npj | breast cancer

Article

Published in partnership with the Breast Cancer Research Foundation

https://doi.org/10.1038/s41523-024-00663-1

A panoptic segmentation dataset and deep-learning approach for explainable scoring of tumor-infiltrating lymphocytes

Shangke Liu[1,4], Mohamed Amgad[1,4] ✉, Deeptej More[1], Muhammad A. Rathore[1], Roberto Salgado[2,3] & Lee A. D. Cooper[1] ✉

Tumor-Infiltrating Lymphocytes (TILs) have strong prognostic and predictive value in breast cancer, but their visual assessment is subjective. To improve reproducibility, the International Immuno-oncology Working Group recently released recommendations for the computational assessment of TILs that build on visual scoring guidelines. However, existing resources do not adequately address these recommendations due to the lack of annotation datasets that enable joint, panoptic segmentation of tissue regions and cells. Moreover, existing deep-learning methods focus entirely on either tissue segmentation or cell nuclei detection, which complicates the process of TILs assessment by necessitating the use of multiple models and reconciling inconsistent predictions. We introduce *PanopTILs*, a region and cell-level annotation dataset containing 814,886 nuclei from 151 patients, openly accessible at: sites.google.com/view/panoptils. Using PanopTILs we developed *MuTILs*, a neural network optimized for assessing TILs in accordance with clinical recommendations. MuTILs is a concept bottleneck model designed to be interpretable and to encourage sensible predictions at multiple resolutions. Using a rigorous internal-external cross-validation procedure, MuTILs achieves an AUROC of 0.93 for lymphocyte detection and a DICE coefficient of 0.81 for tumor-associated stroma segmentation. Our computational score closely matched visual scores from 2 pathologists (Spearman R = 0.58–0.61, $p < 0.001$). Moreover, computational TILs scores had a higher prognostic value than visual scores, independent of TNM stage and patient age. In conclusion, we introduce a comprehensive open data resource and a modeling approach for detailed mapping of the breast tumor microenvironment.

Advances in digital imaging of glass slides and machine learning have increased interest in histology as a source of data in cancer studies[1,2]. Tissue morphology contains important prognostic and diagnostic information and reflects underlying molecular and biological processes. This work presents approaches for the computational discovery of interpretable predictive histologic biomarkers, focusing on invasive breast carcinomas and immune response. Histopathology is a medical field where medical experts (i.e., pathologists) examine stained microscopic tissue sections to make diagnostic decisions, most often from tumor biopsies. While much of medicine relies on the clinical examination of patients, histopathology is a visual-focused field, like radiology, where much of the focus is on visual pattern recognition.

The term biomarker refers to a biological feature that we can use to indicate a clinical outcome. For example, prognostic biomarkers are biological features associated with good (or bad) prognosis, while predictive biomarkers predict response to therapy in randomized controlled trials[3]. Typically, when a histologic trait is related to outcomes in cancer, it is incorporated into the grading criteria, though this is not always the case. For example, there has been a strong focus on tumor-infiltrating lymphocytes (TILs) as a prognostic and predictive biomarker in breast cancer and other solid tumors in recent years[4]. This is because TILs infiltration can be a somewhat direct visualization of how well the host (patient) body can respond to the growing tumor by immune cells.

[1]Department of Pathology, Northwestern University, Chicago, IL, USA. [2]Department of Pathology, GZA-ZNA Ziekenhuizen, Antwerp, Belgium. [3]Division of Research, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia. [4]These authors contributed equally: Shangke Liu, Mohamed Amgad. ✉e-mail: mtageld@northwestern.edu; lee.cooper@northwestern.edu

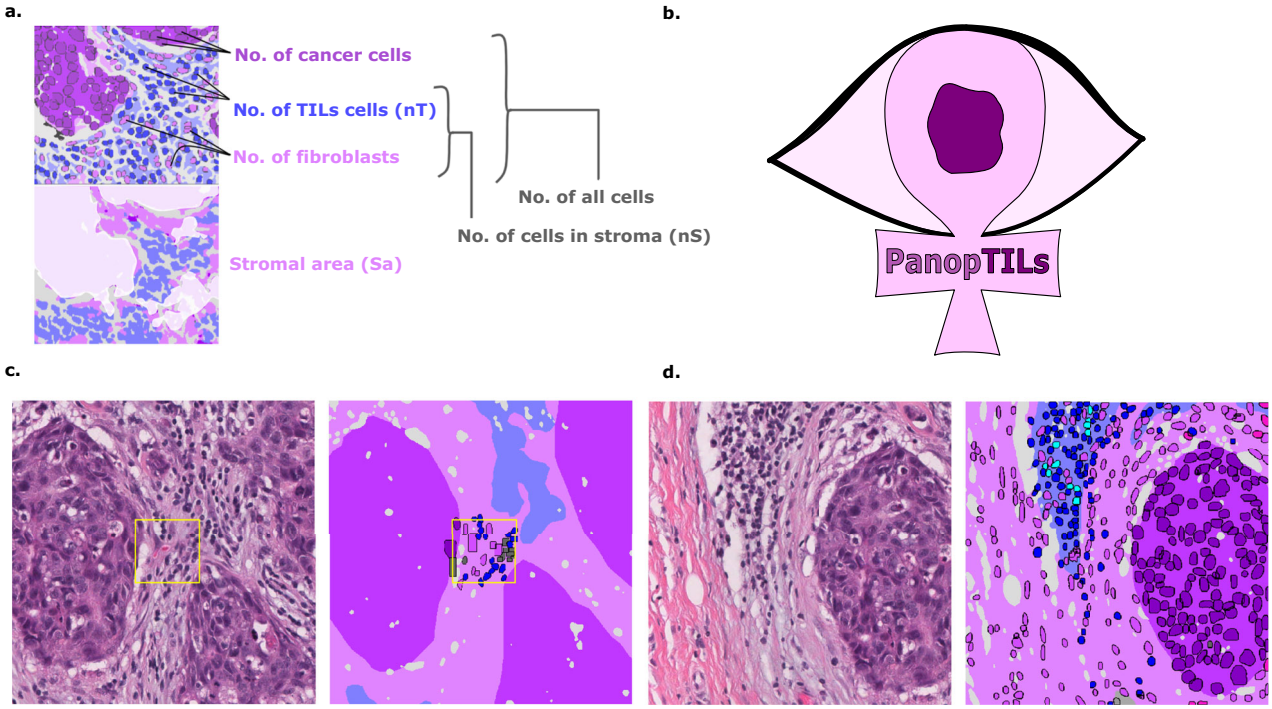

Fig. 1 | Construction of the PanopTILs dataset to facilitate computational scoring of TILs. a Components of various variants of the computational TILs score. **b** Logo of our Panoptic segmentation dataset, PanopTILs, which reconciles and expands the region-level and cell-level annotations from the BCSS[12] and NuCLS[22] datasets to better suit the task of densely mapping the tumor microenvironment for TILs assessment. PanopTILs is openly accessible at: sites.google.com/view/panoptils. **c** The result of combining manual tissue and nucleus annotations from the BCSS and NuCLS datasets. This variant of PanopTILs was used for calculating validation accuracy metrics for our panoptic segmentation model. **d** Expansion of the manual nuclei annotations to facilitate panoptic (MuTILs) model training. This expansion was done by training additional models to extrapolate nuclei annotations beyond the manual annotations as in ref. 13. These extrapolated data were used in MuTILs model training and were not used in validation.

