

Validation of a Label-Efficient Computational Model for Tumour-Infiltrating Lymphocyte Assessment in Breast Cancer: A Multicentre Study of 2,340 Patients

Wei Zhang^{1*}, Chen Hui², Michael Tan¹

¹ Department of Breast Pathology and Computational Pathology, Faculty of Medicine, University of Melbourne, Melbourne, Australia.

² Department of Tumour Immunology and AI Models, Faculty of Medicine, National University of Singapore, Singapore.

*E-mail ✉ wei.zhang@unimelb.edu.au

Abstract

The number of stromal tumor-infiltrating lymphocytes (TILs) is a significant prognostic indicator in triple-negative breast cancer and reflects the immune system's response to the tumor. Automated computational TIL evaluation helps pathologists with a tedious task and delivers fast, consistent results. Yet most existing computational models for TILs depend on extensive, detailed labeling and rely on elaborate deep learning systems, which hinder iterative improvements and real-world clinical application. We present and test a much simpler deep learning solution trained in merely 10 minutes on far fewer — specifically 100 times fewer — pathologist labels. Whole slide images (WSIs) paired with TIL scores and clinical details were obtained for 2340 breast cancer patients, 790 of whom had triple-negative disease. These came from three separate cohorts in three nations (USA, UK, Netherlands) plus three randomized clinical trials from the Netherlands. Feature extraction on the WSIs used a pathology foundation model to capture morphological information. The new model, termed label-efficient computational stromal TIL assessment (ECTIL), predicts the overall WSI TIL score straight from these features. Training variants included one cohort from The Cancer Genome Atlas ($n = 356$, ECTIL-TCGA), only triple-negative samples from four cohorts ($n = 400$, ECTIL-TNBC), and all molecular subtypes across five cohorts ($n = 196464$, ECTIL-combined). We quantified concordance with pathologist scores using Pearson's correlation coefficient (r). AUROC values were computed after splitting pathologist TIL scores into clinically relevant TILs-high ($\geq 30\%$) versus TILs-low ($< 30\%$) categories. Multivariate Cox regression in the PARADIGM cohort with full clinicopathological information ($n = 384$) estimated hazard ratios for overall survival, adjusted for other clinicopathological factors. ECTIL-TCGA-matched pathologist evaluations were well correlated with five heterogeneous external groups ($r = 0.54\text{--}0.74$; AUROC = $0.80\text{--}0.94$). ECTIL-TNBC outperformed ECTIL-TCGA on the PARADIGM set ($r = 0.64$, AUROC = 0.83 compared with $r = 0.58$, AUROC = 0.80). ECTIL-combined delivered the strongest results on an independent external test collection ($r = 0.69$, AUROC = 0.85). In multivariate Cox models, each 10% rise in ECTIL-combined TIL scores was linked to better overall survival (hazard ratio 0.85 , 95% CI: $0.77\text{--}0.93$; $P = 0.0007$), remaining independent of clinicopathological factors and closely matching the pathologist-derived score (0.86 , $0.81\text{--}0.92$; $P < 0.0001$). This investigation shows that ECTIL successfully assigns TIL scores to hematoxylin and eosin-stained, formalin-fixed, paraffin-embedded WSIs in a single step and achieves high agreement with experienced pathologists. By avoiding deep learning segmentation or detection pipelines entirely, ECTIL still yielded survival hazard ratios equivalent to manual pathologist scoring in analyses independent of clinicopathological variables. Such computational TIL tools may eventually support patient pre-screening for future de-escalation trials in triple-negative breast cancer or help pathologists and doctors during breast cancer diagnostic assessments. The model is openly accessible online under an open-source license, allowing researchers to test and apply it in future work on breast cancer or other tumor types.

Keywords: Tumor-infiltrating lymphocytes, Breast cancer, Triple-negative breast cancer, Computational pathology, Whole-slide imaging, Deep learning

Access this article online

<https://smerpub.com/>

Received: 17 September 2024; Accepted: 29 November 2024

Copyright CC BY-NC-SA 4.0

How to cite this article: Zhang W, Hui C, Tan M. Validation of a Label-Efficient Computational Model for Tumour-Infiltrating Lymphocyte Assessment in Breast Cancer: A Multicentre Study of 2,340 Patients. *J Med Sci Interdiscip Res.* 2024;4(2):167-82. <https://doi.org/10.51847/Q9Wh3RZHA1>

Introduction

Breast cancer is the second most frequent reason for cancer deaths in women internationally [1]. Studies of tumor-immune dynamics have identified stromal tumor-infiltrating lymphocytes (TILs) as a crucial prognostic and predictive factor, especially in HER2-positive and

triple-negative subtypes [2]. In triple-negative breast cancer, TIL counts correlate with rates of pathological complete response, disease-free survival, and overall survival [2-4]. High baseline TIL levels (greater than 50% for patients over 40 years and 75% or above for those under 40) can indicate very favorable overall survival in stage I triple-negative breast cancer without any adjuvant or neoadjuvant chemotherapy [5, 6].

Research in context

Evidence before this study

We queried Google Scholar and PubMed on Nov 1, 2023, with the search string ("Deep Learning" OR "Artificial Intelligence") AND "Pathology" AND ("lymphocytes" OR "stromal tumor infiltrating lymphocytes" OR "TILs") AND "breast cancer", inspecting the initial 30 hits. No limits were set on language or publication date. Included were studies validating deep learning for stromal TIL quantification on H&E-stained WSIs from invasive breast cancer (n=11) and relevant reviews (n=1). These either benchmarked automated TIL measures against pathologist readings or developed fresh lymphocyte-related biomarkers to boost prognostic or predictive accuracy. Published TIL scores aligned reasonably with pathologists (Spearman's rank correlations of 0.57 to 0.79 depending on study and cohort) and showed independent prognostic benefit. However, their overall prognostic strength typically fell short of that of a pathologist. Every externally tested method involved multi-step pipelines that performed segmentation and detection to aggregate pixel information into the final biomarker score. Training required thousands of expensive tissue- and cell-level expert pathologist annotations.

Added value of this study

This appears to be the initial external validation of a computational TIL model focused on early triple-negative breast cancer that fundamentally departs from prior designs by eliminating the detection and segmentation deep learning stages. Training uses only WSI-level TIL scores instead of pixel-wise semantic segmentations or individual cell labels. It uses only a few hundred pathologist annotations—a hundredfold fewer than earlier techniques that required thousands. Despite minimal labels, its basic structure and total lack of segmentation masks or cell detections, the model achieved similar agreement levels (Spearman's rank correlations 0.47–0.72 on external cohorts) and retained independent prognostic significance in young patients (<

40 years) with node-negative triple-negative breast cancer who avoided neoadjuvant or adjuvant treatment. This represents the most extensive retrospective validation effort (2340 patients, 2354 slides, six cohorts) for any computational TIL method so far. Performance remained stable across clinical characteristics, molecular subtypes, institutions, and specimen type (biopsy versus resection).

Implications of all the available evidence

Results indicate that a straightforward, annotation-light model can closely match pathologists and maintain independent prognostic power on par with complex multi-step segmentation–detection systems, while requiring far fewer labels. This challenges the assumption that heavy data requirements and detection/segmentation steps are essential for computational TIL scoring. The approach mitigates several existing shortcomings, with outstanding issues fixable with a small number of additional labels and minor design tweaks. Prospective testing of the model is planned in a phase 3 de-escalation trial for low-stage triple-negative breast cancer.

Pathologists score TILs by direct microscopic review of H&E-stained FFPE WSIs. The score quantifies the stromal fraction occupied by mononuclear inflammatory cells inside the invasive tumor edge per International Immuno-Oncology Biomarker Working Group on Breast Cancer (TILs Working Group) standards [7]. Extra training has boosted consistency, but observer variability persists, partly due to tumor heterogeneity. Kos *et al.* [8] reported intraclass correlation coefficients ranging from 0.77 to 0.94 across three groups.

