npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-00899-5

Augmented reality microscopy to bridge trust between AI and pathologists

Check for updates

Sunil Badve[1,14] ✉, George L. Kumar[2,14] ✉, Tobias Lang[3,14] ✉, Eli Peigin[4], James Pratt[2], Robert Anders[5], Deyali Chatterjee[6], Raul S. Gonzalez[1], Rondell P. Graham[7], Alyssa M. Krasinskas[1], Xiuli Liu[8], Alexander Quaas[9], Romil Saxena[1], Namrata Setia[10], Laura Tang[11], Hanlin L. Wang[12], Josef Rüschoff[13], Hans-Ulrich Schildhaus[13], Khalid Daifalla[3], Marc Päpper[3], Patrick Frey[3], Felix Faber[3] & Maria Karasarides[2] ✉

Diagnostic certainty is the cornerstone of modern medicine and critical for maximal treatment benefit. When evaluating biomarker expression by immunohistochemistry (IHC), however, pathologists are hindered by complex scoring methodologies, unique positivity cut-offs and subjective staining interpretation. Artificial intelligence (AI) can potentially eliminate diagnostic uncertainty, especially when AI "trustworthiness" is proven by expert pathologists in the context of real-world clinical practice. Building on an IHC foundation model, we employed pathologists-in-the-loop finetuning to produce a programmed cell death ligand 1 (PD-L1) CPS AI Model. We devised a multi-head augmented reality microscope (ARM) system overlayed with the PD-L1 CPS AI Model to assess interobserver variability and gauge the pathologists' trust in AI model outputs. Using difficult to interpret regions on gastroesophageal biopsies, we show that AI-assistance improved case agreement between any 2 pathologists by 14% (agreement on 77% vs 91%) and among 11 pathologists by 26% (agreement on 43% vs 69%). At a clinical cutoff of PD-L1 CPS ≥ 5, the number of cases diagnosed as positive by all 11 pathologists increased by 31%. Our findings underscore the benefits of fully engaging pathologists as active participants in the development and deployment of IHC AI models and frame the roadmap for trustworthy AI as a bridge to increased adoption in routine pathology practice.

The acceleration of deep learning and generative artificial intelligence (AI) applications in digital pathology are transformational and can potentially outperform the accuracy and reproducibility benchmarks of the typical pathology laboratory[1]. Notable examples of emerging biomedicine AI models include the detection of mitotic figures[2,3] and tumor proliferation[4], prostate cancer grading[5], lymph node metastasis detection[6], and prognosis prediction[7–10], among other applications[11]. Especially compelling are machine learning foundation models trained on both visual image analysis and natural language text[12,13] with recent progress aiming to produce high performance models that integrate a patient's medical record with their histology[14,15]. Real-world adoption of AI pathology systems is significantly lagging, however, and increasingly positioning pathologists to deal with black box outputs that spark questions about AI reliability in routine clinical practice[16,17]. Establishing the trustworthiness of AI systems in the hands of practicing pathologists, therefore, is necessary to realize improvements in patient care.

The need for establishing trust between AI and pathologists is particularly high in diagnostic immunohistochemistry (IHC), a methodology that is burdened by two critical challenges: (1) the absence of analytical reference standards across the diagnostic IHC industry and (2) a heavy reliance on subjective staining interpretation ascertained by manual scoring methodologies with outputs (e.g. H-scores) that are decoupled from the

[1]Emory University School of Medicine, Atlanta, GA, USA. [2]Bristol Myers Squibb, Princeton, NJ, USA. [3]Mindpeak, Hamburg, Germany. [4]Augmentiqs, D.N, Misgav, Israel. [5]Johns Hopkins University Baltimore, Baltimore, MD, USA. [6]MD Anderson Cancer Center, Houston, TX, USA. [7]Mayo Clinic, Rochester, MN, USA. [8]Washington University School of Medicine, St Louis, MO, USA. [9]Cologne University Hospital, Cologne, Germany. [10]University of Chicago, Chicago, IL, USA. [11]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [12]UCLA David Geffen School of Medicine, Los Angeles, CA, USA. [13]Discovery Life Sciences Biomarker Services GmbH, Kassel, Germany. [14]These authors contributed equally: Sunil Badve, George L. Kumar, Tobias Lang. ✉e-mail: sbadve@emory.edu; george.kumar@astrazeneca.com; tobias.lang@mindpeak.ai; maria@delphina.io

