scientific reports

Check for updates

OPEN

Crowdsourcing enables robust cell annotation for breast cancer pathology

Pascal Klöckner[1,2,3,4,9], Melis Erdal Cesur[1,9], Bart de Rooij[1], Zainab Ameziane[1], Idris Iritas[1], Rolf Harkes[1], Siamak Hajizadeh[2], Ibrahim Aloglu[5], Joyce Sanders[6], Marcos da Silva Guimaraes[6], Sybren Lodewijk Meijer[7], Rafael Bidarra[8], Jelle ten Hoeve[2], Laurens Pels[2], Sara Pires Oliveira[1]✉ & Hugo Mark Horlings[1,2]✉

The tumor microenvironment (TME) contains important morphological and molecular cues that help determine prognosis and therapeutic responses. Deep learning models can assist pathologists in assessing such biomarkers. However, generating high-quality annotations for computational pathology remains a significant bottleneck due to the level of expertise and considerable time required for fine-grained labeling. In this study, we evaluate the feasibility of using non-expert crowdsourcing for cell annotation in hematoxylin and eosin (H&E) samples of breast cancer tissue. Unlike prior work, performance assessment was based on high-fidelity ground-truth labels derived from multiplexed immunofluorescence using CODEX data, allowing for accurate benchmarking of both experts and non-experts. We collected cell annotations from experts, semi-experts, and non-experts through Tilly, a gamified annotation application designed to train and engage users in identifying major cell types within the TME. Overall, our results show that non-expert crowdsourcing is a scalable and effective strategy for generating training data for the classification of major cell types in H&E images: tumor cells, lymphocytes, and fibroblasts. Moreover, combining large, crowdsourced datasets with smaller, high-quality subsets annotated using spatial proteomics may offer a practical annotation approach to develop more robust models while minimizing biases.

The tumor microenvironment (TME) plays a crucial role in understanding tumor behavior and contains valuable information that can be used to predict both patient outcomes and treatment response. For example, scoring of tumor-infiltrating lymphocytes (TIL) has prognostic and predictive value for response to immune checkpoint inhibitors in patients with triple-negative breast cancer (TNBC) and melanoma[1,2]. Traditionally, pathologists assess and report such biomarkers through microscopic examination. However, the advent of deep learning and digital pathology enabled the development of numerous computational tools that can assist pathologists across a wide range of tasks, including automatic TIL scoring[3–5]. Nevertheless, for more detailed information on the number and location of the different cell types, it is necessary to perform cell detection using supervised models[6–8], which require large volumes of data, in this case, finely annotated. Usually, such annotations are provided manually by pathologists, a process that demands considerable time and effort, especially when diagnostic workloads are already high. This, combined with the global shortage of pathologists[9], represents a major bottleneck in advancing the field of computational pathology, particularly for fine-grained tasks such as cell classification.

One potential solution to this problem is the use of non-experts for annotation tasks, a strategy commonly applied in other domains[10]. For example, Ji et al.[11] used non-expert annotations to detect and classify mitotic figures in breast cancer images, while Albarqouni et al.[12] developed *AggNet*, which was augmented with crowd-sourced labels. Irshad et al.[13] included non-experts via a crowd-sourcing platform and showed good performance for nuclei detection in H&E-stained images. Deng et al.[14] demonstrated that non-experts can perform better

[1]Computational Pathology group, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [2]ChangeGamers, Amsterdam, The Netherlands. [3]Department of Radiation Oncology, Center for AI in Radiation Oncology, Bern University Hospital and University of Bern, Bern, Switzerland. [4]Department of Digital Medicine, University of Bern, Bern, Switzerland. [5]University of Health Sciences, Mehmet Akif Inan Training and Research Hospital, Sanliurfa, Türkiye. [6]Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [7]Department of Pathology, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [8]Computer Graphics and Visualization Group, Delft University of Technology, Delft, The Netherlands. [9]Pascal Klöckner and Melis Erdal Cesur have equally contributed to this work. ✉email: s.oliveira@nki.nl; h.horlings@nki.nl

than experts in a multiclass cell segmentation task on PAS-stained images when provided with additional immunoflorescence data, compared with experts who used PAS-stained images alone. Amgad et al.[15] found that medical students achieved good performance in classifying tumor, stromal, and immune cells in H&E-stained images. In addition to cell- and tile-level tasks, crowdsourcing has also been used for whole slide image (WSI)-level analyses in histopathology, such as malignancy classification[16].

Another critical factor affecting annotation quality in pathology is the inter-rater variability among annotators. Often, two pathologists will assign different scores or classifications to the same case, leading to variability that can introduce bias into both model training and performance assessment, as demonstrated by Kang et al.[17] for a cell classification task. On the other hand, training models based on a single pathologist's annotations is prone to reproducing biases. Therefore, there is a need for more accurate and consistent ground-truth, especially when evaluating the performance of automated methods or non-expert annotators in these tasks.