Advances in AI and deep learning over the last 10 years, alongside the digitization of pathology slides into WSIs, have spurred the development of computational TIL tools. The TILs Working Group also provided computational recommendations to address the full scoring complexity, facilitating potential clinical uptake of such systems [9]. AI assistance for TIL assessment might also ease some pathologist burdens, including workload pressures on pathology services.

Typical current models produce detailed pixel-level segmentations, cell detections, and classifications [10-13]. While they correlate strongly with pathologist TIL scores and are associated with better survival, they encounter various technical and guideline-related obstacles, as noted by Thagaard *et al.* [14]. Common problems include handling slide artifacts and strictly

following pathologist scoring rules [9]. Large, labeled datasets are necessary for training, which increases costs and effort. Greater complexity for pixel-level outputs makes comprehensive clinical validation of every submodule harder. Even with explicit guideline implementation and cell localization, discordant results still arise, e.g., due to analysis of the wrong tissue zones or cells.

We introduce and thoroughly validate ECTIL, a computational TIL assessment model with a distinct architecture from earlier publications. Instead of pixel-level cell segmentation within the tumor stroma or direct guideline enforcement, it efficiently predicts the full WSI TIL score in a single pass from features generated by an open-source foundation model on the WSI, without additional annotations or manual post-processing rules. Thus, ECTIL trains solely on whole-slide TIL scores. Building on earlier findings, it refines a prior WSI-level TIL regression framework by providing a stronger foundation model, gated attention, and an additional projection head for feature tuning [15]. Only several hundred WSI-level TIL scores suffice for training, making ECTIL a practical, quick-to-deploy solution for accurate TIL quantification on H&E WSIs (biopsies or resections from different scanners) across breast cancer molecular and histologic variants. Validation was performed on the largest computational TIL dataset assembled to date, with direct comparison to an established open-source model using traditional cell segmentation. ECTIL can streamline TIL scoring in breast cancer, advance biomarker use in research, and ease clinical integration. It is also scheduled for inclusion in a prospective phase 3 de-escalation trial for early-stage triple-negative breast cancer (NCT06476119).

Materials and Methods

Study design and cohorts

The Institutional Review Board approved the present investigation and associated data collection at the Netherlands Cancer Institute. Participants enrolled in the clinical trials gave written informed consent covering both the trial interventions and subsequent use of their data for additional translational investigations. The PARADIGM project received ethics approval from the Netherlands Cancer Institute (NKI) ethics committee and the NKI translational research board (CFMPB173). Every cohort followed Good Clinical Practice standards and the Declaration of Helsinki.

Clinical, histopathological, and WSI data from primary tumors with existing pathologist-assigned TIL scores were gathered across six groups of early breast cancer patients. No specific eligibility rules were applied to the chosen cohorts. Two groups were drawn from publicly accessible sources. The Cancer Genome Atlas (TCGA) is a comprehensive cancer genomics resource that has molecularly profiled primary tumors from multiple cancer types, including breast cancer (n=1134). Relevant clinical information was retrieved from the Genomic Data Commons repository. The other public group came from the Breast Cancer Somatic Genetics Study (BASIS, n=560), as reported by Nik-Zainal *et al.* [15]. TCGA patients were treated in the USA, while BASIS represented a UK-led international effort [15]. Three additional groups involved patients from randomized clinical trials performed in the Netherlands (N4+ [16], n=855; MATADOR [17], n=664; and TRAIN II [18], n=438).

The PARADIGM cohort served exclusively as an external validation set. It included retrospectively assembled clinical records and primary tumor samples from patients under 40 years old diagnosed with triple-negative breast cancer (any T stage, node-negative, metastasis-free) who received no systemic treatment and were managed between 1989 and 2000 in the Netherlands within the PARADIGM initiative [19].

Study design and modeling choices were established in advance and remained unchanged, although no formal protocol document was created. The study was not registered. No patients or members of the public participated in study planning, execution, analysis, interpretation, or result sharing.

Data extraction

The clinical variables needed for the analysis were provided directly by each study's principal investigator. Tumor grading followed the Bloom-Richardson system [20]. Lymphovascular invasion status, histological type, and molecular subtype (based on hormone receptor and HER2 status) were evaluated either locally, per national standards [21], or centrally, per trial protocol. Ethnicity information was unavailable owing to privacy regulations in the Netherlands.

TIL scoring on H&E-stained FFPE WSIs adhered to established recommendations from the TILs Working Group [7]. All scores were assigned retrospectively. TCGA, BASIS, and PARADIGM slides were evaluated by one pathologist (RS). N4+ and MATADOR slides

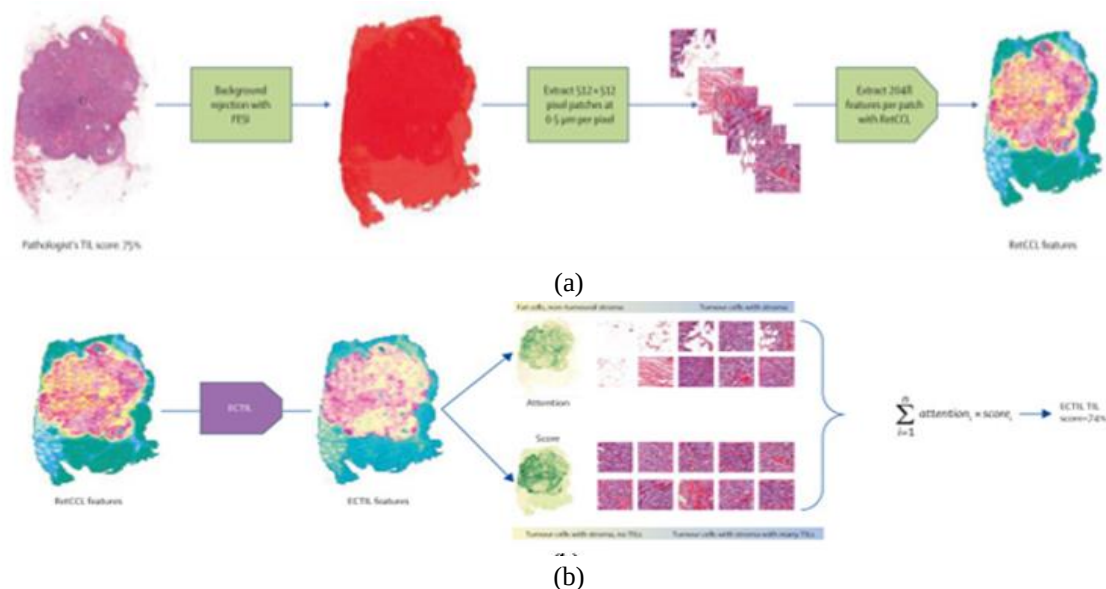
were scored by pathologist HMH and by MO under HMH's supervision, given their demonstrated strong interobserver agreement. For TRAIN II, the scoring procedure was previously outlined [22]. Briefly, each slide received independent scores from RS and another pathologist (HMH or MO). When the absolute difference exceeded 10%, the slide underwent joint review, and RS provided the definitive score. When the difference was 10% or smaller, geometric means were computed after adding 0.5 to any zero values, consistent with the published TRAIN II protocol. Interobserver reliability was high between RS and HMH, as well as between RS and MvS. In some patients, two WSIs were present, but only one TIL score was recorded. For TCGA, we retained both slides from 14 patients with the shared label after manual confirmation that the slides differed, remained high quality, and represented the given score. In all other cohorts, the single highest-quality WSI was selected, as judged by the responsible pathologist. TILs were assessed on primary tumors excised surgically before any neoadjuvant therapy, except in TRAIN II where diagnostic biopsy material was used.