The majority of breast cancers are carcinomas. Based on morphology, breast carcinomas include many variants; the most common are infiltrating ductal carcinoma (which originates from breast duct epithelium) and infiltrating lobular carcinoma (from breast acini/glands)[5,6]. There are numerous morphological elements within a single breast cancer slide. Integrative genomic analysis of breast cancer identified four main subtypes, including Luminal-A, Luminal-B, Her2-Enriched/Her2+, and Basal[7]. These subtypes have distinct alterations and are associated with distinct patient survival prospects[8]. TILs are particularly prognostic and predictive of therapeutic response in basal and Her2+ breast carcinomas[4].

The stromal TILs score is the fraction of stroma within the tumor bed occupied by lymphoplasmacytic infiltrates (Fig. 1). TILs are assessed visually by pathologists through examination of formalin-fixed paraffin-embedded, hematoxylin and eosin (FFPE H&E) stained slides from tumor biopsies or resections. They are subject to considerable inter- and intraobserver variability, and hence a set of standardized recommendations was developed by the International Immuno-Oncology Working Group[9,10]. Nevertheless, observer variability remains a critical limiting factor in the widespread clinical adoption of TILs scoring in research and clinical settings. Therefore, a set of recommendations was published for developing computational tools for TILs assessment[11]. Several existing computational algorithms have been developed to score TILs. However, most diverge from clinical scoring recommendations, as summarized by Amgad et al.[11]. This report describes the PanopTILs dataset and MuTILs, an interpretable deep-learning model for breast cancer WSIs, with a special emphasis on evaluating TILs (Fig. 2, Supplementary Table 1). MuTILs jointly classifies both tissue regions and individual cell nuclei to produce a panoptic segmentation for TIL scoring and other applications.

Results

Accurate panoptic segmentation of the breast cancer tumor microenvironment

MuTILs has a strong emphasis on explainability; it segments individual regions and nuclei, which are then used to calculate the computational scores (Supplementary Fig. 1). Table 1 shows the region segmentation and nucleus classification accuracy on the testing sets. MuTILs achieves a high classification performance for components of the computational TILs score, including stromal region segmentation (DICE = 80.8 ± 0.4) as well as the classification of fibroblasts (AUROC = 91.0 ± 3.6), lymphocytes (AUROC = 93.0 ± 1.1), and plasma cells (AUROC = 81.6 ± 6.6). Region segmentation performance is variable and class-dependent, with the predominant classes (cancer, stroma, and empty) being the most accurate. The region constraint improves nuclear classification AUROC by ~2–3% overall, mainly by reducing the misclassification of immature fibroblasts and large TILs/plasma cells as cancer. A detailed performance analysis of the impact of region constraint is presented in Supplementary Fig. 2 and Supplementary Tables 2–7. The generalization accuracy of MuTILs predictions is also supported by a qualitative examination of model predictions on the ROIs from BCSS and NuCLS datasets (Fig. 3) and the full WSI (Fig. 4). Note that in Fig. 4, the predictions show full WSI inference for illustration.

We compared the performance of MuTILs to previously published models for tissue region segmentation[12] and nuclei instance segmentation[13]. The region segmentation performance of MuTILs was compared to the fully convolutional network (VGG-FCN8) of[12] on common testing slides from both papers (see Supplementary Table 8). We note that while MuTILs segments tissue regions at 10× objective magnification, the VGG-FCN8 model performs segmentation at 40× objective magnification. MuTILs improves segmentation of stromal regions while sacrificing some performance on epithelial and TIL regions (see Supplementary Table 8). A per-

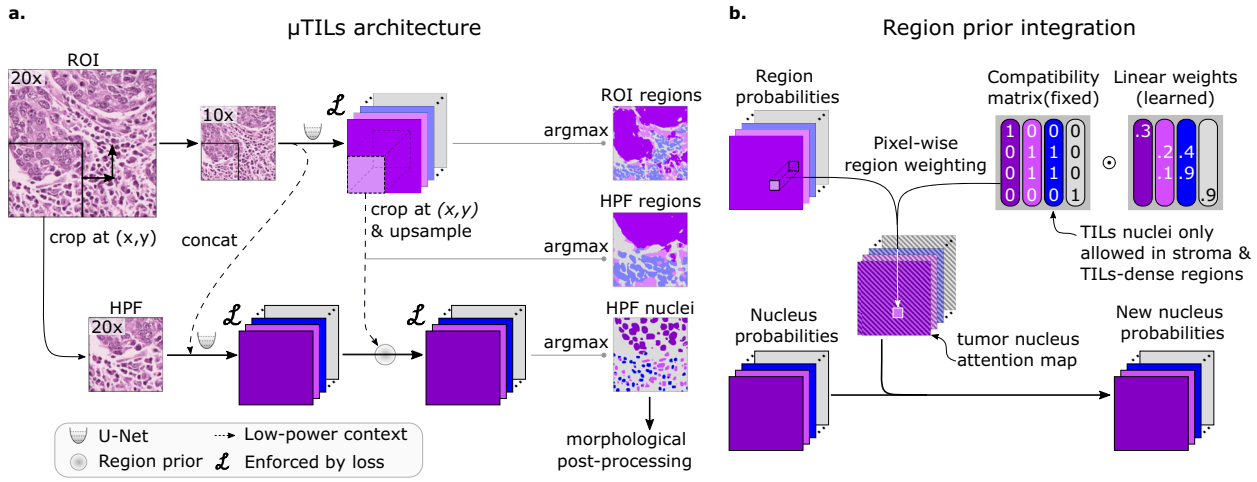

Fig. 2 | MuTILs model architecture. a The MuTILs architecture utilizes two parallel U-Net models to segment regions at 10× objective magnification and nuclei at 20× objective magnification. Inspired by HookNet, we passed information from the low-resolution region segmentation branch to the high-resolution nuclei classification branch by concatenation. This concatenation, as indicated by the dashed arrow, enriches the high-resolution data with contextual details. Additionally, region predictions from the low-resolution branch are upsampled and used to constrain the possible nucleus classifications in the high-resolution branch. The model was trained using a multi-task loss that gives equal weight to ROI and HPF region predictions, unconstrained HPF nuclear predictions, and region-constrained nuclear predictions. **b** Region predictions are used to constrain nucleus predictions to enforce compatible cell type predictions through class-specific attention maps. These maps represent the likelihood for each nuclei class occurring at different points in space based on the region prediction, user-defined hard constraints on what cell types can occupy what tissue regions and learned prior probabilities describing cell type and region type associations. Hard constraints can be used to define rules that prohibit, for example, a nucleus from being classified as a fibroblast within a tumor region.

slide performance comparison of tissue region segmentation is presented in Supplement Data 1. In nuclear classification, MuTILs performs better than the mask-RCNN model of ref. 13 on all nuclei types, including by 2% for TIL nuclei (see Supplementary Table 9). As discussed earlier, for TILs scoring, the most clinically relevant classes are stromal regions and TILs nuclei.