In this study, we comprehensively evaluate crowdsourcing for cell type annotations in H&E images by non-experts. In contrast to similar studies, the annotation performance of both experts (pathologists) and non-experts was assessed using a set of high-quality ground-truth annotations derived from multiplexed immunofluorescence (mIF), in particular Co-Detection by Indexing (CODEX) data[18]. The labels were assigned by two pathologists who reviewed the H&E and (registered) CODEX images side by side, simultaneously analyzing each cell's morphology and protein expression. This step was necessary because current automated cell-type classification methods remain insufficient to generate accurate ground-truth for every cell. In particular, although we explored automated pipelines based on CODEX data (*e.g.*, gating and clustering-based approaches), they sometimes produced inconsistent labels when compared to pathologists' visual evaluation. This is potentially due to cell segmentation quality and phenotypic overlap between cell types, limiting the reliability of those automated pipelines at single-cell resolution. Consequently, expert-guided annotation integrating both morphology and protein expression was required to ensure high-quality ground-truth labels. This approach also enables the evaluation of experts' performance and direct comparison with non-experts. Annotations from non-experts, semi-experts, and experts on H&E-stained images were collected using a game called *Tilly*, which uses gamification elements to train and engage users in annotating cell types common in cancer samples (such as tumor cells, lymphocytes, and fibroblasts). Finally, a deep-learning model was trained for automatic cell classification based on the ground-truth labels and the labels derived from non-expert consensus (Fig. 1).

Methods

Dataset

The TMA was sectioned at a thickness of 3μm and stained with H&E, then digitized using a 3DHistech P1000 scanner at 20X magnification (0.24μm per pixel). Following scanning, the H&E stain was removed from the tissue section, and the slide was subjected to antigen retrieval. Subsequently, the section was stained with an antibody cocktail containing 32 markers (*e.g.* DAPI, PanCK, CD8, and vimentin - full panel in the Supplementary Table) and imaged using the PhenoCycler system (v2.3.1), at 0.51μm per pixel, which employs CODEX technology for high-plex spatial imaging[18].

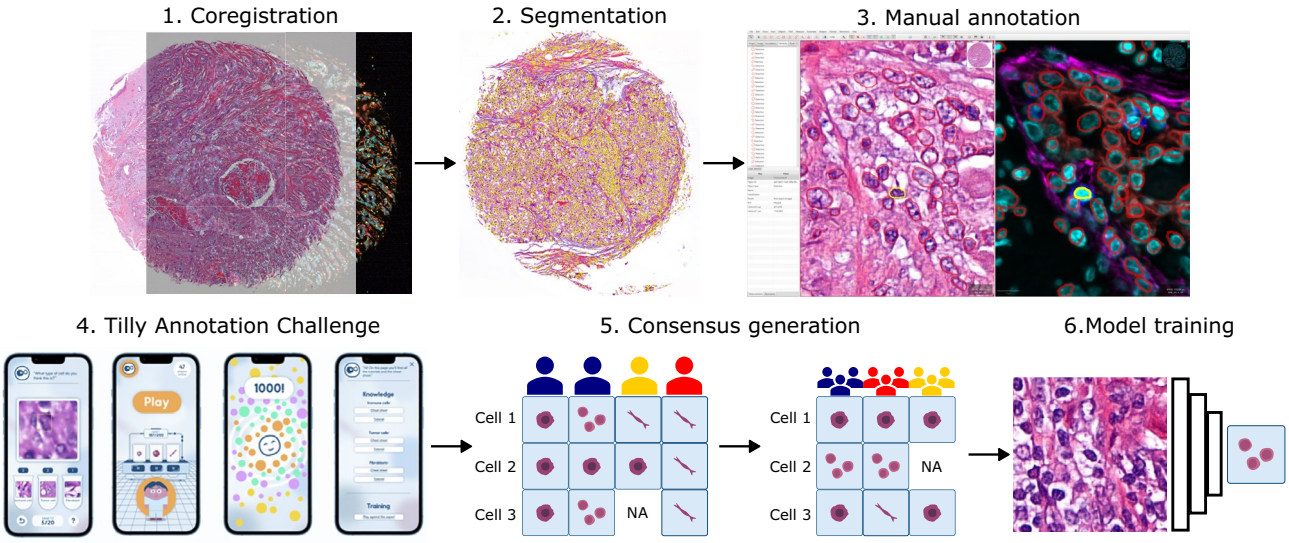

Fig. 1. Study workflow: (1) the CODEX TMA cores are registered to the corresponding H&E-stained images; (2) cells are segmented in the H&E image using the CellViT++ model[19]; (3) cell segmentations are transferred to the CODEX images and are manually annotated by two pathologists (within defined regions of interest); (4) each cell is cropped (centered in a square of 275 × 275 pixels) and presented to participants during the Tilly Annotation Challenge, to be annotated as a tumor cell, lymphocyte or fibroblast; (5) the final cell labels are generated based on a consensus of each group (experts, semi-experts and non-experts); (6) a cell classification model is trained using the generated labels.

After cropping the individual core areas, the CODEX images were registered to the corresponding H&E using the Warpy workflow[20], resulting in near-pixel-level correspondence of cells and structures. Several models, such as StarDist[21], InstanSeg[22] and CellViT++[19] were compared for cell segmentation. Based on the overlap of the generated segmentation and the DAPI channel that marks the cell nuclei by binding to DNA, the CellViT++ model was selected to segment the individual cells on the H&E cores. The model outcome was visually assessed by two pathologists.

Five regions of interest (ROIs) were selected per core by a pathologist, with each ROI containing approximately 200 segmented cells from areas representing different cell types, resulting in about 1000 cells per core and 8045 cells across all eight cores. Cells were annotated using QuPath, with the H&E and CODEX images displayed side by side. This setup enabled pathologists to assess cellular morphology in the H&E image and corresponding protein expression profiles in the CODEX image. Cells were classified as tumor cells, lymphocytes, fibroblasts, other, or no cell (indicating the absence of a DAPI signal in the CODEX image).