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

biomarker's analyte concentration[1,18,19]. Consequently, diagnostic IHC has no definable absolute 'ground truth' and must rely on inter-observer agreement as an estimate of scoring reliability. While regulatory approvals of IHC companion diagnostics and associated scoring methodologies are based on rigorous validation against clinical responses in registrational trials, real world implementation of IHC testing is front-loaded with variables and widely disparate assessments of biomarker positivity[18,20,21]. Importantly, this absence of ground truth is also problematic for training and fine-tuning AI models since the annotated training sets are derived from pathologists' subjective interpretations of physical IHC stains and not decision rules-based criteria safeguarded by universal reference standards.

Additionally, poor biopsy sampling, heterogeneous histology, complex scoring methodologies, and categorical positivity cut-offs add difficulty to the qualitative, and, at best, semi-quantitative nature of diagnostic IHC patient selection. In the case of PD-L1, a prognostic and predictive patient selection biomarker for checkpoint inhibitor immunotherapy, four FDA approved PD-L1 IHC kits[22–27] are commercially available and multiple PD-L1 antibodies are in use as laboratory developed tests (LDTs). Initially, the Blueprint Project[28,29] provided cross-assay concordance estimates but was ultimately ineffective in guiding real-world PD-L1 testing harmonization. Since then, a plethora of indication-specific concordance studies continue to leave the core challenges unaddressed but highlight that complex scoring methods such as combined positive score (CPS) and unique PD-L1 positivity cut-offs are problematic for pathologists[30–35].

Despite these challenges, IHC testing is a pathology workhorse with renewed efforts aiming to improve testing reproducibility and quantitative accuracy. The Consortium for Analytical Standardization in Immunohistochemistry initiated work to define IHC sensitivity thresholds for use in the development of IHC reference materials and standardization calibrators[36] with the first in class breast cancer IHC reference materials FDA cleared for clinical use[37–40]. In parallel, healthcare stakeholders are recognizing the impact of implementing AI capabilities and seeking to identify integration levers, while pathologists, in their unique position as clinical decision makers, keenly aware of IHC pitfalls, are grappling with the trustworthiness of automated AI models and their appropriate placement into their workflows[41,42].

To understand the factors that shape AI trustworthiness in pathology practice, we devised a framework implementing an augmented reality microscope (ARM) enabled with our PD-L1 CPS AI Model and tested the comparative effects of scoring PD-L1 CPS manually and with AI-assistance. The ARM-AI system allowed us to utilize a familiar glass slide workflow while reducing sources of scoring variability by ensuring that pathologists operated under identical conditions, including assessment of the same field of view (FOV), using uniform microscope settings and identical scoring times.

Results

Multi-organ foundation model for cell detection

Following the general research trend of large language models, AI foundation models are currently dominating the computational pathology research space[12,15,43–47]. Existing pathology AI foundation models using patch-level feature extraction do not necessarily learn explicitly about fine-grained cellular and subcellular structures, even though this is key for the precise detection of cells, cell membranes, subcellular staining and nuanced tissue segmentation. We built a cell-based multi-organ IHC foundation model using deep neural networks for clinical tasks requiring the detection of fine-grained histological structures. Employing a hybrid learning approach, we combined self-supervised learning to detect general tissue structures and supervised learning to detect explicit biological cell structures with significant medical relevance. Using multi-head knowledge distillation, we achieved model regularization and robustness. For supervised learning, we used hand-crafted cell annotations and nuanced tissue segmentation structures across a wide variety of organs. Leveraging 1.4 million annotations from 518 multi-organ biopsy cases across multiple organs sourced from 16 institutions (Fig. 1a), our AI model was able to detect individual cells, classified as tumor or immune, and detect PD-L1 staining positivity. Our emphasis on using heterogeneous data was strategically employed to compensate for differences in staining variability and scanner/microscope diversity to achieve the required robustness for clinical-grade AI.