Computational TIL scores are moderately concordant with pathologist TIL scores

Computational TILs score variants had a modest to high correlation with the visual scores, with Spearman correlations ranging from 0.55 to 0.61 (all p-values < 0.001) (Fig. 5). Points in red are outliers that contributed to the correlation metric but were not used in calibration. Some slides were outliers with discrepant visual and computational scores; the causes for this discrepancy are discussed below. Both global and saliency-weighted scores were significantly correlated with the visual scores ($p < 0.001$). We further analyzed pathologist-pathologist concordance using Bland–Altman analysis. For the pathologist-pathologist comparison, most points fall within the +/−two standard deviations interval, with the strongest differences seen in the moderate scores ranging from 20 to 60% with no evidence of proportional bias (Supplementary Fig. 3). Score-score concordance was evaluated to measure agreement between scoring methods composed of score variants (nTSa, nTnS, nTnA) and score aggregation methods (global, saliency weighted). Correlations are high when comparing aggregation methods for the same score variant (Spearman, 0.89–0.92) and across score variants (0.72–0.86) (Supplementary Fig. 4).

Computational TIL scoring improves prognostic accuracy for infiltrating ductal carcinomas and Her2+ carcinomas

We examined the prognostic value of MuTILs on infiltrating ductal carcinomas and Her2+ carcinomas (Fig. 6). While we had access to visual scores from the basal cohort, the number of outcomes was limited, and neither visual nor computational scores had prognostic value. All metrics were obtained by saliency-weighted averaging of computational scores from 300 ROIs. Both visual and computational scores had good separation within the infiltrating ductal cohort, although only the nTnS and nTnA computational scores had significant log-rank p-values ($p = 0.009$ and $p = 0.006$, respectively). Within the Her2+ cohort, all metrics had good separation on the Kaplan–Meier, although the visual score had a borderline p-value. All computational scores were significant within this cohort ($p = 0.018$ for nTSa, $p = 0.002$ for nTnS, and $p = 0.006$ for nTnA).

We also examined the prognostic value of the continuous (non-thresholded) TILs scores using Cox proportional hazards regression, with and without controlling for clinically relevant covariates, including patient age, AJCC pathologic stage, histologic subtype, and basal status (Table 2). The analysis was restricted to slides where visual TILs scores were available for a fair comparison. In the multivariable setting, a model was built for each metric combined with clinically salient covariates. We controlled all multivariable models for patient age and AJCC pathologic stage I and II status. Additionally, we controlled models using the infiltrating ductal carcinoma subset for basal genomic subtype status, and we controlled models using the Her2+ subset for infiltrating ductal histologic subtype status. Within the infiltrating ductal cohort, the only metric with significant independent prognostic value on multivariable analysis was the nTnS computational score. Within the Her2+ cohort, the visual score was not independently prognostic ($p = 0.158$), while the computational scores all had independent prognostic value, with the most prognostic being the nTnS variant ($p = 0.003$, HR < 0.001). Saliency-weighted ROI scores almost always had better prognostic value than global computational scores.

Discussion

We present PanopTILs, a segmentation dataset that enables the joint segmentation of tissue regions and cell nuclei. This dataset enabled us to train a panoptic segmentation model, MuTILs, which is a lightweight deep-learning model for reliable assessment of TILs in breast carcinomas in accordance with clinical scoring recommendations. It jointly classifies tissue regions and cell nuclei at different resolutions and uses these predictions to derive patient-level scores. We show that MuTILs can produce predictions with good generalization for the predominant tissue and cell classes relevant for TILs scoring. Furthermore, computational scores correlate significantly with visual assessment and have strong independent prognostic value in infiltrating ductal carcinoma and Her2+ cancer.

One of the difficulties facing widespread adoption of state-of-the-art DL in medical domains is their opacity. There is a broad consensus that explainability is critical to trustworthiness, especially in clinical

Table 1 | Generalization accuracy for region segmentation and nucleus classification using manual ground truth

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
Regions at 10× objective (DICE)							
Cancer	84.4	82.1	83	82.8	82.8	82.7	0.4
Normal[c]	1.6	2.3	2.1	2.3	2.3	2.3	0.1
Stroma[a]	81.3	80.2	81	80.8	81	80.8	0.4
TILs-dense[b]	64.8	64	65.3	65.6	65.6	65.1	0.8
Necrosis/ Debris[b]	64.1	55.6	56.7	57.3	57.1	56.7	0.8
Empty	83.5	83.5	84	84.2	84.3	84.0	0.4
Nuclei at 20× objective (AUROC)							
Cancer	96.5	97.2	98	97.4	91.1	95.9	3.2
Normal[c]		84.6	89.3	80	74.7	82.2	6.3
Fibroblast[a]	90.4	93	91.8	93.5	85.8	91.0	3.6
Lymphocyte[a]	93.3	92.3	93.6	91.9	94.2	93.0	1.1
Plasma Cell[a]	80.9	73.5	88	78.9	85.8	81.6	6.6
Debris[b]	82.8	84.9	80.1	93.9	57.1	79.0	15.7
Micro-avg.	91.9	92.2	95.6	93.5	88.9	92.6	2.8
Macro-avg.	85.4	83.9	86.3	85.2	75.3	82.7	5.0
Nuclei without region constraint (AUROC)							
Micro-avg.	90.5	91.1	95.4	91.9	86.2	91.2	3.8
Macro-avg.	84.5	78.1	86.9	81.5	73.1	79.9	5.8

Results are on testing sets from the internal-external 5-fold cross-validation scheme (separation by hospital). Fold 1 contributed to hyperparameter tuning, so it is not included in the mean and standard deviation calculation. MuTILs achieves a high classification performance for components of the computational TILs score. Region segmentation performance is variable and class-dependent, with the predominant classes (cancer, stroma, and empty) being the most accurate. The region constraint improves nuclear classification AUROC by ~2–3% overall, mainly by reducing the misclassification of immature fibroblasts and large TILs/plasma cells as cancer (see qualitative examination Fig. 7). Classes and the simplified merged classes are indicated in the first and second columns respectively.

[a]Classes that contribute to the computational TILs score.

[b]Performance for Necrosis/Debris and TILs-dense regions is modest, primarily because of the inherent subjectivity of the task and variability in the ground truth. Infiltrated stromal regions do not have clear boundaries and necrotic regions also often have TILs infiltrates at the margin or adjacent areas of fibrosis, which are inconsistently labeled as necrosis, stroma, or TILs-dense in the ground truth. Nonetheless, classifying cells/material that comprise necrotic regions (neutrophils, apoptotic bodies, debris, etc.) is reasonable at higher magnification.

[c]The model fails to distinguish normal and neoplastic breast epithelium at 10× magnification. This failure is likely caused by: 1. The low representation of normal breast tissue in the validation data from NuCLS and BCSS datasets; 2. Inconsistency in defining "normal," which is sometimes used in the sense of "non-cancer" (including benign proliferation), and sometimes only refers to terminal ductal and lobular units (TDLUs). At high resolution, the distinction between cancer versus normal/benign epithelial nuclei is reasonable.

applications[1,13–15]. The standard application of DL models in histopathology involves the direct prediction of targets from the raw images. For example, we may predict patient survival given a WSI scan[16]. However, an alternative paradigm is beginning to emerge that combines the strong predictive power of opaque DL models and the interpretable nature of handcrafted features, a technique called Concept bottleneck modeling[17]. The fundamental idea is simple: 1. Use DL to delineate various tissue compartments and cells; 2. Extract handcrafted features that make sense to a pathologist; 3. Learn to predict the target variable, say patient survival, using an interpretable ML model that takes handcrafted features as its input. Hence, the most challenging task is handled using powerful DL models, while the terminal prediction task uses highly interpretable models.