The "tumor cell" label was assigned to cells displaying atypical epithelial morphology on H&E and showing PanCK or E-cadherin expression on the CODEX image. The "lymphocyte" label was applied to small cells with inconspicuous cytoplasm and clumped chromatin on H&E, expressing CD3e, CD4, CD8, or CD20 on CODEX images. The "fibroblast" label was assigned to spindle-shaped cells located within the stroma, exhibiting an open chromatin pattern. Markers evaluated for this class included vimentin, aSMA, FAP, and PDGFR-b. In addition, spatial organization on H&E images was considered to avoid mislabeling myoepithelial cells or vascular smooth muscle cells, which may share similar marker profiles. Finally, the label "other" was reserved for macrophages, plasma cells, mast cells, myoepithelial cells, endothelial cells (vascular or lymphatic), and smooth muscle cells.

The final dataset comprised 4076 tumor cells (2800 IBC-NST and 1276 ILC subtypes), 1945 lymphocytes, 760 fibroblasts, and 1219 cells categorized as "other" or "no cell". For the annotation task, a cropped image of 275 × 275 pixels centered on each cell was generated, and a yellow square with a small central dot was overlaid (see Fig. 2) to direct annotators' attention to the target cell.

Tilly game

The Tilly game[1] is a puzzle-based application in which players assume the role of apprentice pathologists tasked with classifying cells by type. Towards the player, Tilly's ultimate rationale for "consuming your classified data" is clearly stated upfront: to reach very accurate advice on a patient's eligibility for immunotherapy. This clarity strongly promotes players' motivation and facilitates onboarding, even when players don't have any background related to the healthcare domain.

[1] https://www.changegamers.nl

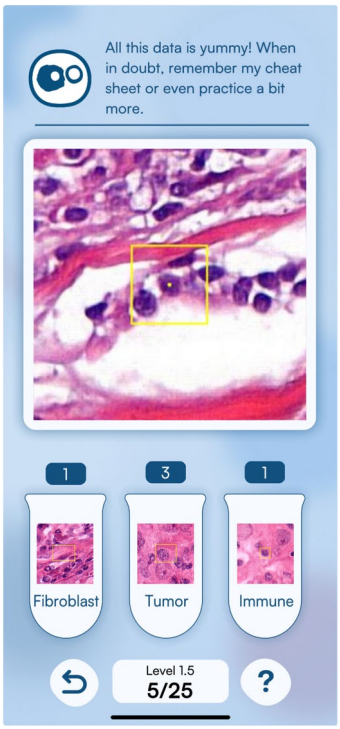

Fig. 2. A screenshot of the Tilly game. Players have to sort cells into different buckets. The specific cell of interest is marked by a small yellow dot within a yellow rectangle. Using the question mark (bottom right), players can access a cheat sheet with additional tips on how to recognize the different cell types or select the option "I don't recognize this cell".

Players self-identified their level of expertise as one of three categories: non-experts (no prior exposure to pathology images), semi-experts (*e.g.*, medical students or researchers working with pathology images), or experts (pathologists). Random spot checks of participants' backgrounds were performed to verify these self-reported classifications.

The basic game mechanics consist of displaying an image that highlights a single cell at a time, which the player must classify by dragging and dropping it into one of three categories: fibroblasts, tumor cells, or lymphocytes (Fig. 2). Non-expert players, lacking pathology experience, are introduced to the game through brief tutorials integrated into the initial levels, so that they can immediately apply the learned concepts. In essence, this initial instruction describes the three main cell types as follows:

- Tumor cells (center) are often larger than normal cells and frequently have an irregular or variable shape, with ragged edges. The nucleus is dark, enlarged, and irregular in shape. Compared to an immune cell (right), the core is not evenly colored. A tumor cell usually has a small, dark dot in its center, called the nucleolus.
- Lymphocytes are small, with a large, round nucleus that's dark purple-almost like a little button. They are usually about 1/3 the size of a tumor cell, as shown on the right.
- Fibroblasts (left) are spindle-shaped cells with long extensions. The nucleus is light purple, elongated, and centrally located in the cell. The spindle shape clearly differs from the round lymphocytes and the larger ragged tumor cells.

Players can also find these instructions in a cheat sheet, allowing them to consult it whenever they have doubts. Even if the player does not open the cheat sheet, these instructions reappear throughout the game after completing a level to remind the player of key features of the cells.

Tilly's game design integrates this central learning element with two other essential features: player engagement and retention. Players start as a basic trainee and evolve up to a senior pathologist: at each stage, they are faced with increasingly challenging classification tasks, *i.e.* cells which type is progressively less clear or intuitive. For this purpose, a careful preparation and selection of cell images has been performed in the backend.

For each game level, players are asked to classify 25 cells. This is a reasonable amount for a short effort, allowing players to be subtly encouraged to proceed for another round at the conclusion. At regular intervals, the game provides feedback in the form of encouragement, progress, or reward messages.

The Tilly challenge for this study was run by ChangerGamers at the end of June 2025 and included a total of 280 participants. They used a closed beta version of the Tilly game, released for both Android and iOS devices.

Annotation consensus

Each cell was labeled a distinct number of times by different players. To generate consensus labels, we employed a simple majority voting method. Additionally, we tested labels generated by an expectation maximization algorithm[23] as implemented by Amgad et al.[15]. However, this approach did not improve performance and barely changed the resulting labels. Thus, we used the labels from the majority voting method for the following analysis, as well as for model training.