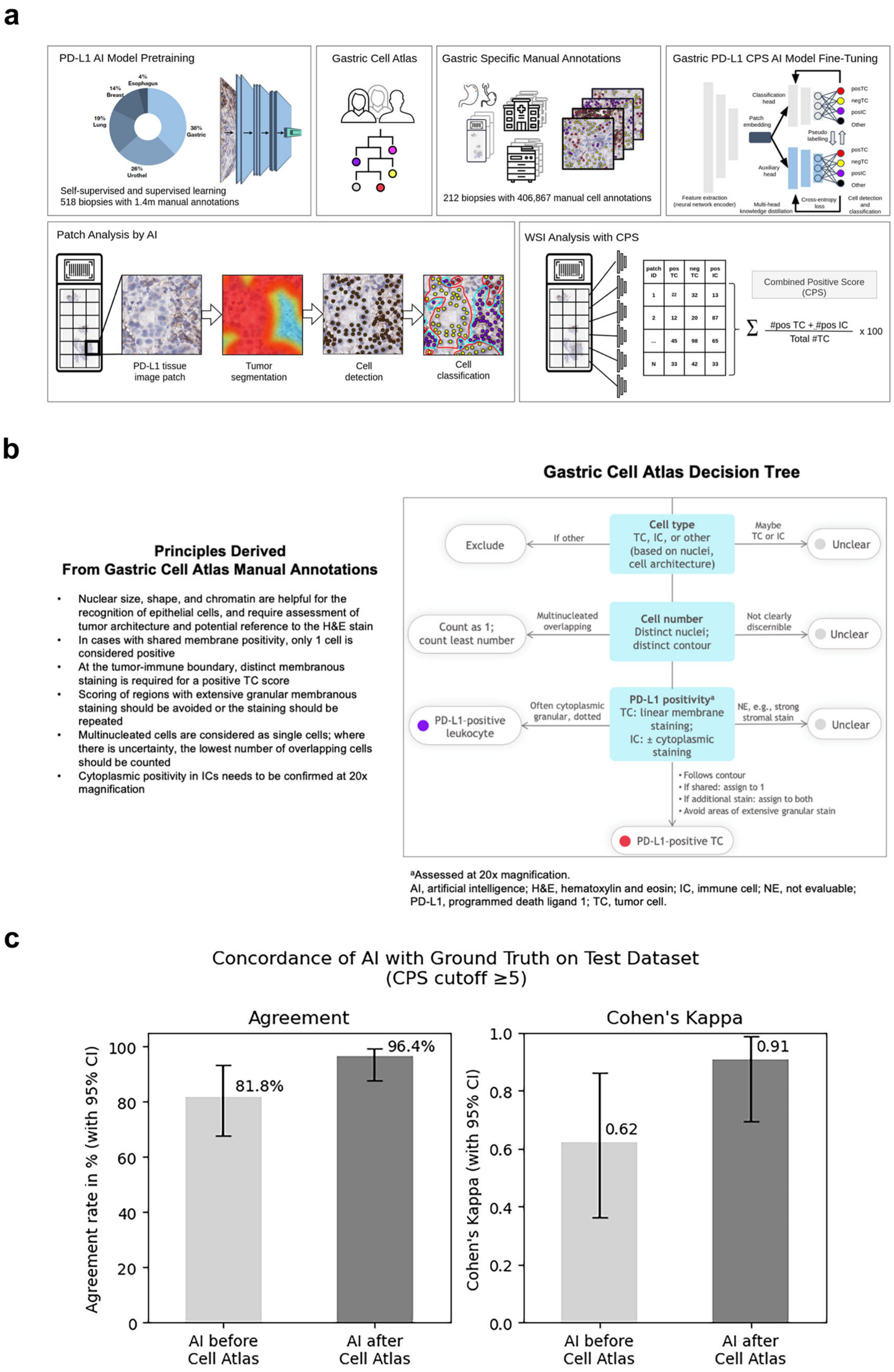
Our AI model analyzes PD-L1 stained tissue images in a patch-based approach following a stepwise procedure by first identifying assessable tissue, then segmenting tumor tissue, followed by cell detection and classification, and finally aggregating across all analyzed patches (Fig. 1a). To deploy our AI model on gastric cancer samples, we adapted the AI model to detect PD-L1 expression by the CPS methodology using gastric, gastroesophageal junction, and esophageal adenocarcinoma (GC/GEJC/EAC) biopsies. Based on cell counts, the CPS methodology requires the identification of PD-L1 on tumor cells and immune cells (lymphocytes and macrophages) with an output that is mathematically derived and defined as the number of positive tumor cells (posTC) plus the number of positive immune cells (posIC), divided by total viable tumor cells (TC), multiplied by 100. Without fine-tuning, our IHC foundation model was empirically evaluated on a test-set of ($n = 55$) PD-L1 gastric cancer biopsy cases stained with the PD-L1 IHC 28-8 pharmDx assay, obtained from three institutions and three different scanners (Fig. 1c). In terms of case classification as positive or negative using the clinically relevant cutoff of CPS ≥ 5, a PD-L1 reference manual consensus score was considered ground truth. For each case, consensus was derived from the scores of two pathologists and adjudicated through discussion when opinions differed. The outputs of the IHC foundation model were compared to the consensus score based on the prespecified cutoff resulting in an agreement rate of 81.8% (95% CI, 68.6%, 94.3%) and a Cohen's kappa value of 0.62 (95% CI, 0.36, 0.86). Visual inspection of the IHC foundation model outputs, in terms of individual cells, revealed limitations in the interpretation of challenging gastric specific tumor tissue characteristics and PD-L1 staining positivity. Our findings aligned with our previous study by Robert et al.[48] showing similar inconsistencies when pathologists manually scored PD-L1 CPS on gastric cancer biopsy WSIs.