MuTILs is a concept bottleneck model; it learns to predict the individual components that contribute to the TILs score (i.e., peritumoral stroma and TILs cells) and uses those to make the final predictions. This setup makes its predictions explainable and helps identify sources of error. The region constraint helped provide context for the nuclear predictions at high resolution, which helped reduce the misclassification of immature fibroblasts and plasma cells as cancer (Fig. 7). To improve the reliability of tissue and cell classifications, we grouped model predictions into a simplified set of labels necessary for the core task of TIL quantification. For example, normal breast acini are not well represented in the training data, and so MuTILs model predictions are not reliable for distinguishing normal and cancer acini (Fig. 7, bottom row). Hence, we assessed performance at the level of grouped classes with reliable ground truth (epithelium, stroma, TILs) at evaluation. A richer set of predicted labels can be achieved by expanding the training set or downstream modeling of architectural patterns, which is beyond the scope of this work. The MuTILs model analyzes a typical slide in approximately 15 min on a system equipped with a single NVIDIA A4000 graphics processor. This time can be further reduced by parallelizing tiles over multiple graphics processors on a server system.

A qualitative examination of slides with discrepant visual and computational TILs scores shows there are three major contributors to discrepancies:

1. Misclassifications of some benign or low-grade tumor nuclei as TILs.
2. Variations in TILs density in different areas within the slide, which cause inconsistencies in visual scoring. This phenomenon is also a well-known contributor to inter-observer variability in visual TILs scoring[10].
3. Variable influence of tertiary lymphoid structures on the WSI-level score.

Our results show that the most prognostic TILs score variant (nTnS) is derived from dividing the number of TILs cells by the total number of cells within the stromal region. The visual scoring guidelines rely on the nTSa, which is reflected in the slightly higher correlation of the nTSa variant with the visual scores compared to nTnS[9]. So why is nTnS more prognostic than nTSa? There are two potential explanations. First, it may be that nTnS is better controlled for stromal cellularity since it would be the same in low- vs. high-cellularity stromal regions if the proportion of stromal cells that are TILs is the same. Second, nTnS may be less noisy since it relies entirely on nuclear assessment at 20x objective, while stromal regions are segmented at half that resolution.

Finally, we note that this validation was done only using the TCGA cohort, and future work will include validation on more breast cancer cohorts. In addition, we note that MuTILs cannot distinguish cancer from normal breast tissue at low resolution, which may necessitate manual curation of the analysis region, especially for low-grade cases.

Methods

MuTILs model design

MuTILs jointly classifies tissue regions and cell nuclei and extends our earlier work on this topic (Fig. 2)[18]. It acts as a panoptic segmentation algorithm[19]; that is, it uses semantic segmentation to delineate tissue regions and instance segmentation to segment and classify individual cell nuclei to enable a holistic, context-aware assessment of TILs. MuTILs comprises two parallel U-Net models[20] (each with a depth of 5) for segmenting tissue regions and nuclei at 10X objective and 20X objective magnifications, respectively. Inspired by the HookNet method, information is shared from the tissue region segmentation to inform nucleus segmentation by providing low-power context[21]. Additionally, region predictions from the low-resolution branch are upsampled to 20X magnification and used to constrain the predicted nucleus classes. Tissue region predictions are used to infer attention maps that define the likelihood of different nuclei types occurring based on learned prior probabilities. These attention maps also incorporate user-defined compatibility kernels that prohibit biologically implausible predictions, for example, a fibroblast nucleus in a tumor tissue region. The MuTILs model was trained using a multi-task loss that gives equal weight to Regions of Interest (ROI) ROI and High-Power Field (HPF) region predictions, unconstrained HPF nuclear predictions, and region-constrained nuclear predictions.

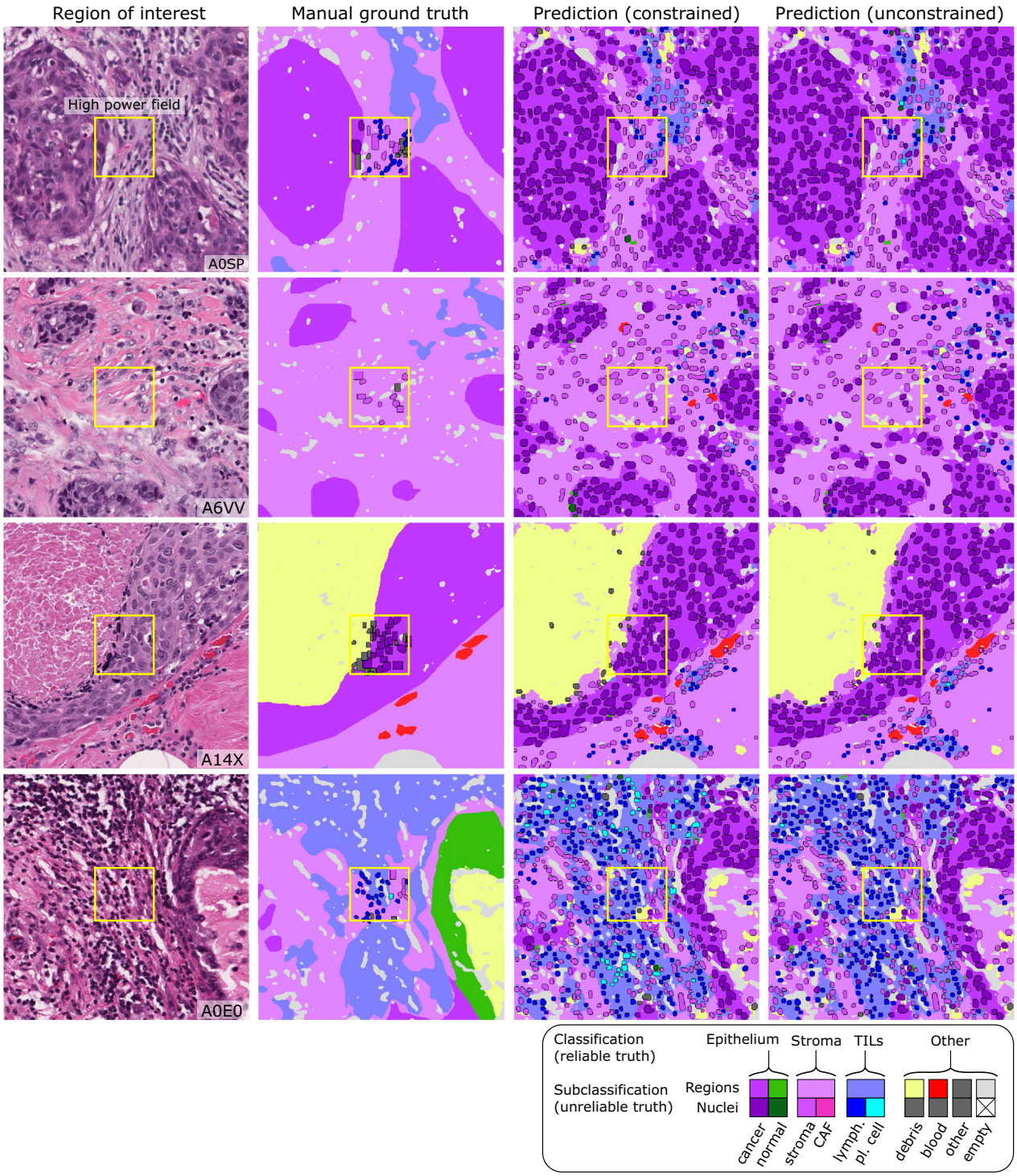

Fig. 3 | Reconciliation of manual region and nucleus ground truth to produce the PanopTILs validation dataset. Each high-power field from the pathologist-corrected single-rater NuCLS dataset was padded to 1024 × 1024 at 0.5 MPP resolution (20× objective). As a result, each ROI had region segmentation for the entire field (from the BCSS dataset) and nucleus segmentation and classification for the central portion (from the NuCLS dataset). Note that the nucleus ground truth contains a mixture of bounding boxes and segmentation. The fields shown here are from the testing sets.