WSIs were included in the analysis only when both a digitized slide and a corresponding pathologist TIL score were already available, and when matching clinical data for that patient existed. This criterion led to exclusion of some patients in certain cohorts. No additional patient or slide selection occurred. Scanning was performed on

various devices: Leica Biosystems Aperio (all TCGA samples and 425 N4+ samples), Hamamatsu C9600-12 (BASIS), and 3DHistech P1000 (PARADIGM, MATADOR, TRAIN II, and 127 N4+ samples). Fourteen TCGA patients contributed two representative WSIs each, both of which were incorporated into the ECTIL models. In total, 5799 WSIs were collected across the six cohorts. Pathologist TIL scores existed for 2480 patients. Complete clinical data were available for 2340 patients, yielding 2354 slides suitable for concordance evaluations. Within PARADIGM, 476 WSIs corresponded to 390 patients. For univariate and multivariate survival analyses in this cohort, 384 patients were retained after excluding 6 patients with incomplete clinical information (5 lacking T-stage data and 1 lacking lymphovascular invasion data). Individuals missing clinical variables were omitted from analyses.

Deep learning model: ECTIL

All components of the ECTIL pipeline are illustrated in **Figure 1**. Non-tissue areas were first identified and removed using the Foreground Extraction from Structure Information algorithm from the open-source Deep Learning Utilities for Pathology library (version 0.3.22), developed at the Netherlands Cancer Institute [23, 24]. Tiles containing no foreground pixels (i.e., no tissue) were discarded.



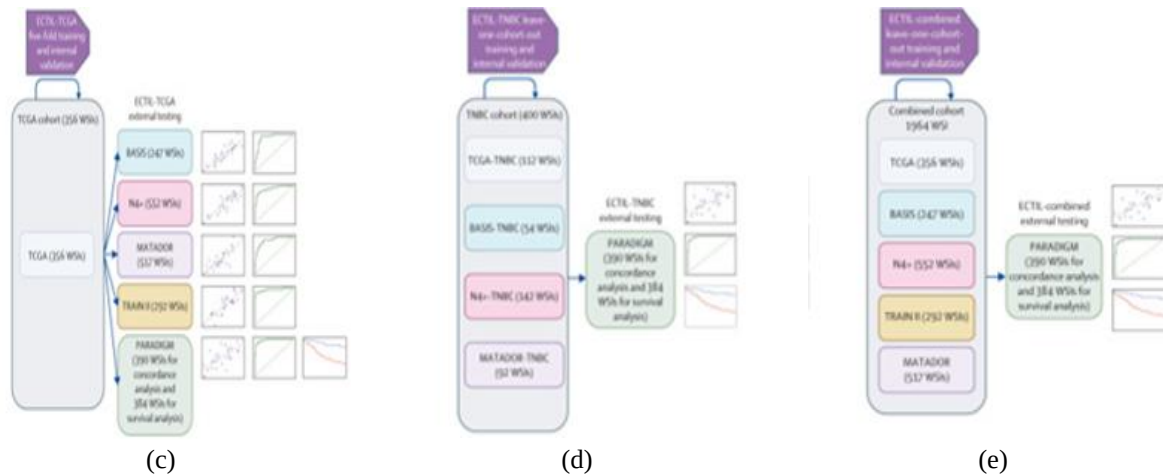


Figure 1. Overview diagram showing the complete pipeline including data preparation stages, model training procedures, internal validation steps, external evaluation arrangements, and the one-pass inference workflow used by the ECTIL system

(a) Preparation of input data through publicly released software tools; (b) Direct prediction of TIL score by ECTIL. In the inference example, RetCCL features from a WSI (pathologist reference TIL score of 75%) were projected into a 512-dimensional embedding space to derive per-tile attention weights and scores. Heatmaps visualize attention and score distributions for the sample; randomly chosen patches across the range are highlighted with short content descriptions. The ultimate prediction equals the attention-weighted average across all tiles, yielding 74% in this instance — nearly identical to the reference; (c) Training/validation/testing scheme applied for ECTIL-TCGA; (d) Training/validation/testing scheme applied for ECTIL-TNBC; (e) Training/validation/testing scheme applied for ECTIL-combined. All plots (scatter, ROC, survival) rely on artificial data created only for demonstration. Abbreviations: ECTIL = label-efficient computational stromal TIL assessment; FESI = Foreground Extraction from Structure Information; TCGA = The Cancer Genome Atlas; TIL = tumor-infiltrating lymphocyte; TNBC = triple-negative breast cancer; WSI = whole slide image.

The feature extraction step relied on RetCCL, a freely available convolutional neural network pre-trained via self-supervised methods on histopathological images by Wang *et al.* [25]. It generated 2048-dimensional vectors from non-overlapping 512×512 -pixel regions at $0.5 \mu\text{m}$ per pixel magnification, after applying ImageNet normalization statistics (**Figure 1**) [26]. Selection of RetCCL was based on its leading results at the time, as

documented by Niehues *et al.* [27]. Training of RetCCL was confined to TCGA slides alone, thereby eliminating any risk of information leakage into the independent external test groups examined here.

ECTIL employs a gated attention multiple instance learning framework to predict TIL scores directly from patch-level features, drawing on prior methodological foundations [28, 29]. Following the base feature vectors, a non-linear projection incorporating rectified linear unit activation created adapted tile representations, and a final sigmoid function restricted outputs to the interval $[0,1]$. Comparison was made with EasTIL, an available open-source pipeline for cell segmentation that had previously been evaluated as a TIL assessment tool [30]. Tumor bed region annotations for EasTIL were collected from a subset of PARADIGM material.

Training of ECTIL-TCGA used the TCGA collection (356 slides) and employed five-fold cross-validation, with each fold separated by center. Training was performed in three folds, with one fold used for monitoring during learning and the last for internal evaluation (**Figure 1**). External performance checks then covered the other cohorts. ECTIL-combined training involved all cohorts except PARADIGM (1964 slides) in a five-fold leave-one-cohort-out scheme, holding out one cohort per fold for validation, with final testing restricted to PARADIGM. ECTIL-TNBC training focused exclusively on triple-negative cases drawn from TCGA, BASIS, N4+, and MATADOR (400 slides) using four-fold leave-one-cohort-out splitting, reserving one cohort for internal validation each time, and testing solely on

PARADIGM. For each external test, we selected the best-performing model per validation fold (highest Pearson's r) and deployed an ensemble of these on the held-out data. EasTIL operated with a pre-trained segmentation network and its released public weights.

Statistical analysis

Agreement metrics included mean squared error, Pearson's correlation, Spearman's rank correlation, and concordance correlation coefficient, all computed in Python via torchmetrics [31]. For binary classification performance distinguishing high versus low TIL categories, we used pathologist scores divided at $< 30\%$ (low) and $\geq 30\%$ (high), acknowledging the lack of a universally accepted threshold. AUROC was derived from these binary labels against continuous model outputs. Additional metrics of average precision and AUROC were generated for alternative thresholds (10%, 50%, 75%). Visual summaries of predictions involved scatter diagrams, calibration graphs, and attention/score heatmaps at both dataset and individual slide levels.

Survival modeling followed the strategy published by de Jong *et al.* [5] and was implemented in R using the survival package on the PARADIGM data. Effects of pathologist-assigned and ECTIL-generated TIL scores on overall survival were quantified through hazard ratios (HRs) and concordance indices in univariate and multivariate Cox models, incorporating adjustments for key clinicopathological factors (T stage, grade, histology type, lymphovascular invasion, and local treatment). Model scores from ECTIL and EasTIL were normalized to the range observed in their respective training data. Kaplan–Meier survival curves for pathologist scores employed established cut points (30% and 75%), previously linked to prognostic significance in node-negative triple-negative breast cancer among patients

younger than 40 years [5]. Group differences were tested with log-rank statistics through the ggsurvplot R package. Direct comparison between pathologist and ECTIL scores used median-based dichotomization for each, followed by survival curve comparison between resulting subgroups. Proportionality of hazards was verified using Schoenfeld residuals.