Pathologist-in-the-loop finetuning of the gastric cancer PD-L1 CPS AI Model

As previously reported, accurate and reproducible IHC based biomarker readouts are notoriously difficult to obtain, largely due to analytical parameters and staining interpretation variability. CPS methodologies are especially problematic because they account for PD-L1 staining in tumor and immune cells and typically produce highly discordant outputs among pathologists. These discrepancies also pose challenges for developing accurate IHC AI models where consistent and reliable training data are required. To address this challenge, we curated scoring guidelines for difficult gastric cancer tissue architectures that could be utilized systematically for pathology annotations. The direct participation of expert pathologists, not annotating pathologists, allowed us to compile accurate and exhaustive decision rules on how to annotate tissue areas and single cells with expert-level precision. These decision rules were given to annotating pathologists and were used to develop the fine-tuned PD-L1 CPS AI Model.

We identified distinct ROIs from 31 GC/GEJC/EAC biopsy cases with ambiguous outputs from the AI foundation model training. All individual cells in each ROI were assessed independently by three expert pathologists for PD-L1 IHC analysis. Adjudication of all 31 biopsy cases took place in-person with extensive discussion to derive consensus and create decision rules on cell type, cell number and PD-L1 positivity. The resulting decision rules summarized as the Gastric Cell Atlas (Fig. 1b) address areas of difficulty encountered by pathologists and AI models when assessing PD-L1 CPS, including gastric-specific tissue and cell architectures, recognition and distinction of tumor cells and a variety of immune cells, counting of overlapping and multinucleated cells, determination of PD-L1 positivity on tumor cell membranes and number of PD-L1 positive tumor cells in the case of shared membrane positivity.

By reconciling heterogeneous histology and complex cellular features, we employed the Gastric Cell Atlas to guide further annotations and achieve


a
PD-L1 AI Model Pretraining
4% Esophagus
14% Breast
19% Lung
38% Gastric
26% Urothel
Self-supervised and supervised learning
518 biopsies with 1.4m manual annotations
Gastric Cell Atlas
Gastric Specific Manual Annotations
212 biopsies with 406,867 manual cell annotations
Gastric PD-L1 CPS AI Model Fine-Tuning
Classification head
Patch embedding
Pseudo labelling
Auxiliary head
posTC
negTC
posIC
Other
Feature extraction (neural network encoder)
Multi-head knowledge distillation
Cross-entropy loss
Cell detection and classification
Patch Analysis by AI
PD-L1 tissue image patch
Tumor segmentation
Cell detection
Cell classification
WSI Analysis with CPS
patch ID
pos TC
neg TC
pos IC
1 22 32 13
2 12 20 87
... 45 98 65
N 33 42 33
Combined Positive Score (CPS)
∑ #pos TC + #pos IC / Total #TC x 100
b
Gastric Cell Atlas Decision Tree
Exclude
If other
Cell type
TC, IC, or other (based on nuclei, cell architecture)
Maybe TC or IC
Unclear
Count as 1; count least number
Multinucleated overlapping
Cell number
Distinct nuclei; distinct contour
Not clearly discernible
Unclear
PD-L1-positive leukocyte
Often cytoplasmic granular, dotted
PD-L1 positivity[a]
TC: linear membrane staining;
IC: ± cytoplasmic staining
NE, e.g., strong stromal stain
Unclear
• Follows contour
• If shared: assign to 1
• If additional stain: assign to both
• Avoid areas of extensive granular stain
PD-L1-positive TC
c
Concordance of AI with Ground Truth on Test Dataset (CPS cutoff ≥5)
Agreement
Agreement rate in % (with 95% CI)
0
20
40
60
80
100
81.8%
96.4%
AI before Cell Atlas
AI after Cell Atlas
Cohen's Kappa
Cohen's Kappa (with 95% CI)
0.0
0.2
0.4
0.6
0.8
1.0
0.62
0.91
AI before Cell Atlas
AI after Cell Atlas