PanopTILs dataset

We created a panoptic segmentation dataset that fuses the annotations from two public datasets: the Breast Cancer Semantic Segmentation dataset (BCSS)[12] and the Nucleus classification, localization, and segmentation dataset (NuCLS)[22]. These datasets were produced through a crowdsourcing process that engaged an international network of medical students, pathology residents, and pathologists using a web-based platform as described in refs. 12,22. These datasets annotated regions selected in WSIs from 125 infiltrating ductal breast carcinoma patients from The Cancer Genome Atlas. We call this combined dataset *PanopTILs*, since it enables the panoptic segmentation of tissue regions and cell nuclei necessary for TIL assessment (Fig. 1). The PanopTILs dataset contains manual annotations comprising 16,322 cancer cells, 9596 lymphocytes, 6945 fibroblasts, 5943 debris, and 4641 plasma cell nuclei (see Supplementary Table 1), along with

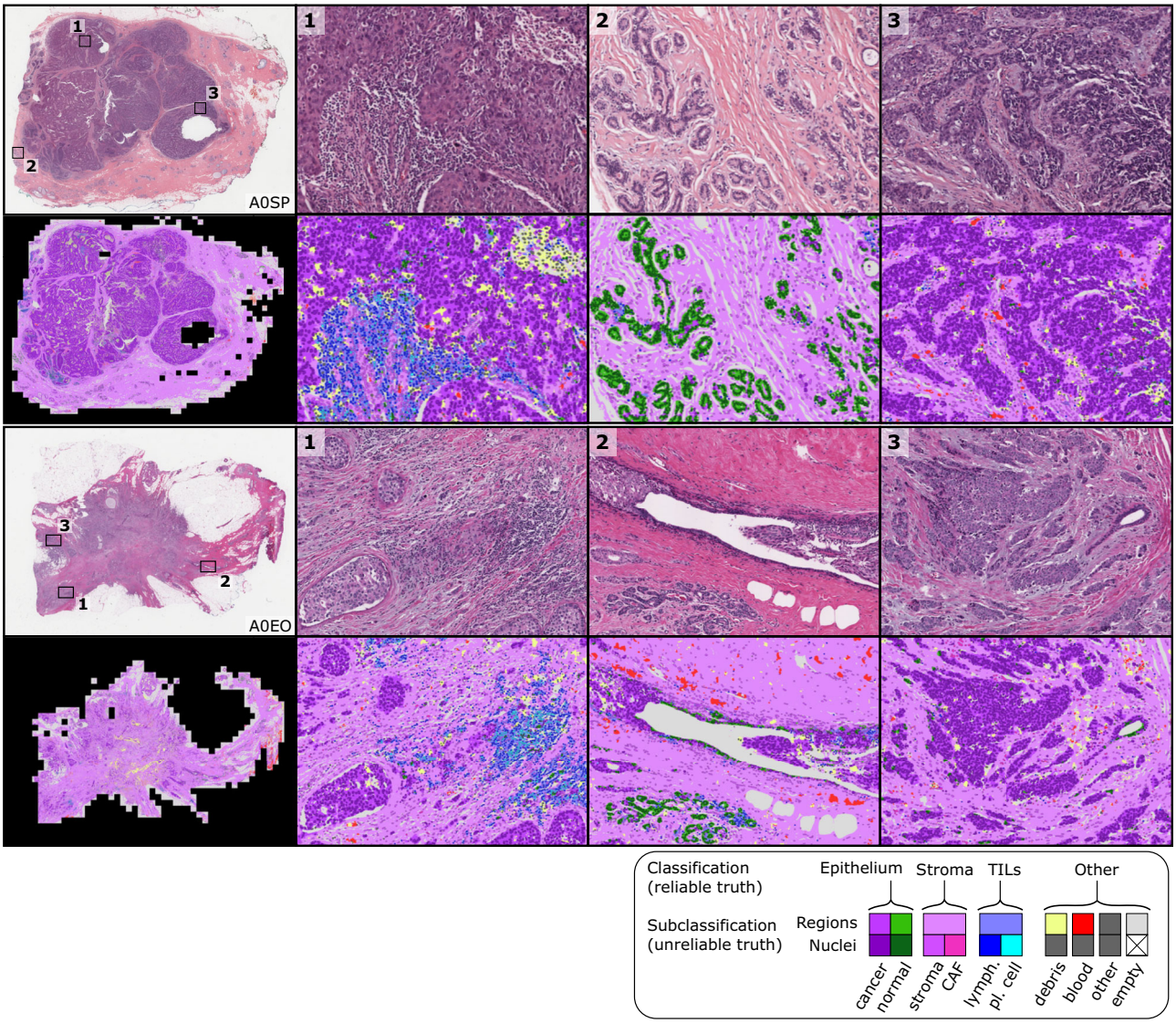

Fig. 4 | Sample whole-slide predictions from trained MuTILs models. The predictions show full WSI inference for illustration, however, our analysis only admitted the 300 most informative ROIs to the MuTILs model to limit run time to fifteen minutes per slide for practical applicability. ROI saliency was measured at a very low resolution (2 MPP) during WSI tiling and favored ROIs with more peritumoral stroma. The training set annotations of cells and tissue regions are more granular than what is necessary for TIL scoring, for example distinguishing between lymphocytes and plasma cells. Some granular classifications have lower inter-rater agreement ("unreliable") or are not abundant enough for a model to learn. Therefore, we assessed performance by grouping several classes to form a more reliable and practical ground truth (epithelium, stroma, TILs).

semantic tissue annotations within 1317 regions of interest where cancer and normal epithelium, stroma, immune infiltrates, and necrosis were annotated. Manually annotated nuclei are concentrated in 256 × 256 pixel regions-of-interest (0.5 MPP resolution, 20× objective) centered within a larger 1024 × 1024 pixel area defining the semantic segmentation (see Fig. 3, center).

MuTILs model training

For the purposes of training MuTILs models, nuclei annotations were extrapolated to the full 1024 × 1024 ROI using models for nuclear instance classification from ref. 13 to infer nuclear boundaries and classes in the periphery. The extrapolation models were trained using the manual nuclear annotations from the central 256 × 256 regions of training images and then applied to margins of the 1024 × 1024 ROI to infer nuclei annotations there (see Fig. 1d). During MuTILs model training, we also supplement PanopTILs with annotations from 85 slides from the Cancer Prevention Study II cohort to enrich the training data with lower-grade and normal tissue examples (these annotations are not included in the PanopTILs release)[23].

Analytical validation of MuTILs panoptic segmentation

Slides were separated into training and testing sets using 5-fold internal-external cross-validation[24], using the same folds described in ref. 13. This ensures that slides from a single hospital never appear in both training and testing to better estimate generalizability. In all experiments fold 1 contributed to hyperparameter tuning and so it is not included in reporting of mean and standard deviation for performance metrics. Metrics calculated include the Sørensen–Dice (DICE) coefficient for tissue segmentation, and accuracy, area under ROC curve (AUROC), sensitivity and specificity, precision and recall, F1 score, and Matthews correlation coefficient (MCC) for nucleus classification. Model performance was assessed entirely on manual cell annotations (Fig. 3). Extrapolated nuclei annotations were not used for validation. The fields depicted in Fig. 3 are from the application testing sets.