Cell classifier

The cell type classifier is based on the ResNet-50 architecture[24], initialized with ImageNet weights and trained using Pytorch Lightning[25]. Hyperparameters were optimized using the Optuna framework[26], and the final models were trained using an initial learning rate of 3×10^{-4}, batch size of 64, crop size of 256 pixels, and a learning rate scheduler based on stalling (*ReduceLROnPlateau*) of the validation loss. Models were cross-validated (6-fold) with cells from one core being held out as the validation set for each run. The model was first trained with ground-truth labels derived from the CODEX data, and then trained with non-experts' consensus annotations, for comparison. The ground-truth labels were independently provided by two experienced breast pathologists. Each pathologist annotated distinct cell sets, following an annotation protocol based on marker intensity and morphological features, using CODEX and H&E imaging data side-by-side, ensuring consistency and reliability in the labeling process across the full dataset. Additionally, they performed random cross-checks to verify annotation quality and consistency. The final test runs were performed on a holdout set of two cores, with a balanced distribution of the different cell types, and that were neither used for hyperparameter selection nor training.

Results

Annotation performance across groups

Across all participants (n=280), we collected a total of 162,020 unique annotations of individual cells. This resulted in a median of 19 annotations per cell, with each cell being annotated at least 7 times (Table 1). The overall inter-rater agreement ranges from low (0.53) to moderate (0.68) Krippendorf alpha values (Table 2), with a good mean consensus, reflecting the percentage of votes received by the majority options. Additionally, higher levels of expertise are associated with increased inter-rater agreement.

All groups of annotators demonstrated good accuracy (non-experts: 82%, semi-experts: 83%, and experts: 83%) and weighted F1-score (non-experts: 0.82, semi-experts: 0.83, and experts: 0.83), with the best performance being achieved for tumor cells, with high recall and precision (Table 3). However, all expertise groups struggled with correctly classifying fibroblasts, which were most frequently misidentified as tumor cells (Fig. 3A). While both experts and semi-experts showed slightly better performance at detecting fibroblasts compared to non-experts (F1=0.47 and 0.46 vs. 0.41), performance metrics across all classes and expertise levels are quite similar, with only minor differences. When analyzing the performance of individual annotators, non-experts show a higher variance compared to semi-experts and experts (Fig. 3B).

	Non-experts	Semi-experts	Experts
Participants	227	42	11
Annotations	132,696	23,075	6249
Unique cells	8045	4125	1700
Annotations/cell*	16	6	5

Table 1. Descriptive statistics of the different subgroups of participants during the Tilly challenge. *median values.

	Non-experts	Semi-experts	Experts
Krippendorfs alpha	0.53	0.62	0.68
Mean consensus	79%	84%	89%

Table 2. Inter-rater agreement of the different subgroups of participants during the Tilly challenge.

Class	Group	Precision	Recall	F1-Score
Fibroblasts	Non-exp.	0.39	0.43	0.41
	Semi-exp.	0.45	0.50	0.47
	Experts	0.42	0.51	0.46
Lymphocytes	Non-exp.	0.87	0.82	0.84
	Semi-exp.	0.86	0.84	0.85
	Experts	0.87	0.79	0.83
Tumor cells	Non-exp.	0.89	0.89	0.89
	Semi-exp.	0.90	0.89	0.89
	Experts	0.91	0.91	0.91

Table 3. Annotations performance across cell types and expertise groups when compared to CODEX-derived labels.

There was no correlation ($r = -0.04$, $p = 0.55$) between the proportion of cells that an annotator labeled as "I don't recognize this cell" and their performance (Fig. 3C). Across the 914 individual annotations in which players selected "I don't recognize this cell," every instance corresponded to cells whose ground truth labels were "other" or "not a cell," underscoring their discriminative ability.

Misclassified cells

Non-experts generally demonstrated a good understanding of the basic rules introduced in the game. For example, lymphocytes are small, round cells often found in clusters, whereas fibroblasts are spindle-shaped cells that typically appear more isolated within the stroma and are surrounded by fewer neighboring cells. However, most misclassifications resulted from insufficient knowledge of additional features, including chromatin pattern, formation of structures, and the spatial context of the cell. Figure 4A presents an example of a fibroblast misclassified as a tumor cell, representing the most common type of misclassification. The ground truth label was determined to be fibroblast by integrating morphological assessment (Fig. 4A.1) with protein expression data (Fig. 4A.2). This cell appears to belong to a specific subtype of fibroblasts, cancer-associated fibroblasts (CAFs), which are located within the tumor stroma and exhibit signs of activation. The presence of an atypical nucleus may have contributed to its misclassification as a tumor cell. However, the absence of cytoplasm and its position within the stroma are features indicative of a fibroblast.

Although less common, there were also instances in which tumor cells were mistaken for fibroblasts (Fig. 4B). While cells in both images display a spindle shape, the cell in Fig. 4B.3 clearly forms a tubular structure with surrounding tumor cells. Similarly, the cell in Fig. 4B.5 is part of a tumor nest and exhibits a hyperchromatic nucleus. While these features are characteristic of tumor cells, focusing solely on shape can lead to misclassification.

Expert misclassifications, on the other hand, seem to occur when comparing to nearby, morphologically distinct cell types. While being a logical approach, it should ideally follow an initial assessment of fundamental cell features, such as size and shape. Figure 4C shows a lymphocyte (confirmed in image Fig. 4C.8 by CD4 positivity, a marker expressed on the membrane of helper T lymphocytes) that was misclassified by the expert group as a fibroblast. Compared with neighboring lymphocytes, this cell exhibits more open chromatin and an elongated shape, features suggestive of a fibroblast, which may have led to its misclassification. It is worth noting that in regions with dense inflammatory infiltration within the stroma, all groups struggled to classify cells correctly.

