer histopathology slides. Cancers (Basel). 2023;15(14).
36. Choi S, Cho SI, Jung W, Lee T, Choi SJ, Song S, et al. Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer. NPJ Breast Cancer. 2023;9(1):1–13.
37. Rong R, Sheng H, Jin KW, Wu F, Luo D, Wen Z, et al. A deep learning approach for histology-based nucleus segmentation and tumor microenvironment characterization. Mod Pathol [Internet]. 2023;36(8):100196. Available from: https://doi.org/10.1016/j.modpat.2023.100196
38. Makhlouf S, Wahab N, Toss M, Ibrahim A, Lashen AG, Atallah NM, et al. Evaluation of tumour infiltrating lymphocytes in luminal breast cancer using artificial intelligence. Br J Cancer. 2023;(September).
39. Jahan S, Al-saigul AM, Abdelgadir MH. Tumor-infiltrating lymphocyte scoring in neoadjuvant-treated breast cancer. Cancers (Basel). 2024;16.

 doi:10.20944/preprints202507.0874.v1

40. Eunkyung Han, Hye Yeon Choi, Hyun Jung Kwon, Yul Ri Chung, Hee-Chul Shin, Eun-Kyu Kim, Koung Jin Suh, Se Hyun Kim JHK& SYP. Characterization of tumor-infiltrating lymphocytes and their spatial distribution in triple-negative breast cancer. Breast Cancer Research. 2024;
41. Danielle J. Fassler 1,† , Luke A. Torre-Healy 1,† RG 1, 2 AMH, 1 SK, 2 SCVA, 1 YZ, 1 TK, et al. Spatial characterization of tumor-infiltrating lymphocytes and breast cancer progression. Cancers (Basel). 2022;
42. Kilmartin D, O'Loughlin M, Andreu X, Bagó-Horváth Z, Bianchi S, Chmielik E, et al. Intra-tumour heterogeneity is one of the main sources of inter-observer variation in scoring stromal tumour infiltrating lymphocytes in triple negative breast cancer. Cancers (Basel). 2021;13(17):1–16.
43. Xu Q, Adam A, Abdullah A, Bariyah N. A Review of Advanced Deep Learning Methods of Multi-target segmentation for breast cancer WSIs. IEEE Access. 2025;13(May):76016–37.
44. Xu Q, Adam A, Abdullah A, Bariyah N. Advanced deep learning approaches in detection technologies for comprehensive breast cancer assessment based on WSIs: A systematic literature review. Diagnostics. 2025;15(9):1–24.