The classes in the PanopTILs training set are more granular than what is required for TIL scoring (for example, discriminating lymphocyte from plasma cell nuclei). Some of these finer subclassifications

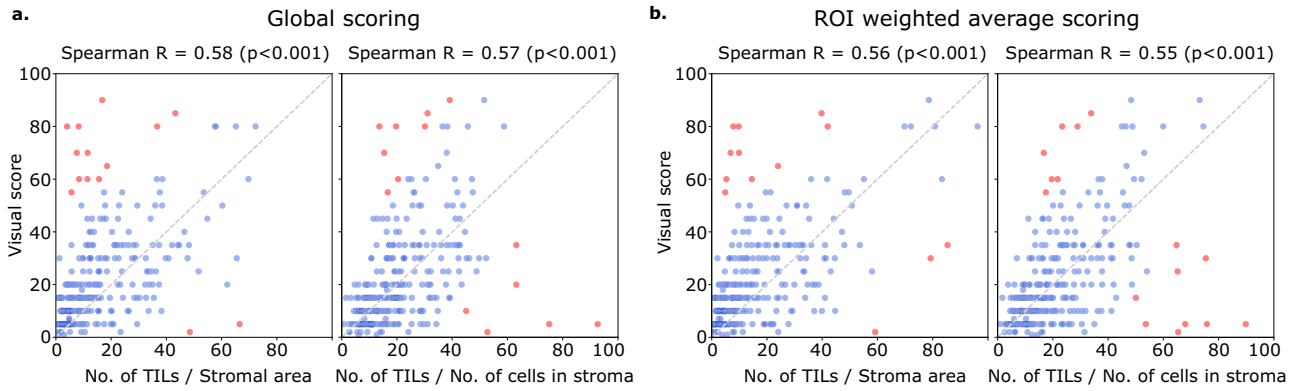

Fig. 5 | Correlation between visual and computational TILs scores. Visual scores were obtained from two pathologists using scoring recommendations from the TILs Working Group. Each point in the scatter plots above represents a single patient. Each plot above illustrates the correlation between the visual scores of one pathologist against either nTSa or nTnS scoring (using either global or salience weighted aggregation). Computational TIL scores were calibrated for the sake of interpretation to map them to a similar value range as the visual scores. Points in red are outliers that contributed to the correlation calculations but were not used during calibration. **a** Scores obtained globally by aggregating data from all ROIs. **b** Scores obtained by saliency-weighted averaging using estimated peritumoral stroma to weight each ROI.

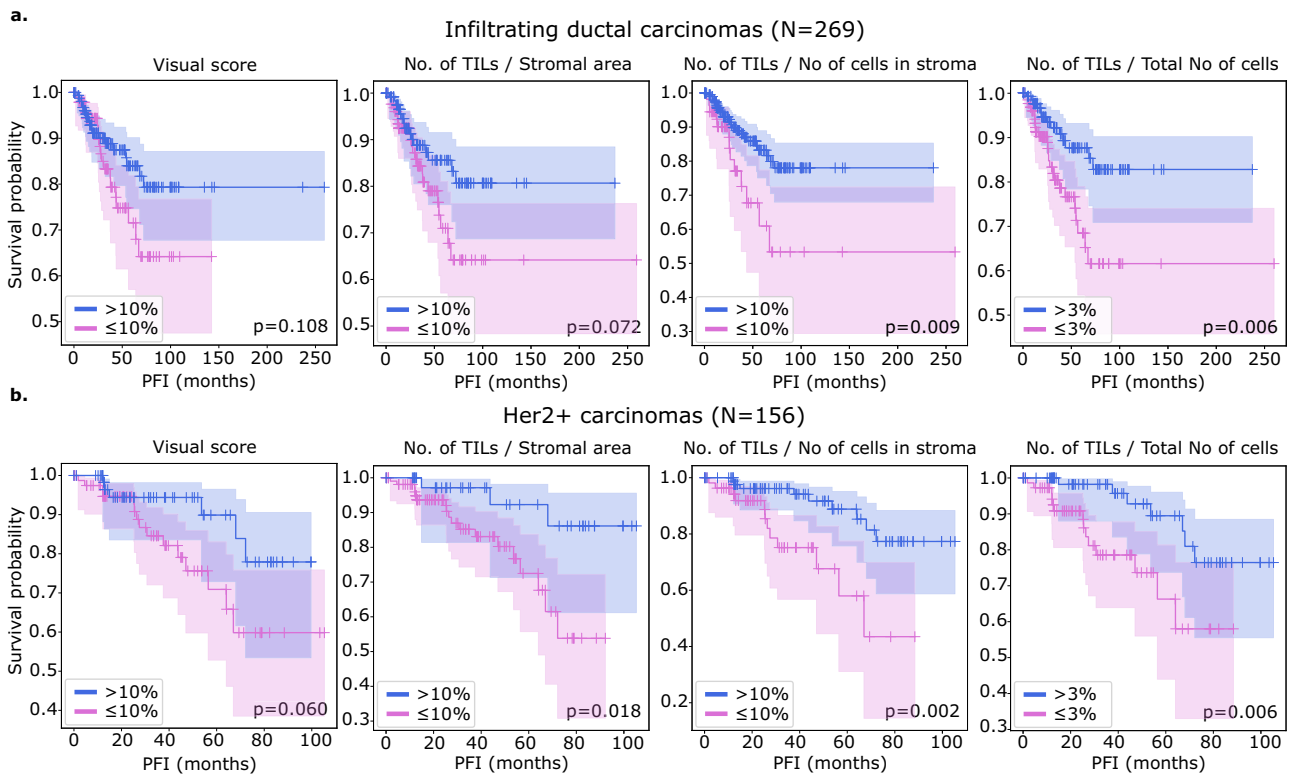

Fig. 6 | Kaplan–Meier analysis of visual and computational TILs assessment in predicting breast cancer progression. A threshold of 10% was used to define low-score and high-score patients for scores estimated in stromal regions which includes the visual, nTnS, and nTSa scores. For comparison, the nTnA score was included where the denominator includes all cells, not just those in the stromal compartment. For the nTnA score used a 3% threshold to account for the larger denominator. An analysis of the nTnA score distributions is presented in Supplementary Fig. 1 to justify this choice. We utilized visual scores from pathologist 1 as pathologist 1 is a breast cancer subspecialist. Both visual and computational scores effectively stratify outcomes in the (**a**) infiltrating ductal and (**b**) HER2+ carcinomas. Stratification for visual scores is clear but not statistically significant at the $p = 0.05$ level. Computational scores generally improve stratification over visual scores and are statistically significant, except for the nTSa in infiltrating ductal carcinoma.

have lower inter-rater agreement in the annotation datasets ("unreliable"), or are not abundant enough for a model to learn (normal epithelium). Therefore, we assessed performance by grouping several classes to form a more reliable and practical ground truth (epithelium, stroma, TILs). Predictions for normal and cancer epithelium are combined into a single "epithelium" class.