Principles Derived From Gastric Cell Atlas Manual Annotations

- Nuclear size, shape, and chromatin are helpful for the recognition of epithelial cells, and require assessment of tumor architecture and potential reference to the H&E stain
- In cases with shared membrane positivity, only 1 cell is considered positive
- At the tumor-immune boundary, distinct membranous staining is required for a positive TC score
- Scoring of regions with extensive granular membranous staining should be avoided or the staining should be repeated
- Multinucleated cells are considered as single cells; where there is uncertainty, the lowest number of overlapping cells should be counted
- Cytoplasmic positivity in ICs needs to be confirmed at 20x magnification

[a]Assessed at 20x magnification.

AI, artificial intelligence; H&E, hematoxylin and eosin; IC, immune cell; NE, not evaluable; PD-L1, programmed death ligand 1; TC, tumor cell.

refined PD-L1 assessments to produce an accurate PD-L1 CPS IHC AI Model. 212 GC/GEJC/EAAC biopsy cases sourced from 11 institutions, 7 scanners and inclusive of 406,867 cells, including tissue areas that challenge pathologists when reviewing gastric cancer diagnoses requiring PD-L1 CPS readouts, were used to finetune the IHC foundation model and produce the Gastric PD-L1 CPS AI Model. Preliminary performance of the Gastric PD-L1 CPS AI Model was evaluated on the same test-set of 55 gastric cancer biopsy cases used previously. Compared to the original IHC foundation

Fig. 1 | Development of an IHC AI Model finetuned to detect PD-L1 CPS in gastric cancer biopsies. a Gastric PD-L1 CPS AI Model. A multi-organ foundation IHC AI model was developed (Mindpeak, Hamburg, Germany) and adapted to detect PD-L1 expression by CPS methodology in GC/GEJC/EAC biopsy tissues. Pretraining of the neural network (NN) PD-L1 AI model employed 1.4 million annotations from 518 multi-organ biopsy cases sourced from 16 labs. Refinement of cell annotations was accomplished through construction of a PD-L1 CPS specific decision tree (Gastric Cell Atlas) to reconcile heterogenous histology and complex features. Gastric specific manual annotations were determined from 212 GC/GEJC/EAC biopsy cases (28-8 IHC stained WSIs) sourced from 11 labs, 7 scanners and inclusive of 406,867 GC/GEJC/EAC cells. Finetuning of the Gastric PD-L1 CPS AI Model was performed by using multi-head knowledge distillation. Patch analysis was employed stepwise to identify PD-L1 tissue image patches, then tumor segmentation (red invasive tumor, blue tumor-associated immune cells), then cell identification and finally cell classification. For CPS, single cell classification was defined for PD-L1 positive TC (posTC, red), PD-L1 negative TC (negTC, yellow) and PD-L1 positive IC (posIC, purple). Total tumor cells (TC) were calculated from posTC plus negTC counts. CPS equals the number of posTC, plus the number of posIC, divided by total viable TC, multiplied by 100 and in all cases was mathematically derived. WSI analysis by Gastric PD-L1 CPS AI Model as final output. **b** Gastric Cell Atlas. Three pathologists with deep expertise in PD-L1 staining interpretation reviewed 31 cases of GC/GEJC/EAC with challenging histology and complex cellular features including at least one of the following criteria: overlapping cells; faint, shared, and/or granular staining; presence of tumor associated mononuclear inflammatory cells; indiscernible tumor or non-tumor cells; presence or absence of tumor invasiveness; and accurate quantification of cell numbers. Decision criteria were devised for cell type, cell number and PD-L1 positivity. Principles from the Gastric Cell Atlas manual annotations were summarized and along with the Gastric Cell Atlas Decision Tree were provided to pathologists involved in the training phase of the Gastric PD-L1 AI model. **c** Gastric PD-L1 CPS AI Model Performance. Preliminary performance of the Gastric PD-L1 CPS AI Model was tested on 55 gastric cancer cases, independent from training set cases, sourced from 3 labs and 3 scanners and stained with PD-L1 IHC 28-8 pharmDX assay. The WSIs were reviewed by 2 expert pathologists and assigned a CPS ≥ 5 PD-L1 reference consensus score. Left, agreement of Gastric PD-L1 CPS AI Model against reference consensus score before Gastric Cell Atlas decision rules (81.8%) and after (96.4%). Right, Cohen's Kappa before Gastric Cell Atlas decision rules (0.62) and after (0.91).