Role of the funding source

Funding bodies played no part in study planning, data gathering, statistical processing, result interpretation, or report drafting.

Results and Discussion

Table 1 presents patient and disease features at baseline for the full cohort, covering 2340 cases with both digitized WSIs and comprehensive clinical records. Median age, excluding the PARADIGM group (corresponding to the ECTIL-combined column), stood at 49 years (IQR: 43–56). Outside PARADIGM, 1943 (99%) of the 1964 patients were women. Disease stage distribution showed that 842 (43%) of 1964 patients were stage II and 864 (44%) were stage III at diagnosis or enrolment. Known stage IV disease at entry affected 12 (1%) of 1964 cases. The predominant histology was invasive breast carcinoma of no special type (1308 [67%] of 1964), with grade II tumors in 655 (33%) and grade III in 675 (34%) of 1964 patients. Lymphovascular invasion information was missing for 1742 (89%) of 1964 tumors. Receptor status breakdown indicated 1175 (60%) ER-positive HER2-negative, 230 (12%) ER-positive HER2-positive, 123 (6%) ER-negative HER2-positive, and 400 (20%) triple-negative tumors. The median value for pathologist-determined TIL score was 10% (IQR: 5–25).

Table 1. Baseline characteristics of all clinical cohorts

Empty cell	TCGA (n = 356*)	BASIS (n = 247)	N4+ (n = 552)	MATADOR (n = 517)	TRAIN II (n = 292)	PARADIGM (n = 390)
Age, years						
Median (IQR)	50.0 (43.0–62.0)	57.0 (48.0–69.0)	46.0 (40.0–49.0)	51.0 (46.0–57.0)	49.0 (43.0–57.0)	35.0 (32.0–38.0)
Missing	14 (4%)*	0	0	0	0	0
Gender						
Female	338 (95%)	244 (99%)	552 (100%)	517 (100%)	292 (100%)	390 (100%)
Male	4 (1%)	3 (1%)	0	0	0	0
Missing	14 (4%)*	0	0	0	0	0
Menopausal status						
Premenopausal	0	58 (23%)	0	266 (51%)	160 (55%)	390 (100%)

ER+ and HER2+	32 (9%)	3 (1%)	0	8 (2%)	187 (64%)	0
ER+ and HER2–	168 (47%)	188 (76%)	410 (74%)	409 (79%)	0	0
ER– and HER2+	9 (3%)	2 (1%)	0	7 (1%)	105 (36%)	0
Triple negative	112 (31%)	54 (22%)	142 (26%)	92 (18%)	0	390 (100%)
Missing	35 (10%)*	0	0	1 (<1%)	0	0
Histology						
Breast cancer of no special type	321 (90%)	193 (78%)	115 (21%)	414 (80%)	265 (91%)	333 (85%)
Invasive lobular carcinoma	18 (5%)	17 (7%)	50 (9%)	71 (14%)	13 (4%)	1 (<1%)
Metaplastic	0	3 (1%)	0	0	0	20 (5%)
Other	3 (1%)	34 (14%)	39 (7%)	25 (5%)	14 (5%)	36 (9%)
Missing	14 (4%)*	0	348 (63%)	7 (1%)	0	0
Pathologist TIL score (%)						
Median (IQR)	12.5 (5–30)	13.0 (8–30)	15.0 (10–30)	10.0 (5–20)	5.00 (2–20)	20.0 (5–70)
Missing	14 (4%)*	0	0	0	0	0

Data are presented as n (%), unless noted otherwise. ER: estrogen receptor. HER2: human epidermal growth factor receptor 2. TCGA: The Cancer Genome Atlas. TIL: tumor-infiltrating lymphocyte.

* Within the TCGA collection, 14 patients contributed 2 slides each, yielding 356 samples from 342 distinct individuals. The table reports clinical features at the patient level; therefore, the 14 extra slides (duplicates) lack separate clinical information.

† Individuals showing no detectable tumor within the breast yet with cancer involvement of the axillary lymph nodes.

For the PARADIGM cohort (**Table 1**), the median patient age was 35 years (IQR: 32–38). The group consisted entirely of female premenopausal individuals. Stage I was the most common presentation (232 out of 390 patients, 59%), followed by stage II (157 out of 390, 40%). All cases were node-negative at diagnosis and free of distant metastases. Invasive breast carcinoma of no special type made up the clear majority of tumors (333 of 390, 85%), while metaplastic carcinoma occurred in 20 patients (5%). Grade 3 tumors accounted for 336 cases (86%), and grade 2 for 52 (13%). Lymphovascular invasion was detected in 40 tumors (10%). Every tumor in this cohort was triple-negative breast cancer. The median TIL score assigned by pathologists was 20% (IQR: 5–70).

Figure 1 breaks down the dataset volumes and compositions used to develop ECTIL-TCGA, ECTIL-TNBC, and ECTIL-combined.

We began by assessing how well ECTIL-TCGA outputs aligned with pathologist assessments in the TCGA internal test portion, as well as in the independent BASIS, N4+, TRAIN II, MATADOR, and PARADIGM collections. Within the TCGA internal test data, ECTIL-TCGA achieved a Pearson correlation coefficient (r) of 0.61 and an AUROC value of 0.84. When applied to the five separate external cohorts, the model yielded r values ranging from 0.54 to 0.74 and AUROC scores ranging from 0.80 to 0.94 (**Figure 2**). The strongest match was observed in the MATADOR group ($r = 0.74$; AUROC = 0.94), with the weakest in either PARADIGM ($r = 0.58$; AUROC = 0.80) or N4+ ($r = 0.54$; AUROC = 0.82), depending on the evaluation metric. Patient-specific prediction visualizations are presented as scatter plots.

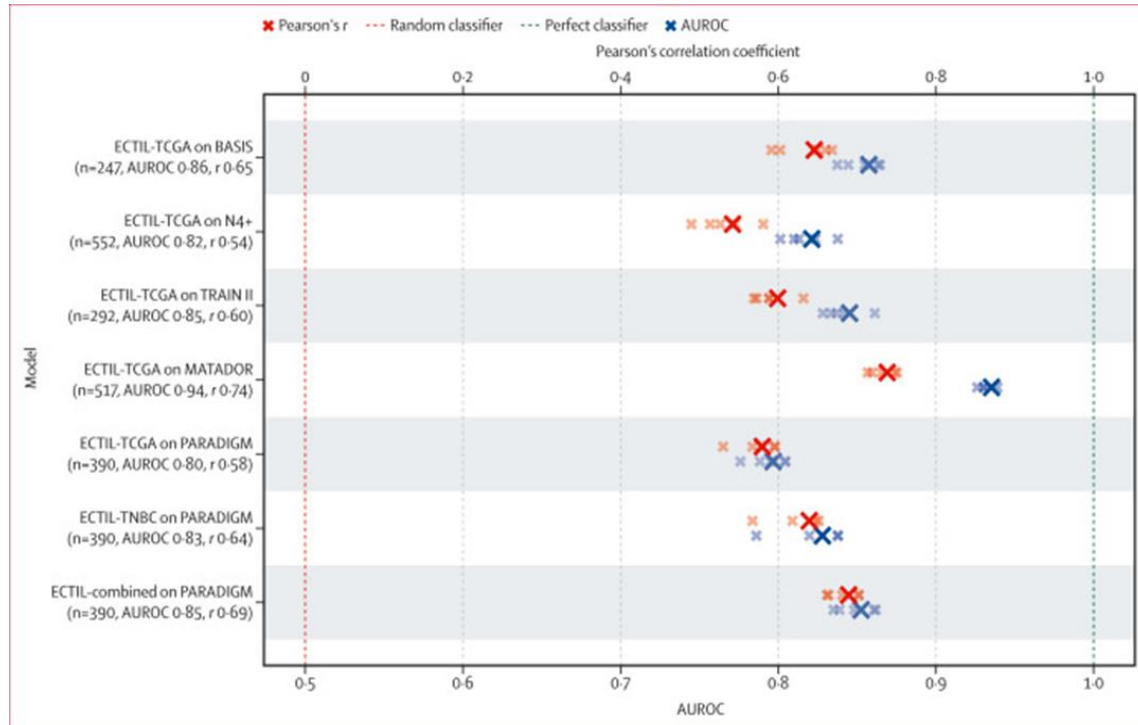


Figure 2. Summary of how all models performed across every dataset.