Pathologist visual TIL scoring

For whole-slide image (WSI) inference, we relied on data from 305 breast carcinoma patients for validation, 269 of whom were infiltrating ductal carcinomas, and 156 were Her2+. Visual scores were assessed by two pathologists and used as the baseline. Scores were performed in accordance with recommendations of the International TILs Working Group, which

Table 2 | Cox regression survival analysis of the predictive value of visual and computational TILs scores for breast cancer progression

Metric	Type	Univariable					Multivariable				
		HR	95% CI		*P*-value	C-index	HR	95% CI		*P*-value	C-index
Infiltrating ductal carcinoma (*N* = 269)											
Visual score		0.466	0.074	2.951	0.418	0.520	0.334	0.039	2.881	0.318	0.681
No of TILs/Stromal area	Global	*	*		0.287	0.548	*	*	*	0.321	0.667
No of TILs/No of cells in stroma	Global	0.098	0.004	2.711	0.170	0.546	0.081	0.002	3.428	0.188	0.670
No of TILs/Total No of cells	Global	0.078	*	16.98	0.353	0.526	0.073	*	29.87	0.393	0.667
No of TILs/Stromal area	ROI avg.	*	*		0.159	0.577	*	*	*	0.192	0.668
No of TILs/No of cells in stroma	ROI avg.	0.005	*	0.832	**0.042**	0.600	0.002	*	0.722	**0.038**	0.675
No of TILs/Total No of cells	ROI avg.	0.001	*	11.56	0.151	0.579	0.001	*	18.33	0.164	0.679
Her2+ carcinoma (*N* = 156)											
Visual score		0.073	0.001	3.919	0.198	0.581	0.029	*	3.952	0.158	0.725
No of TILs/Stromal area	Global	*	*		**0.039**	0.644	*	*	*	**0.011**	0.816
No of TILs/No of cells in stroma	Global	*	*	0.201	**0.015**	0.673	*	*	0.057	**0.007**	0.813
No of TILs/Total No of cells	Global	*	*	0.719	**0.045**	0.621	*	*	0.001	**0.007**	0.800
No of TILs/Stromal area	ROI avg.	*	*		**0.020**	0.679	*	*	*	**0.010**	0.837
No of TILs/No of cells in stroma	ROI avg.	*	*	0.010	**0.005**	0.704	*	*	0.002	**0.003**	0.837
No of TILs/Total No of cells	ROI avg.	*	*	0.014	**0.021**	0.660	*	*	*	**0.006**	0.833

Each metric was combined with clinically salient covariates to create a separate multivariable model. All multivariable models were controlled for patient age and AJCC pathologic stage I and II status. Additionally, we controlled models using the infiltrating ductal carcinoma subset for basal genomic subtype status, and we controlled models using the Her2+ subset for infiltrating ductal histologic subtype status. Significant *p*-values are outlined in bold, using a significance threshold of 0.05. The * symbol indicates values < 0.001.
HR Hazard Ratio, *95% CI* upper and lower bounds of the 95% confidence interval, *C-index* concordance index, *No.* number, *Avg* weighted average.

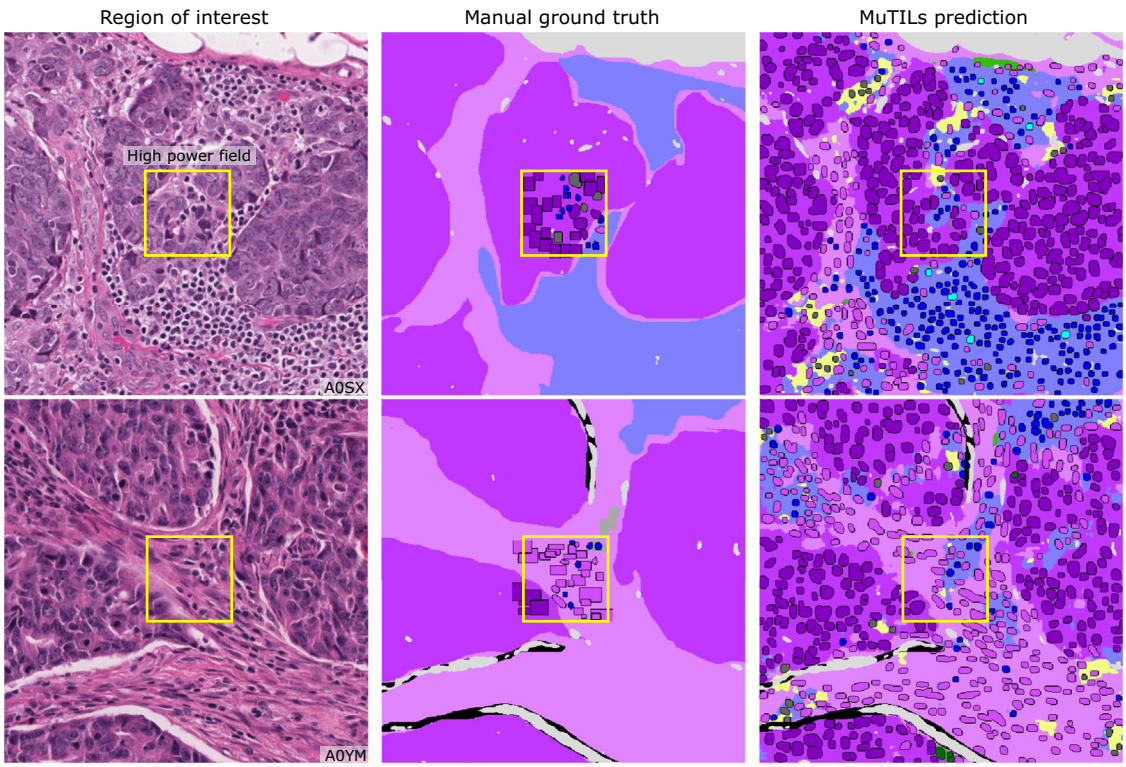

Fig. 7 | Qualitative examination of sample testing set predictions and sources of misclassification. The training set annotations of cells and tissue regions are very granular, for example distinguishing between lymphocytes and plasma cells. Some of these finer subclassifications have lower inter-rater agreement ("unreliable"), or are not abundant enough for a model to learn. Furthermore, the classes that a model needs to distinguish for the core task of TIL quantification is much simpler. Therefore, we assessed performance by grouping several classes to form a more reliable and practical ground truth (epithelium, stroma, TILs). The low abundance of normal breast acini in the training data makes it difficult for MuTILs models to distinguish normal and cancerous epithelial tissue (bottom row). We combine predictions of cancer and normal epithelium regions into a single "epithelium" class. Note how the region constraint improves nuclear classifications (third vs fourth column). This improvement is most notable for large TILs (first row) and immature fibroblasts (second row), which are misclassified as cancer without the region constraint.

recommends scoring stromal TILs as a percentage of the stromal areas between nests of carcinoma cells[9]. Scoring was performed within the border of the invasive cancer, but areas occupied by malignant cells are not included in the total assessed area. All mononuclear cells are scored (including both plasma cells and lymphocytes). Pathologists were blinded to each others' scores, and the scores of the algorithms in these experiments. Scores from reader 1 were used in clinical correlations as this reader is a breast cancer subspecialist.