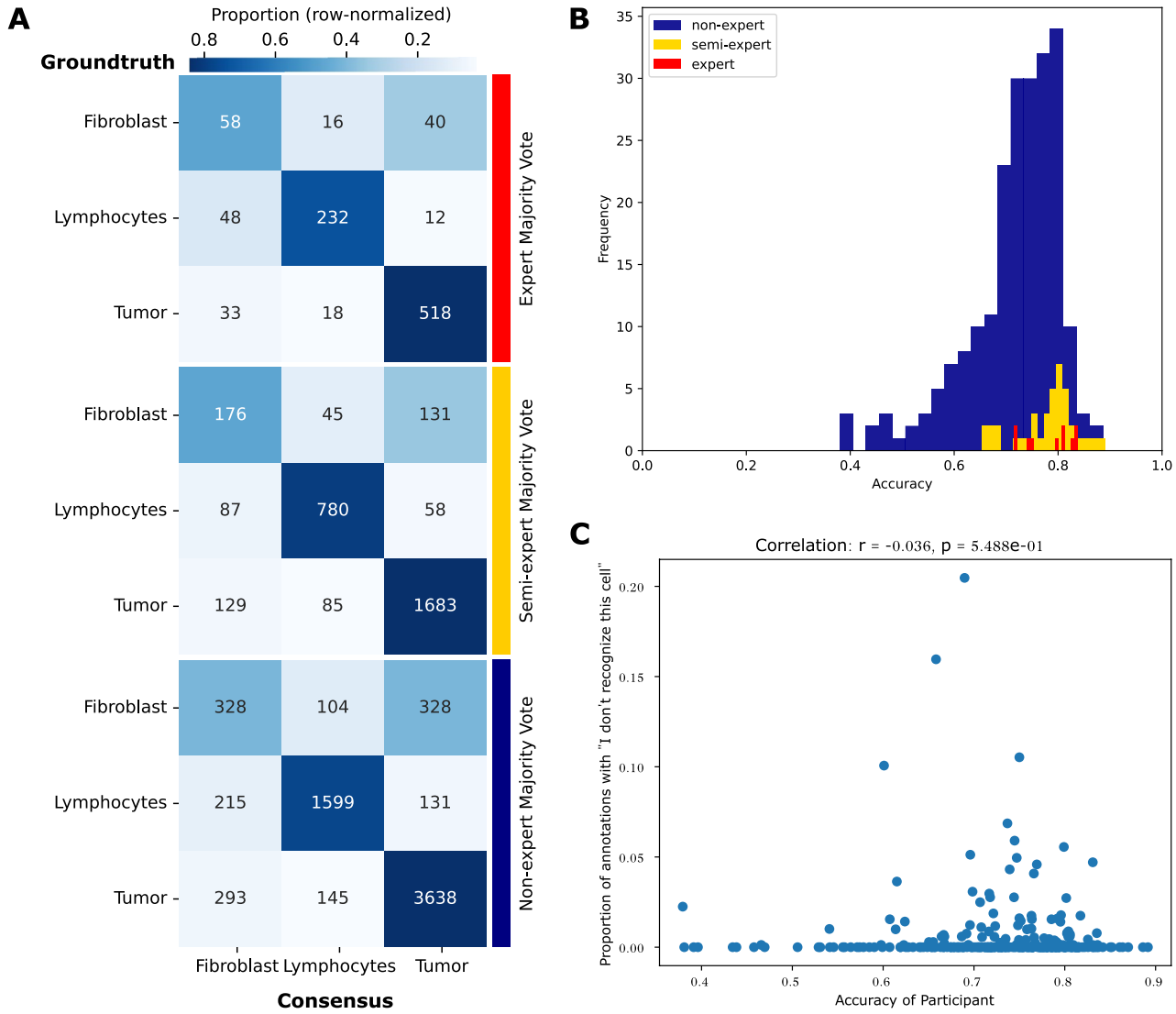

Fig. 3. (**A**) Confusion matrices per expertise group. Consensus is achieved by majority voting, and the ground-truth is derived from CODEX data. A similar pattern emerges across all groups, with lymphocytes and tumor cells showing good correspondence with the ground-truth labels, but fibroblasts often being misclassified as tumor cells by annotators. (**B**) Histogram of accuracies of participants who annotated more than 50 cells. (**C**) Scatter plot of the proportion of cells that a participant labeled with "I don't recognize this cell" and their overall annotation accuracy compared to the ground truth.

CODEX labels vs. non-expert labels

Given that non-experts performed well when compared to CODEX-derived labels, we were interested in how a model trained on non-expert labels performs against a model trained on the actual ground-truth. The model using CODEX-derived labels achieved an accuracy of 90% and an F1 score of 0.89, outperforming the annotations provided by experts. Conversely, the model trained on non-experts' labels achieved an accuracy of 82% and an F1 score of 0.82, which is comparable to the experts' (and other groups) annotation performance (Table 3).

We hypothesized that using only "top-annotators" might enhance the results. Thus, we retrained the non-experts model using only annotations from the top 30% of annotators, reaching similar performance, with an accuracy of 82% and a marginally improved F1 score of 0.83. The decision to implement such a cutoff value was a compromise between having a sufficient number of unique cell labels (thereby ensuring enough and diverse samples within the training set) and achieving good performance metrics (based on mean accuracy). Upon detailed examination of performance across individual cell types, we observe a consistent pattern aligned with the annotators' performance, as illustrated in Fig. 5. Similar to the labeling accuracy of the annotators, the model shows better performance classifying tumor cells and lymphocytes, while fibroblasts are the most challenging class to identify. This trend was also reflected in the model trained on the actual ground-truth (CODEX-derived labels), indicating that it probably focuses on similar features as the human annotators.