Large cross symbols indicate the metric obtained from the final ensemble model selected for subsequent analyses. Smaller transparent crosses show outcomes from individual models trained on different folds, illustrating the degree of variation associated with training data composition. The lower x-axis shows AUROC, calculated by dividing pathologist scores at the 30% TIL threshold into high and low categories. The upper x-axis presents Pearson's correlation coefficient (r). The y-axis lists test results for the various models on each cohort. A model performing at chance level would show r near 0 and an AUROC of 0.5, while a flawless model would achieve r of 1.0 and an AUROC of 1.0. N denotes the sample count used for each metric calculation. AUROC: area under the receiver operating characteristic curve. ECTIL: label-efficient computational stromal TIL assessment. TCGA: The Cancer Genome Atlas. TIL: tumor-infiltrating lymphocyte. TNBC: triple-negative breast cancer.

Next, the performance of ECTIL-TNBC and ECTIL-combined was specifically examined in the PARADIGM group. ECTIL-TNBC delivered improved results ($r = 0.64$; AUROC = 0.83) relative to ECTIL-TCGA ($r = 0.58$; AUROC = 0.80; **Figure 2**). ECTIL-combined, developed using all available slides from all groups except PARADIGM ($n = 1964$), attained $r = 0.69$ and

AUROC = 0.85 (**Figure 2**). In the PARADIGM set, scores from both ECTIL-TNBC and ECTIL-combined successfully predicted binarised pathologist TIL labels at 10%, 30%, 50%, and 75% thresholds, producing AUROC values of 0.86, 0.85, 0.85, and 0.90 respectively for ECTIL-combined, and 0.82, 0.83, 0.82, and 0.88 for ECTIL-TNBC.

To explore ECTIL-combined predictions more closely in the PARADIGM patients, a calibration diagram was generated (**Figure 3**). Pathologist scores were plotted on the y-axis against binned ECTIL-combined predictions (intervals of 0.05) on the x-axis. The range of ECTIL-combined outputs proved narrower (0–0.7) compared with the broader spread of pathologist scores (0.1–0.95). Additionally, ECTIL-combined tended to overestimate cases with very low pathologist scores (< 0.2) and underestimate those with higher scores (> 0.3). Several mismatched predictions were noted: samples with a pathologist score of 0.9 could receive ECTIL predictions ranging from 0.2 to 0.7. At the same time, cases predicted by ECTIL in the 0.2–0.5 range sometimes corresponded to original pathologist scores of either 0 or 0.9.

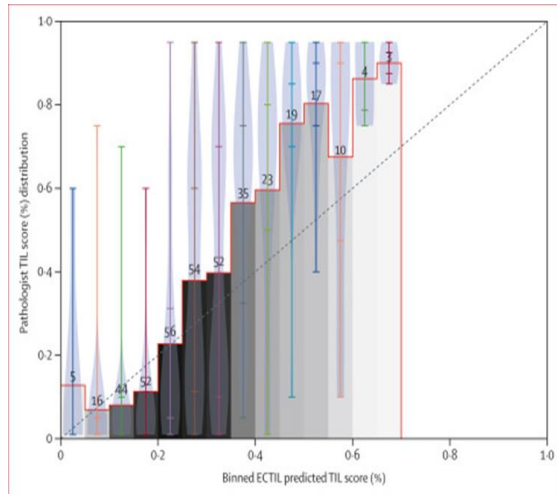


Figure 3. Calibration diagram for ECTIL-combined on the PARADIGM group.

This diagram presents the calibration performance of the ECTIL model trained across all cohorts and tested on PARADIGM (ECTIL-combined). It depicts the distribution of pathologist scores given the binned model predictions $\hat{p}(y|y)$, where y represents the actual score and $\hat{y}$ the predicted score interval. An ideal model, both accurate and well-calibrated, would align the bars precisely along the diagonal $x = y$ line across the full range. The x-axis consists of 20 evenly spaced bins of ECTIL predictions, while the y-axis indicates the spread of corresponding pathologist scores within each bin. Bar color intensity reflects sample count per bin (darkest for the bin containing the most cases, lightest for the fewest), with exact counts shown above each bar. Violin plots within bins illustrate the pathologist score distribution (minimum, 10th percentile, 90th percentile, maximum), and the embedded bar indicates the mean score for that bin. ECTIL: label-efficient computational stromal TIL assessment. TIL: tumor-infiltrating lymphocyte.

On the subset of PARADIGM samples that included annotations, ECTIL-combined reached r of 0.77 and AUROC of 0.88, whereas the open-source cell segmentation, detection, and post-processing pipeline EasTIL achieved r of 0.53 and AUROC of 0.78.

Subsequently, we examined cases showing substantial disagreement between ECTIL outputs and pathologist

scores to identify potential model weaknesses. For every cohort and within two categories (pathologist TIL scores $< 30\%$ termed pTIL-l and $> 50\%$ termed pTIL-h), one sample was randomly selected from the ten with the greatest prediction-reference differences. Heatmaps of ECTIL TIL scores and attention weights were generated for these cases and reviewed by a board-certified pathologist. One observed limitation involved invasive lobular carcinoma slides, in which tumor cells were incorrectly assigned high TIL scores. Patches displaying lymphovascular invasion, additional immune cell aggregates, fibrotic regions, or lymphocytes along tumor edges also received elevated scores and attention, even though pathologists exclude such areas. In very large slides containing small tumor regions, accumulated attention from numerous low-attention patches could shift the overall score. Aggregated subgroup confirmed quantitatively reduced performance of ECTIL-TCGA on invasive lobular carcinoma cases ($r = 0.36$, AUROC = 0.79) compared with breast cancer of no special type ($r = 0.58$; AUROC = 0.83) across cohorts. No consistent drop in lymphovascular invasion was observed in the overall results. Pearson's correlation was marginally higher when invasion was present ($r = 0.66$) than absent ($r = 0.6$), but AUROC was slightly lower with invasion present (0.79) than absent (0.83).

We then determined whether TIL scores generated by ECTIL-TCGA, ECTIL-TNBC, and ECTIL-combined independently predicted overall survival. Univariate Cox regression per 10% TIL increment demonstrated statistically significant hazard ratios for all three models (ECTIL-TCGA HR: 0.87, 95% CI: 0.79–0.96; $P = 0.0010$; ECTIL-combined HR: 0.86, 95% CI: 0.78–0.94; $P = 0.0002$; ECTIL-TNBC HR: 0.85, 95% CI: 0.77–0.94; $P = 0.0002$), comparable to the pathologist scores (HR: 0.87, 95% CI: 0.82–0.92; $P < 0.0001$); (**Table 2**). In multivariate Cox regression adjusting for all other clinical variables, the HRs per 10% TIL increment stayed significant for every ECTIL version (ECTIL-TCGA HR: 0.87, 95% CI: 0.79–0.95; $P = 0.0032$; ECTIL-combined HR: 0.85, 95% CI: 0.77–0.93; $P = 0.0007$; ECTIL-TNBC HR: 0.84, 95% CI: 0.76–0.93; $P = 0.0007$) and remained similar to the pathologist score (HR: 0.86, 95% CI: 0.81–0.92; $P < 0.0001$); (**Table 2**).