Computational TIL score calculation

Analysis of a whole-slide image to generate scores begins with a tiling procedure that includes: 1. Tissue detection; 2. Exclusion of non-tissue and markers/inking regions; 3. Tiling the slide and generating an informativeness score for each tile at low resolution (2 MPP); 4. Analyzing the regions corresponding to the top 300 most informative tiles at high resolution. Fixing the number of regions ensures a near-constant run time of fifteen minutes per slide. The *large_image*[25] Python library was used to read the whole slide image files, and the *histolab*[26] library was used for the exclusion of marker/inking and non-tissue areas. The informativeness score was calculated as follows. Low-resolution tiles were deconvolved using a masked Macenko method to identify the hematoxylin and eosin components, excluding white space. This was performed with the *color_deconvolution_routine* method from the HistomicsTK package[27,28]. The informativeness score was calculated as the product of the mean hematoxylin and eosin values in each tile. Hence, tiles with a high composition of cellular (hematoxylin-rich) and acellular (eosin-rich) regions received a higher informativeness score, which favors tiles with more peritumoral stroma. Note that the informativeness score is different from the saliency score described below. The informativeness score provides a fast evaluation of where to perform the time-intensive high-resolution segmentation, while the saliency score is derived at high resolution to weigh the relative importance of ROIs in determining the overall TILs score.

At inference the 5 MuTILs models obtained from cross validation are used to perform ensembling. ROIs are assigned to the models in a cyclical manner (every fifth ROI is analyzed by the model from the first cross validation fold). This provides additional robustness without increasing the overall inference time. If runtime is not a constraint, then each ROI can be analyzed by all five models and the model outputs averaged.

Using the nucleus and tissue segmentations obtained from MuTILs models we assessed the following variants of the TILs score (Fig. 1):

1. Number of TILs/Stromal area (nTSa)
2. Number of TILs/Number of cells in stroma (nTnS)
3. Number of TILs/Number of cells anywhere (nTnA)

We obtained these score variants using two aggregation strategies: 1. Globally (aggregating region and nuclear counts from informative tiles) and 2. By saliency-weighted averaging of informative tiles. The saliency score for each tile was obtained using a Euclidean distance transform to identify stroma within 32 microns from the tumor boundary. The fraction of image pixels occupied by this peritumoral stroma was used as a saliency score for each tile. The 32 micron distance was determined by visual comparison of 8, 16, 32, and 64 microns and finding 32 to most closely represent the commonly accepted definition of peritumoral stroma.

Computational TIL score calibration

A simple linear calibration was used to scale computational scores to a similar range of magnitudes as the visual scores. This calibration procedure first z-scores the visual and computational scores to identify outliers where disagreement is greater than 1.96 standard deviations. The remaining inliers are used to define a scaling factor between computational scores and visual scores using linear regression with no intercept. This scaling improves interpretability and enables the value of a threshold intended for pathologist TIL scores to be mapped to a corresponding threshold value for computational scores.

Clinical outcomes analysis

Clinical data analysis used progression-free interval (PFI) as the endpoint used per recommendations from Liu et al. for TCGA, with progression events including local and distant spread, recurrence, or death[29]. Kaplan–Meier curves were examined for patient subgroups using a TILs-score threshold of 10% for stromal TILs scores. While different thresholds are used in the literature, a 10% is often the defining threshold for a low TIL-score. For the nTnA score variant, a threshold of 3% is used to adjust for the larger number of cells included in the denominator (see Supplementary Fig. 1). To avoid having all conclusions rely on a specific choice of threshold, TIL-scores were also included in Cox regression analysis as continuous variables.

Informed consent and ethics

All data was shared with investigators in a deidentified form. All patients participated voluntarily and provided written informed consent. CPS-II data sharing was approved through the Emory University Institutional Review Board, approval number IRB00045780.

We have complied with all relevant ethical regulations including the Declaration of Helsinki.

Data availability

The PanopTILs dataset is made public at: https://sites.google.com/view/panoptils/.

Code availability

Relevant code is publicly available at: github.com/PathologyDataScience/MuTILs_Panoptic.

Received: 18 May 2023; Accepted: 17 June 2024;
Published online: 28 June 2024

References

1. Abels, E. et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J. Pathol.* **249**, 286–294 (2019).
2. van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
3. Ballman, K. V. Biomarker: Predictive or Prognostic? *J. Clin. Oncol.* **33**, 3968–3971 (2015).
4. Savas, P. et al. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. *Nat. Rev. Clin. Oncol.* **13**, 228–241 (2016).
5. Molavi, D. W. *The Practice of Surgical Pathology: A Beginner's Guide to the Diagnostic Process*, 2nd ed. (Springer, 2017).
6. Amin, M. B. et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J. Clin.* **67**, 93–99 (2017).
7. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
8. Fallahpour, S., Navaneelan, T., De, P. & Borgo, A. Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. *CMAJ Open* **5**, E734–E739 (2017).
9. Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **26**, 259–271 (2015).
10. Kos, Z. et al. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *NPJ Breast Cancer* **6**, 17 (2020).
11. Amgad, M. et al. Report on computational assessment of Tumor Infiltrating Lymphocytes from the International Immuno-Oncology Biomarker Working Group. *NPJ Breast Cancer* **6**, 16 (2020).
12. Amgad, M. et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**, 3461–3467 (2019).

13. Amgad, M. et al. Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings. *Bioinformatics* **38**, 513–519 (2022).
14. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
15. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328 (2021).
16. Mobadersany, P., et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018)
17. Wei Koh, P. et al. Concept Bottleneck Models. In: *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. (PMLR, 2020).
18. Amgad, M. et al. Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. In: *Medical Imaging 2019: Digital Pathology* (eds Tomaszewski, J. E., Ward, A. D.) 129–136 (SPIE, 2019)
19. Kirillov, A., He, K., Girshick, R., Rother, C.& Dollar, P. Panoptic Segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 9404–9413 (CVPR, 2019).
20. Ronneberger, O., Fischer, P.& Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Vol. 9351, 234–241 (Springer, LNCS, 2015)
21. van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J. & Ciompi, F. HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med. Image Anal.* **68**, 101890 (2021).
22. Amgad, M. et al. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *Gigascience* **11**, giac037 (2022)
23. Calle, E. E. et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort. *Cancer* **94**, 2490–2501 (2002).
24. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal–external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
25. Kitware Inc. Python modules to work with large multiresolution images. https://github.com/girder/large_image (2024).
26. Marcolini, A. et al. histolab: A Python library for reproducible Digital Pathology preprocessing with automated testing. *SoftwareX* **20**, 101237 (2022).
27. Gutman, D. A. et al. The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research. *Cancer Res.* **77**, e75–e78 (2017).
28. Macenko M. et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1107–1110 (IEEE, 2009).
29. Liu, J. et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400–416.e11 (2018).

Acknowledgements

This work was supported by the U.S. NIH NCI grants U01CA220401 and U24CA19436201, NLM grant R01LM013523, and by the generosity of Ms. Jeanne Lombardo. We acknowledge support from Dr. David Gutman and the American Cancer Society, including Dr. Mia M. Gaudet, Dr. Samantha Puvanesarajah, Dr. Lauren Teras, James Hodge, and Elizabeth Bain.

Author contributions

M.A.: Idea conception, model implementation, validation, and manuscript review. S.L.: validation, manuscript writing. D.M.: code optimization, execution time measurements. M.A.R.: validation, manuscript writing. R.S.: manual scoring of TILs, manuscript approval. L.A.D.C.: Idea conception, manuscript writing. Authors M.A. and S.L. made equal contributions.

Competing interests

The authors declare the following competing interests: L.A.D.C. has invention disclosures registered at the Northwestern Office of Innovation and New Ventures, consults for Tempus. R.S. has received research support from Merck, Roche, Puma; and travel/congress support from AstraZeneca, Roche, and Merck; and he has served as an advisory board member of BMS and Roche and consults for BMS. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41523-024-00663-1.

Correspondence and requests for materials should be addressed to Mohamed Amgad or Lee A. D. Cooper.