model, the finetuned Gastric PD-L1 CPS AI Model showed improved concordance against the reference consensus increasing agreement from 81.8% (95% CI, 68.6%, 94.3%) to 96.4% (95% CI, 88.6%, 100%) ($p < 0.05$, McNemar's test) and Cohen's Kappa κ from 0.62 (95% CI, 0.36, 0.86) to 0.91 (95% CI, 0.71, 1.0) ($p < 0.05$, Bootstrapping) (Fig. 1c). To validate the performance of the Gastric PD-L1 CPS AI Model, we retrospectively compared it to the PD-L1 CPS readouts of 12 expert gastro-intestinal pathologists on an independent dataset of 97 GC biopsies, where we previously reported high variability when using manual scoring[48]. Correlation of the outputs from the PD-L1 CPS AI Model with pathologists in terms of concordance correlation coefficient (CCC) were higher (0.59; 95% CI, 0.49, 0.67) than the average correlation among pathologists' manual scores (0.56; 95% CI, 0.41, 0.62), but without statistical significance, as well as higher than for the original IHC foundation model with statistical significance ($p < 0.001$, Bootstrapping) (Supplementary Fig. S2). At a PD-L1 CPS ≥ 5 cut-off, the concordance between PD-L1 AI scores and pathologists' manual scores was higher (Cohen's κ = 0.46; 95% CI, 0.37, 0.54) than the average concordance among pathologists' manual scores (Cohen's κ = 0.39; 95% CI, 0.29, 0.48) with statistical significance ($p < 0.05$, Bootstrapping) as well as higher than for the original IHC foundation model ($p < 0.001$, Bootstrapping). By validating AI performance on two independent gastric cancer biopsy datasets, we found that performance of the Gastric PD-L1 CPS AI Model was markedly improved compared to the IHC foundation model and was non-inferior to human PD-L1 CPS IHC readouts.