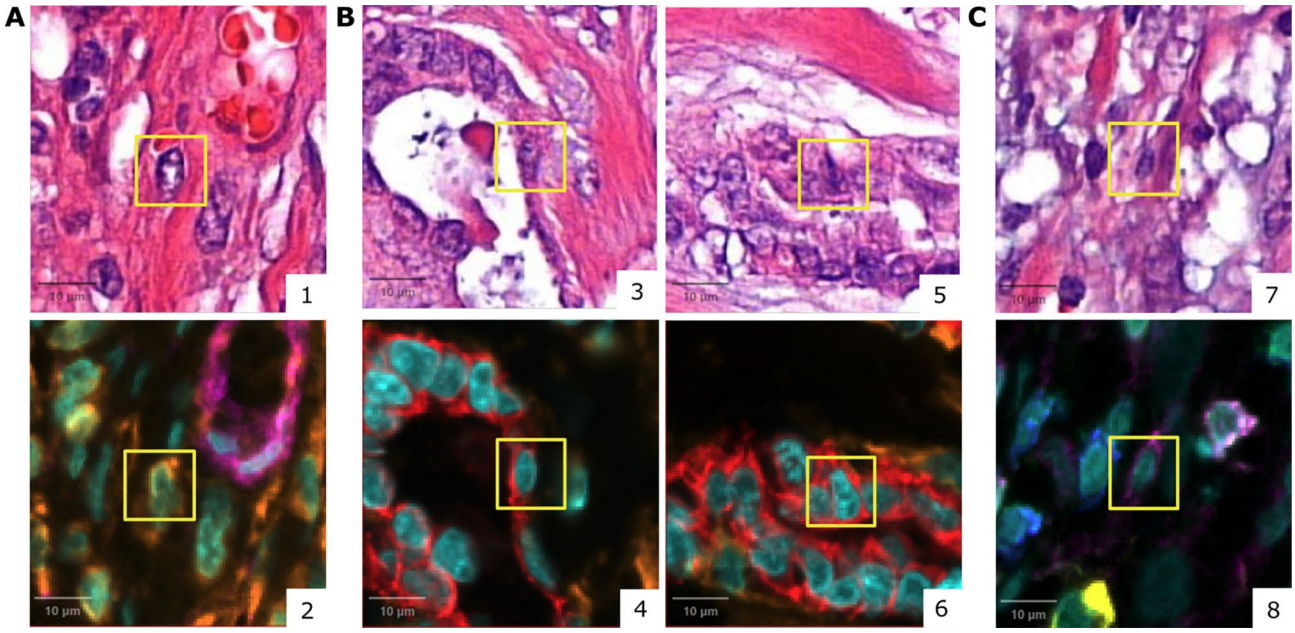

Fig. 4. Examples of cell misclassifications (scale bar: 10 µm). (**A**) Fibroblast misclassified as a tumor cell by non-experts; (1) H&E: oval-to-spindle-shaped cell lacking conspicuous cytoplasm, located within the stroma adjacent to collagen fibers; Atypical nuclear features are suggestive of an activated cancer-associated fibroblast; (2) CODEX: vimentin positivity and absence of PanCK staining indicate fibroblastic origin; (orange: vimentin, magenta: caveolin, cyan: DAPI); (**B**) Tumor cells misclassified as fibroblasts by non-experts; (3) H&E: oval-shaped tumor cell forming part of a tubular structure; (4) CODEX: strong PanCK positivity indicating epithelial origin; (5) H&E: spindle-shaped tumor cell with a hyperchromatic nucleus located within a tumor nest; (6) CODEX: strong PanCK positivity confirming epithelial origin; (red: PanCK, orange: vimentin, cyan: DAPI); (**C**) Lymphocyte misclassified as a fibroblast by experts; (7) H&E: oval-shaped lymphocyte with an open chromatin pattern; (8) CODEX: CD4 expression in the same cell, a marker characteristic of helper T lymphocytes; (blue: CD8, green: CD3e, magenta: CD4, yellow: CD31, cyan: DAPI).