Table 2. Cox regression results ($n = 384$; events = 139).

Empty cell	No TIL score		Pathologist TIL score		ECTIL-TCGA TIL score		ECTIL-combined TIL score		ECTIL-TNBC TIL score	
	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value
Univariate analysis*										
TILs (10% increment)	..	..	0.87 (0.82–0.92)	< 0.0001	0.87 (0.79–0.96)	0.0038	0.86 (0.78–0.94)	0.0012	0.85 (0.77–0.94)	0.0013
Multivariate analysis†										
TILs (10% increment)	..	..	0.86 (0.81–0.92)	< 0.0001	0.87 (0.79–0.95)	0.0032	0.85 (0.77–0.93)	0.0007	0.84 (0.76–0.93)	0.0007
T stage										
T1a/b	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..
T1c	1.02 (0.55–1.91)	0.95	1.02 (0.55–1.91)	0.95	1.11 (0.59–2.09)	0.74	1.09 (0.58–2.05)	0.79	1.11 (0.59–2.09)	0.74
T2/3	1.50 (0.81–2.80)	0.20	1.62 (0.87–3.02)	0.13	1.62 (0.87–3.02)	0.13	1.62 (0.87–3.02)	0.13	1.66 (0.89–3.10)	0.11
Tumour grade										
Grade 1–2	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..
Grade 3	0.82 (0.52–1.28)	0.38	1.01 (0.64–1.60)	0.96	0.88 (0.56–1.39)	0.58	0.92 (0.58–1.46)	0.73	0.90 (0.57–1.43)	0.66
Histology										
Breast cancer of no special interest	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..
Metaplastic	0.34 (0.11–1.08)	0.068	0.25 (0.08–0.81)	0.020	0.32 (0.10–1.03)	0.055	0.32 (0.10–1.03)	0.056	0.32 (0.10–1.02)	0.055
Other	0.61 (0.32–1.16)	0.13	0.62 (0.32–1.20)	0.16	0.60 (0.31–1.15)	0.13	0.62 (0.32–1.18)	0.15	0.62 (0.32–1.18)	0.15
Lymphovascular invasion										
Absent	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..
Present	2.05 (1.30–3.23)	0.0021	1.87 (1.18–2.97)	0.0076	1.98 (1.25–3.14)	0.0036	2.05 (1.29–3.24)	0.0022	2.00 (1.26–3.17)	0.0031
Local treatment										
Breast-conserving surgery plus radiotherapy	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..	1 (ref)	..
Mastectomy	1.39 (0.96–2.00)	0.16	1.41 (0.98–2.04)	0.067	1.47 (1.02–2.12)	0.041	1.47 (1.02–2.13)	0.038	1.48 (1.03–2.14)	0.035
Other	1.91 (0.96–2.00)	0.080	1.76 (0.71–4.36)	0.22	1.81 (0.73–4.49)	0.20	1.93 (0.78–4.79)	0.16	1.85 (0.75–4.60)	0.18

Univariate and multivariate Cox regression findings, contrasting the pathologist TIL assessment against the ECTIL versions developed on separate data collections (ECTIL-TCGA, ECTIL-combined, and ECTIL-TNBC). The term "events" denotes deaths. In both univariate and

multivariate settings, every computational variant generated hazard ratios closely mirroring those obtained from pathologist scoring. ECTIL: label-efficient computational stromal TIL assessment. HR: hazard ratio. TILs: tumor-infiltrating lymphocytes. TCGA: The

Cancer Genome Atlas. TIL: tumor-infiltrating lymphocyte. TNBC: triple-negative breast cancer.

* Concordance indices reached 0.62 ($P < 0.0001$) using the pathologist TIL score, 0.57 ($P = 0.0010$) with the ECTIL-TCGA TIL score, 0.58 ($P = 0.00018$) with the ECTIL-combined TIL score, and 0.58 ($P = 0.00017$) with the ECTIL-TNBC TIL score.

† Concordance indices amounted to 0.63 ($P < 0.0001$) in the model lacking a TIL score, 0.68 ($P < 0.0001$) for the pathologist TIL score, 0.65 ($P < 0.0001$) for ECTIL-TCGA TIL score, 0.65 ($P < 0.0001$) for ECTIL-combined TIL score, and 0.65 ($P < 0.0001$) for ECTIL-TNBC TIL score.

Full Cox regression outcomes covering all ECTIL approaches, plus EasTIL within the annotated PARADIGM portion. When applied to complete tissue sections, ECTIL-combined returned an HR of 0.81 (95% CI: 0.69–0.95; $P = 0.0078$). EasTIL, evaluated solely on annotated tumor bed regions, yielded an HR of 0.81 (95% CI: 0.61–1.07; $P = 0.14$). Restricting ECTIL-combined to the same annotated tumor bed regions preserved its statistical significance (HR: 0.83, 95% CI: 0.71–0.97; $P = 0.0018$).

Kaplan–Meier survival plots were constructed for PARADIGM cases through several TIL scoring techniques. One approach followed the earlier published pathologist TIL thresholds ($< 30\%$, $30\%–74\%$, $\geq 75\%$). Another employed a median split on pathologist TIL scores (≥ 20 vs < 20). A third applied the median split to ECTIL-combined predictions (≥ 27 vs < 27). Distinct survival differences emerged between the formed groups. Those with pathologist TIL scores of 75% or higher exhibited 100% 3-year overall survival (95% CI: 100.0–100.0), while the 30–74% group showed 89% (82.9–95.3) and the $< 30\%$ group 81% (75.4–86.2) ($P < 0.0001$). Median-based pathologist scoring produced 94% 3-year survival (95% CI: 90.6–97.5) for scores $\geq 20\%$ versus 80% (74.8–85.8) for $< 20\%$ ($P < 0.0001$). With ECTIL-combined analysis, the $\geq 27\%$ median group achieved 90% 3-year survival (95% CI: 85.8–94.4) compared with 83% (78.3–88.8) in the $< 27\%$ group ($P < 0.0043$). These patterns demonstrate the prognostic usefulness of median-split TIL evaluations from both ECTIL and pathologists among patients under 40 years old who have low-stage, node-negative triple-negative breast cancer without prior systemic treatment. Supporting HR values for the groups reinforced the predictive capacity of these methods for overall survival (Table 2).

Kaplan–Meier graphs for the annotated PARADIGM subset for ECTIL-combined across whole tissue, ECTIL-combined limited to the tumor bed, and EasTIL limited to the tumor bed. Significant survival separation occurred with both ECTIL-combined applications: whole-tissue scores ≥ 26 linked to 92% 3-year survival (95% CI: 86.4–97.4) versus 85% (78.4–93.1) for < 26 ($P = 0.0084$); tumour-bed-only scores ≥ 27 linked to 91% (85.9–97.3) versus 86% (79.3–93.4) for < 27 ($P = 0.0071$). EasTIL on tumor bed tissue failed to achieve significant separation (scores ≥ 1 : 90% [95% CI: 84.3–95.7] versus < 1 : 87% [80.1–94.9]; $P = 0.20$).

This work introduces ECTIL, a data-efficient computational model for stromal TIL evaluation that directly predicts slide-level TIL percentages from features extracted from WSI patches. To the best of our knowledge, it constitutes the most extensive retrospective evaluation to date of any computational TIL assessment tool, encompassing 2340 breast cancer patients with digitized WSIs, pathologist TIL scores, and accompanying clinical records from six distinct cohorts. We demonstrated that ECTIL, after training on as few as 200 WSIs each annotated with a single overall TIL percentage provided by a single pathologist (e.g., 30%), could generate scores that strongly aligned with pathologist judgments across multiple independent external test groups. This alignment was reflected in robust performance indicators, such as Pearson correlation coefficients and AUROC values. Moreover, the TIL scores produced by ECTIL were significantly linked to overall survival among patients younger than 40 years with node-negative triple-negative breast cancer, independent of additional clinicopathological factors. This points to ECTIL's potential for identifying lower-risk triple-negative breast cancer cases that might safely omit adjuvant or neoadjuvant chemotherapy. Our analyses also showed that incorporating broader, more varied training collections improved the model's agreement with pathologist evaluations. Even though ECTIL-TCGA and ECTIL-combined were primarily trained on samples from ER-positive breast cancer patients (200 [56%] of 356 WSIs for ECTIL-TCGA and 1405 [72%] of 1964 WSIs for ECTIL-combined), they still exhibited meaningful concordance when tested on younger node-negative triple-negative breast cancer patients. These observations underscore the value of heterogeneous training data for optimizing ECTIL. While the achieved Pearson correlation values did not indicate flawless agreement, pathologist inter- and intra-

rater variability imposes a natural ceiling on attainable performance. For example, in the MATADOR collection, the correlation between two pathologists reached 0.8, whereas ECTIL-TCGA correlated at 0.74 with the averaged pathologist scores. This suggests ECTIL-TCGA matched the mean pathologist scoring nearly as closely as two pathologists matched each other. Such robustness extended across different slide scanners as well. Although ECTIL-TCGA was trained exclusively on Aperio-scanned slides, it did not show consistent underperformance across other scanner types across the datasets. Within the N4+ group, however, it performed somewhat better on Aperio slides than on 3DHistech scans, indicating incomplete scanner independence. We suspect this partly explains why ECTIL-combined performed better on PARADIGM, having additionally encountered 3DHistech slides during training—the same scanner type used for PARADIGM. We deliberately reserved PARADIGM as a pure external test set because prior work had established TILs as a robust prognostic marker in untreated young node-negative triple-negative breast cancer patients, carrying direct therapeutic implications [5]. Separate validation efforts will be needed to assess ECTIL's survival-prediction capability across other breast cancer subtypes. Our data indicate that ECTIL shows good concordance with pathologist scoring across molecular subtypes.

Unlike conventional approaches that rely on multi-stage detection–segmentation workflows and post-processing, ECTIL consists of a single model trained with roughly two orders of magnitude fewer labels and completes training from scratch in approximately 10 minutes on standard hardware (NVIDIA GeForce RTX 2080 Ti) [10–13]. Despite this streamlined architecture, its concordance with pathologists rivaled that of more elaborate existing systems. For instance, Bai *et al.* [11] reported Spearman's rank correlation coefficients of 0.61 and 0.63 on two triple-negative breast cancer groups. At the same time, our model yielded Spearman's r ranging from 0.47 to 0.72 across five cohorts with mixed molecular subtypes, and 0.61 specifically on our triple-negative test set. When TIL scores were expressed as proportions (0 to 1), Sun *et al.* [10] documented root mean squared error values of 0.15–0.22 for three models on two triple-negative datasets in their supplementary data; ECTIL achieved RMSE between 0.14 and 0.28 on more diverse cohorts. Thagaard *et al.* [14] reported a notably high Spearman's r of 0.79 for commercial software. That system, however, was built using

immunohistochemistry-guided annotations on selected held-out slides from their local collection, making external generalisability difficult to assess. Similarly, Choi *et al.* [13] reported concordance correlation coefficients of 0.71 and 0.75 across two triple-negative sets—higher than most ECTIL results. Their model benefited from extensive training on 2712 WSIs spanning multiple cancer types, including 2334 mm² of dense tissue annotations and 558 225 individual cell labels. After applying calibration to ECTIL scores, the model aligned closely with absolute pathologist TIL values and attained a concordance correlation coefficient of 0.67, narrowing the difference with Choi *et al.* This observation is important, as many prior computational pathology models have used relatively uniform training data. Our findings support incorporating varied datasets during training to boost robustness and accuracy. Overall, ECTIL generalized effectively across cohorts and appears promising for routine clinical integration. While a direct head-to-head comparison remains challenging, only a few commercial TIL tools that rely on large-scale annotation efforts have surpassed ECTIL. By succeeding with markedly fewer labels, ECTIL directly addresses the central difficulty noted by Thagaard *et al.* [14]: the expense and effort of collecting diverse annotations for broadly applicable algorithms.

Even without directly encoding TILs Working Group rules, ECTIL learns to approximate guideline-derived scores through supervised training on pathologist labels, implying that explicit computational translation of those guidelines may be unnecessary. Its successful testing across numerous cohorts with varying patient features, specimen preparation, scanners, and tissue extraction methods highlights strong generalisability. A qualitative review of the most discrepant cases by an experienced, board-certified pathologist trained in TIL scoring helped pinpoint model weaknesses. Patch-level predictions generally reflected expected stromal lymphocyte density, and attention maps typically concentrated on pertinent areas. Major discordances arose mainly from unintended inclusion of regions such as immune cell hotspots, border lymphocytes, necrotic zones, or lymphovascular invasion areas. These problems intensified when tumor regions were small relative to the overall slide size. Notably, the TRAIN II cohort—consisting of biopsies with low tumor-to-tissue ratios—did not show a meaningful reduction in performance compared with resection specimens. Likewise, differences between ECTIL-combined models trained with or without TRAIN II

biopsy data had negligible impact on PARADIGM results. This indicates that ECTIL performance is largely unaffected by tissue extraction approach. Restricting inference to annotated tumor bed regions did not substantially improve most metrics for ECTIL-combined, though absolute scores shifted closer to pathologist values. Scoring invasive lobular carcinoma remained challenging, consistent with pathologists' known difficulties due to morphological overlap between tumor cells and lymphocytes. These specific pitfalls mirror those documented in earlier studies despite our different model design [14]. ECTIL successfully ignored pen marks and air bubbles (assigning them low attention), so such artifacts did not drive discordance. A limitation inherent to ECTIL's architecture involves score dilution in oversized slides containing small tumor beds, where many low-attention tiles can collectively contribute substantial weight (up to 50%). This explains the model's observed maximum predicted score around 70%, which can be mitigated through post-hoc calibration, tumor bed masking, or adoption of context-sensitive architectures.

Several limitations apply to the current study. We included only WSIs that already possessed both a TIL score and complete clinical information. Although excluded cases showed no obvious clinical imbalance, subtle selection bias cannot be fully ruled out—for example, ethnicity data were unavailable due to privacy constraints. Cases lacking TIL scores (owing to poor WSI quality or absent tumor) were omitted. We provide detailed demographic tables for the included training and validation sets to promote transparency in patient selection. Another potential constraint is the involvement of multiple pathologists in TIL scoring across cohorts, which could affect model outcomes. However, all followed standard TIL guidelines and demonstrated good inter-observer agreement. Additionally, the authors tested only one foundation model, despite the availability of newer alternatives.

While encouraging, ECTIL would benefit from further development. Its reliance solely on slide-level labels makes it straightforward to augment training sets with new annotated cases. For example, invasive lobular carcinomas represented only 170 (7%) of 2354 samples in our collections, likely contributing to reduced performance on this histology. Employing more advanced foundation models could also improve morphological representations. ECTIL's current design processes patches independently and therefore lacks

spatial context. Future enhancements might incorporate context-aware architectures, such as vision transformers, that better capture relationships across WSIs and could more effectively implement the TILs Working Group criteria, including excluding lymphocytes outside the tumor bed and the invasive margin.