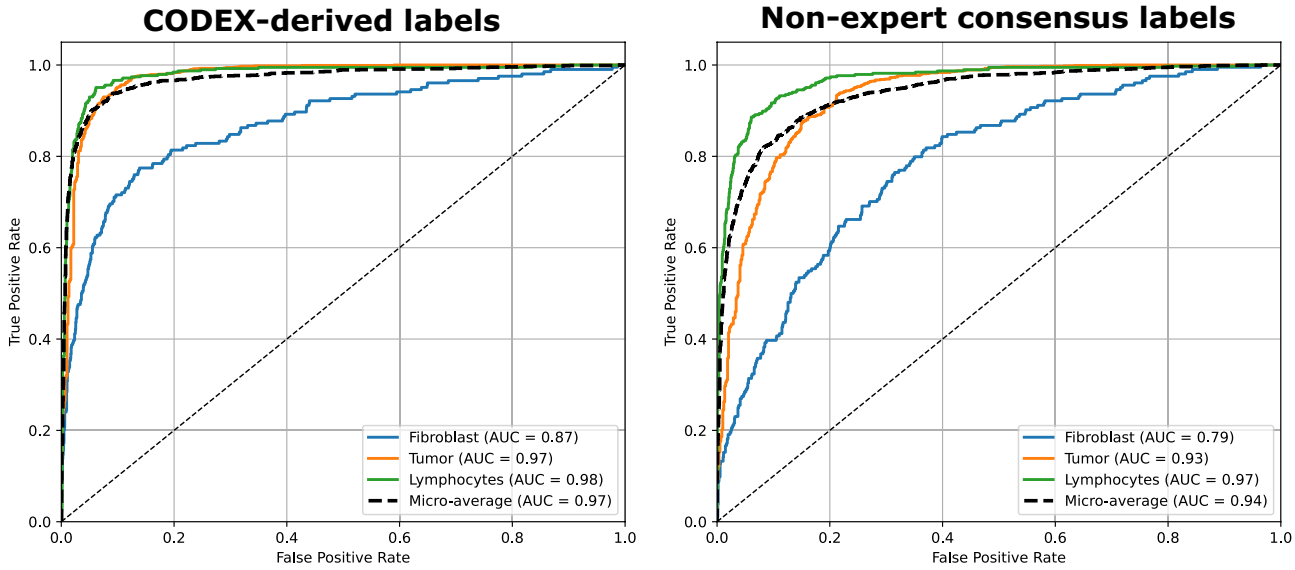

Fig. 5. The receiver operator curves (ROC) with area under the curve (AUC) metrics of the trained models on the test set.

Discussion

In this study, we thoroughly investigated the feasibility of leveraging crowdsourcing with non-expert and semi-expert annotators for cell type classification in H&E-stained images. Unlike prior work, we rely not only on morphological features visible in H&E-stained images for determining cell types, but also on protein expression data obtained from mIF. Given that each cell was manually annotated by a pathologist using both H&E and mIF composites, our dataset represents a unique source of accurate ground-truth data, which enables comparison between experts and non-experts.

The final results were revealing, as participants without expertise performed comparably to experts in annotating the three predominant cell types in H&E-stained images derived from breast cancer cases: tumor cells, lymphocytes, and fibroblasts. While experts seem to perform slightly better in detecting fibroblasts, overall metrics are fairly similar across all expertise groups. Nonetheless, all groups did not perform well on detecting fibroblasts, which is an important cell type in the context of the TME[27,28]. These findings align with those reported by Amgad et al.[15], where medical students performed the same task and demonstrated lower accuracy in labeling stromal cells, even though their ground-truth labels were derived from pathologist annotations on H&E images.

These findings have two important implications. First, while recent computational models such as HistoPlus[8], are advancing in their ability to detect and classify more complex cell types (such as subsets of lymphocytes in H&E images), it is important to keep in mind that even for relatively "simple" (*e.g.* well distinguishable based on morphology) major cell type such as fibroblasts, experts cannot be entirely relied upon to produce accurate labels for model training. Consequently, models may reproduce such errors, and despite achieving high performance, their reliability concerning these specific cell types remains questionable, due to underlying inaccuracies in the ground-truth labels. On the other hand, our research suggests that when relying exclusively on H&E-based annotations, non-expert crowdsourced consensus labels can provide a viable and scalable alternative for the generation of training labels for cell classifiers. It is also important to highlight that the game did not display the option to choose the "other" or "no cell" class, a factor that could potentially impact relative performance, as pathologists might be better at distinguishing cell types that are not covered in the game's tutorials.

In addition to evaluating the performance of groups with different levels of expertise, we also evaluated the performance of models trained on the crowdsourced cell annotations. Only the non-expert group provided enough annotations for training a model, further demonstrating the scalability advantages of the crowdsourcing approach. A model trained on these non-expert consensus labels performed worse than models trained on gold standard labels (CODEX-derived labels), but performed on par with the experts' consensus. Additionally, it showed similar biases to human annotators, with decreased performance metrics for classifying fibroblasts. These results underscore the well-known fact that the quality of a model is directly dependent on the quality of its training data: if the data contains biases, the model will inevitably learn and replicate them.

Conclusion

Deep learning has transformed the field of computational pathology, and data-hungry architectures like vision transformers have become the go-to model for many tasks. However, the availability of (good) annotated and labeled data required to train these models is still a limitation. In this study, we demonstrate that, at least for classifying major cell types from H&E images, non-expert crowdsourcing represents a feasible alternative with similar performance to pathologists. Nonetheless, we also see that even pathologists do not perform as well as expected, and cannot be 100% relied on for producing ground-truth annotations.

Given the global shortage of pathologists, we could envision future efforts to generate datasets primarily labeled by non-experts, who are, in our experience, very motivated to contribute to cancer research, with a smaller subset of samples being then manually annotated by pathologists, using techniques such as spatial proteomics. Such an approach would result in a larger, although noisy, dataset on which models could be trained, complemented by a smaller, high-quality subset serving as a gold standard to finetune and refine the base model. Overall, we believe that crowdsourcing in pathology is still an underexplored but promising method. Future research is needed to validate our findings across different tasks, such as segmentation or slide-level applications.

Data availability

Code for training the classifiers can be found in the GitHub repository: https://github.com/pkloeckner/tilly-public The cropped images (275x275 pixels, centered in the cells) used for the annotation experiments are available upon request.