Clinical deployment of a tool like ECTIL will require a dedicated quality-control module, since even blank slides without tissue would otherwise receive a score—a problem shared by current models. Ultimately, pathologists and clinicians retain responsibility for interpreting model outputs and determining patient management. As with standard practice favoring consensus between two pathologists, combining AI and pathologist assessments should become the norm. This collaborative approach would allow pathologists to detect severe preprocessing artifacts or erroneous AI inclusions, while AI could flag areas overlooked due to time pressure. Together, such synergy is expected to elevate overall scoring quality. Prospective validation trials are essential before using ECTIL for pre-screening or pathologist assistance. ECTIL is scheduled for integration into an upcoming prospective de-escalation phase 3 trial for low-stage triple-negative breast cancer (NCT06476119).

Conclusion

In conclusion, ECTIL evaluates TILs on H&E-stained FFPE WSIs through a single-step process that bypasses deep learning segmentation and detection pipelines. The model is both data- and compute-efficient while achieving TIL scores comparable to those of expert pathologists. It also delivers similar hazard ratios in overall survival analyses independent of clinicopathological variables for patients under 40 years with node-negative triple-negative breast cancer. In the future, computational TIL tools of this kind could support patient pre-screening for de-escalation or immunotherapy studies and aid pathologists and clinicians in treatment decisions.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin.* 2023;73(1):17-48.
2. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, 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.* 2018;19(1):40-50.
3. Loi S, Salgado R, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, et al. Tumor infiltrating lymphocyte stratification of prognostic staging of early-stage triple negative breast cancer. *NPJ Breast Cancer.* 2022;8(1):3.
4. Loi S, Drubay D, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, et al. Tumor-infiltrating lymphocytes and prognosis: a pooled individual patient analysis of early-stage triple-negative breast cancers. *J Clin Oncol.* 2019;37(7):559-69.
5. de Jong VMT, Wang Y, ter Hoeve ND, Opdam M, Stathonikos N, van Diest PJ, et al. Prognostic value of stromal tumor-infiltrating lymphocytes in young, node-negative, triple-negative breast cancer patients who did not receive (neo)adjuvant systemic therapy. *J Clin Oncol.* 2022;20(20):2361-74.
6. Leon-Ferre RA, Jonas SF, Salgado R, Loi S, De Angelis C, Pusztai L, et al. Tumor-infiltrating lymphocytes in triple-negative breast cancer. *JAMA.* 2024;331(14):1135-44.
7. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, 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.* 2017;24(6):311-35.
8. Kos Z, Roblin E, Kim RS, Michiels S, Gallas BD, Chen W, et al. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *NPJ Breast Cancer.* 2020;6:17.
9. Amgad M, Stovgaard ES, Balslev E, Thagaard J, Chen W, Dudgeon S, et al. Report on computational assessment of Tumor Infiltrating Lymphocytes from the International Immuno-Oncology Biomarker Working Group. *NPJ Breast Cancer.* 2020;6:16.
10. Sun P, He J, Chao X, Chen K, Xu Y, Huang Q, et al. A computational tumor-infiltrating lymphocyte assessment method comparable with visual reporting guidelines for triple-negative breast cancer. *EBioMedicine.* 2021;70:103492.
11. Bai Y, Cole K, Martinez-Morilla S, Reisenbichler ES, Rimm DL, Wang C, et al. An open-source, automated tumor-infiltrating lymphocyte algorithm for prognosis in triple-negative breast cancer. *Clin Cancer Res.* 2021;27(20):5557-65.
12. Thagaard J, Stovgaard ES, Vognsen LG, Hauberg S, Dahl A, Christensen IJ, 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;13(12):3050.
13. Choi S, Cho SI, Jung W, Song IH, Park SY, Cho SY, et al. Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer. *NPJ Breast Cancer.* 2023;9(1):71.
14. Thagaard J, Broeckx G, Page DB, Jahangir CA, Verbandt S, Kos Z, 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 on Breast Cancer. *J Pathol.* 2023;260(5):498-513.
15. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature.* 2016;534(7605):47-54.
16. Steenbruggen TG, Steggink LC, Seynaeve CM, van der Hoeven JJM, Hooning MJ, Jager A, et al. High-dose chemotherapy with hematopoietic stem cell transplant in patients with high-risk breast cancer and 4 or more involved axillary lymph nodes: 20-year follow-up of a phase 3 randomized clinical trial. *JAMA Oncol.* 2020;6(4):528-34.
17. van Rossum AGJ, Kok M, van Werkhoven E, van Tinteren H, Mandjes IA, Wesseling J, et al. Adjuvant dose-dense doxorubicin-cyclophosphamide versus docetaxel-doxorubicin-cyclophosphamide for high-risk breast cancer: first results of the randomised MATADOR trial (BOOG 2004-04). *Eur J Cancer.* 2018;102:40-8.
18. van Ramshorst MS, van der Voort A, van Werkhoven ED, Mandjes IA, Kemper I, Dezentjé VO, et al. Neoadjuvant chemotherapy with or

- without anthracyclines in the presence of dual HER2 blockade for HER2-positive breast cancer (TRAIN-2): a multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol.* 2018;19(12):1630-40.
19. Dackus GM, Ter Hoeve ND, Opdam M, Vreuls W, van Diest PJ, van der Wall E, et al. Long-term prognosis of young breast cancer patients (≤ 40 years) who did not receive adjuvant systemic treatment: protocol for the PARADIGM initiative cohort study. *BMJ Open.* 2017;7(8):e017842.
 20. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology.* 1991;19(5):403-10.
 21. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. *Arch Pathol Lab Med.* 2018;142(11):1364-82.
 22. Liefwaard MC, van der Voort A, van Seijen M, Sanders J, van Leeuwen-Stok AE, van de Vijver MJ, et al. Tumor-infiltrating lymphocytes in HER2-positive breast cancer treated with neoadjuvant chemotherapy and dual HER2-blockade. *NPJ Breast Cancer.* 2024;10(1):29.
 23. Bug D, Feuerhake F, Merhof D. Foreground extraction for histopathological whole slide imaging. In: Handels H, Deserno TM, Meinzer H-P, Tolxdorff T, editors. *Bildverarbeitung für die Medizin* 2015. Berlin, Heidelberg: Springer Berlin Heidelberg; 2015. p. 419-24.
 24. Teuwen J, Romor L, Pai A, Schirris Y. DLUP: Deep Learning Utilities for Pathology, version 0.3.22 [Internet]. 2023 [cited 2023 Apr 1]. Available from: <https://github.com/NKI-AI/dlup/releases/tag/v0.3.22>
 25. Wang X, Du Y, Yang S, Zhang J, Wang M, Zhang J, et al. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. *Med Image Anal.* 2023;83:102645.
 26. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. p. 248-55.
 27. Niehues JM, Quirke P, West NP, Grabsch HI, van den Brandt PA, de Bruïne AP, et al. Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: a retrospective multi-centric study. *Cell Rep Med.* 2023;4(1):100980.
 28. Schirris Y, Engelaer M, Panteli A, Horlings HM, Gavves E, Teuwen J. WeakSTIL: weak whole-slide image level stromal tumor infiltrating lymphocyte scores are all you need. In: Tomaszewski JE, Ward AD, editors. *Medical Imaging 2022: Digital and Computational Pathology*. San Diego, CA: SPIE; 2022. p. 13.
 29. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: *International Conference on Machine Learning*. PMLR; 2018. p. 2127-36.
 30. Vidal JM, Tsiknakis N, Staaf J, Loman N, Larsson C, Rydén L, et al. The analytical and clinical validity of AI algorithms to score TILs in TNBC: can we use different machine learning models interchangeably? *EClinicalMedicine.* 2024;78:102928.
 31. Detlefsen N, Borovec J, Schock J, Jha AH, Koker T, Di Liello L, et al. TorchMetrics—measuring reproducibility in PyTorch. *J Open Source Softw.* 2022;7(70):4101.