Received: 6 January 2026; Accepted: 21 April 2026
Published online: 11 May 2026

References

1. Wein, L. et al. Clinical validity and utility of tumor-infiltrating lymphocytes in routine clinical practice for breast cancer patients: Current and future directions. *Front. Oncol.* https://doi.org/10.3389/fonc.2017.00156 (2017).
2. Presti, D. et al. Tumor infiltrating lymphocytes (tils) as a predictive biomarker of response to checkpoint blockers in solid tumors: A systematic review. *Crit. Rev. Oncol./Hematol.* **177**, 103773. https://doi.org/10.1016/j.critrevonc.2022.103773 (2022).
3. Schirris, Y. *et al.* Ectil: Label-efficient computational tumour infiltrating lymphocyte (til) assessment in breast cancer: Multicentre validation in 2,340 patients with breast cancer (2025). arxiv:2501.14379.
4. Sharma, A., Liu, B. & Rantalainen, M. A method for spatial interpretation of weakly supervised deep learning models in computational pathology. *Sci. Rep.* **15**, 19804 (2025).
5. Abousamra, S. et al. Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. *Front. Oncol.* https://doi.org/10.3389/fonc.2021.806603 (2022).

6. Baumann, E. *et al.* HoVer-NeXt: A Fast Nuclei Segmentation and Classification Pipeline for Next Generation Histopathology. In Burgos, N. *et al.* (eds.) *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, vol. 250 of *Proceedings of Machine Learning Research*, 61–86 (PMLR, 2024).
7. Nederlof, I. et al. Spatial interplay of lymphocytes and fibroblasts in estrogen receptor-positive her2-negative breast cancer. *NPJ Breast Cancer* **8**, 56 (2022).
8. Adjadj, B. *et al.* Towards comprehensive cellular characterisation of h&e slides, arXiv:2508.09926 (2025).
9. Walsh, E. & Orsi, N. M. The current troubled state of the global pathology workforce: A concise review. *Diagn. Pathol.* **19**, 163 (2024).
10. Crowston, K. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Bhattacherjee, A. & Fitzgerald, B. (eds.) *Shaping the Future of ICT Research. Methods and Approaches*, 210–221 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
11. Ji, Z. *et al.* Considerations for data acquisition and modeling strategies: Mitosis detection in computational pathology. In Oguz, I. *et al.* (eds.) *Medical Imaging with Deep Learning*, vol. 227 of *Proceedings of Machine Learning Research*, 1051–1066 (PMLR, 2024).
12. Albarqouni, S. et al. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans. Med. Imaging* **35**, 1313–1321. https://doi.org/10.1109/TMI.2016.2528120 (2016).
13. Irshad, H. *et al.* Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. *Pac. Symp. Biocomput.* 294–305, https://doi.org/10.1142/9789814644730_0029 (2015).
14. Deng, R. *et al.* Democratizing pathological image segmentation with lay annotators via molecular-empowered learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 497–507 (Springer, 2023).
15. Amgad, M. et al. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *Gigascience* **11**, giac037 (2022).
16. del Amor, R. et al. Annotation protocol and crowdsourcing multiple instance learning classification of skin histological images: The cr-ai4skin dataset. *Artif. Intell. Med.* **145**, 102686. https://doi.org/10.1016/j.artmed.2023.102686 (2023).
17. Kang, C., Lee, C. C., Song, H., Ma, M. & Pereira, S. Variability Matters: Evaluating inter-rater variability in histopathology for robust cell detection. In *ECCV Workshops* (2022).
18. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell* **174**, 968-981.e15. https://doi.org/10.1016/j.cell.2018.07.010 (2018).
19. Hörst, F. *et al.* Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models (2025). arxiv:2501.05269.
20. Chiaruttini, N. et al. An open-source whole slide image registration workflow at cellular precision using fiji, QuPath and elastix. *Front. Comput. Sci.* **3**, 780026 (2022).
21. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. *Cell Detection with Star-Convex Polygons*, 265–273 (Springer International Publishing, 2018).
22. Goldsborough, T. *et al.* Instanseg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation (2024). arxiv:2408.15954.
23. Dawid, A. P. & Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **28**, 20–28. https://doi.org/10.2307/2346806 (1979).
24. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
25. Falcon, W. & The PyTorch Lightning team. *PyTorch Lightning* https://doi.org/10.5281/zenodo.3828935 (2019).
26. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631, https://doi.org/10.1145/3292500.3330701 (2019).
27. Xu, M., Zhang, T., Xia, R., Wei, Y. & Wei, X. Targeting the tumor stroma for cancer therapy. *Mol. Cancer* **21**, 208 (2022).
28. Zhang, Q., Wang, Y. & Liu, F. Cancer-associated fibroblasts: Versatile mediators in remodeling the tumor microenvironment. *Cell. Signal.* **103**, 110567 (2023).

Acknowledgements

The authors would like to thank all the participants of the Tilly challenge, and the BioImaging Facility and Core Facility of Molecular Biobanking and Pathology for their support in conducting the CODEX experiments.

Author contributions

P.K., B.d.R., R.H., S.H., J.t.H., L.P., and H.H. conceived and designed the study. P.K., B.d.R., Z.A., I.I., and M.E.C. developed the methodology and performed the analysis. R.B. was responsible for the game design. M.E.C, H.H, I.A., J.S., and M.S.G. provided the (expert) annotations. P.K. and M.E.C. drafted the manuscript. S.P.O. and H.H. reviewed and revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

This work was financed by the Antoni van Leeuwenhoek Foundation and Vodafone Netherlands.

Declarations

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-50544-9.

Correspondence and requests for materials should be addressed to S.P.O. or H.M.H.