Trust facilitated through augmented reality

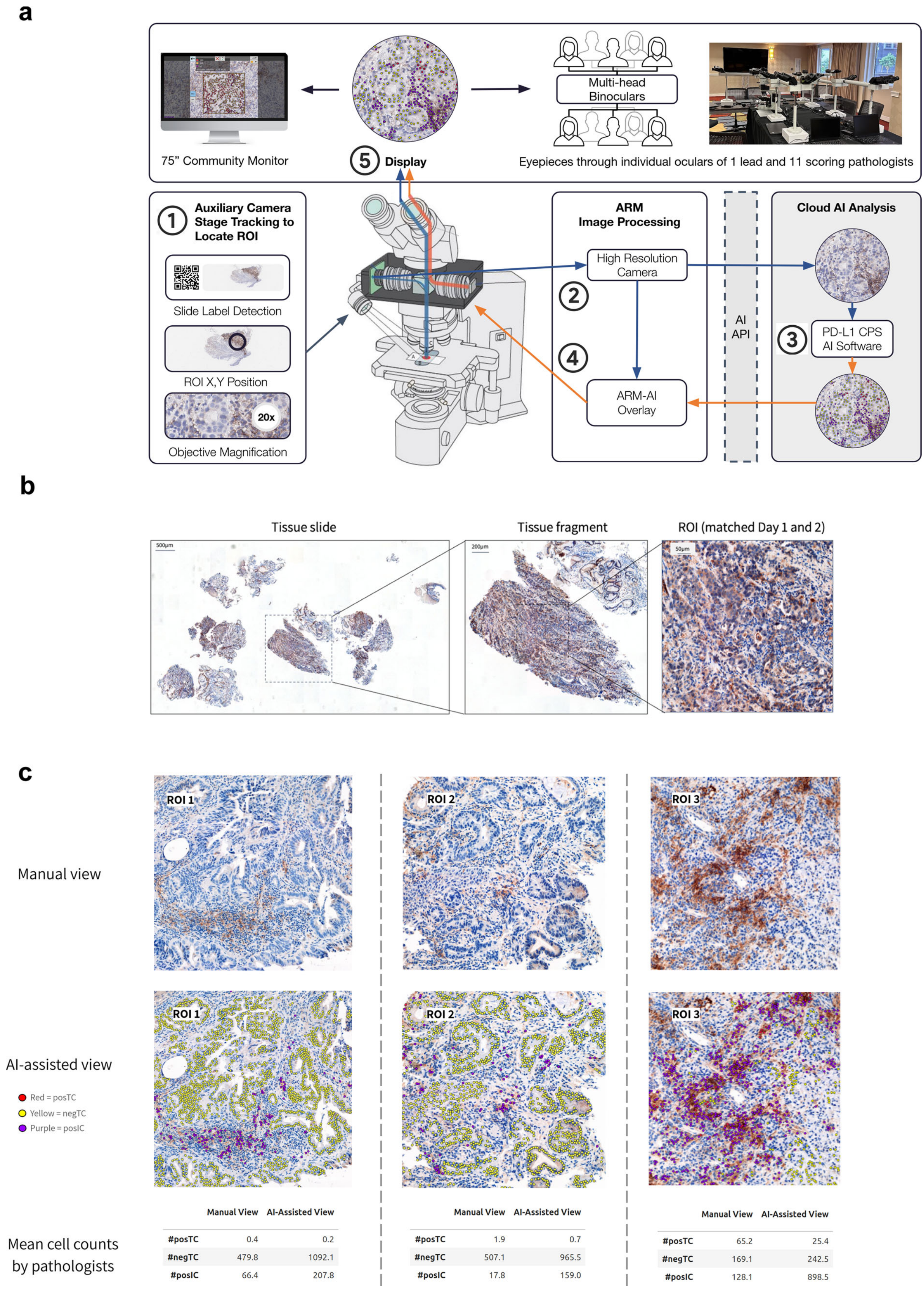
Next, we wanted to study the influence of our Gastric PD-L1 CPS AI Model as an AI-assistance tool for pathologists scoring PD-L1 CPS on glass slides. One of the leading reasons for the low uptake of published AI algorithms in routine clinical practice is the lack of appropriate assessment in clinical grade settings. Even with FDA cleared and Conformité Européenne (CE) certified AI systems, pathologists tend to require a level of trust, possibly gained through hands-on experience, prior to approving integration into their practice.

Using the ARM-AI system we investigated the utility and trustworthiness of the Gastric PD-L1 CPS AI Model in a controlled manner. Twelve expert pathologists, 1 lead pathologist who did not score samples and 11 scoring pathologists, participated in a two-day experiment to evaluate the same ROIs in real-time comparing manual vs AI-assisted scoring on an independent set of 35 GC/GEJC/EAC biopsy cases. For this, we devised an ARM-AI system using an Olympus BX53 light microscope retrofitted with an Augmentiqs ARM unit that was connected to a local computer workstation for ARM image processing. Multi-person viewing of the same ROIs was made possible by attaching a 16-unit binocular head system to the light microscope (Fig. 2) with simultaneous display on a 75-inch monitor for community viewing. The lead pathologist maneuvered the ARM-AI system to ensure the lighting conditions were optimal and the images, both through the microscope and the digital monitors, were clear and sharp. Live feed from the auxiliary camera enabled stage tracking to locate each ROI through x, y coordinates at 20x magnification ensuring ROIs matched from day 1 to day 2. An overlay of the PD-L1 CPS AI analysis could be triggered by the lead pathologist for the tissue displayed in the FOV by the push of a button. The ARM-AI system was programmed to send the respective tissue image to the API of a cloud-based AI server operating the Gastric PD-L1 CPS AI Model.

The Gastric PD-L1 CPS AI Model analyzed tissue images by capturing sufficiently large viewing areas, including peripheral margins, beyond the core ROIs. The AI outputs consisted of (a) individual cells with their coordinates, (b) cell type and (c) corresponding CPS score. Visually through the oculars, pathologists were able to see the PD-L1 CPS AI analysis as a real-time digital overlay superimposed on the glass slide containing the tissue sample, including the visualization of PD-L1 positive tumor cells (posTC, red), PD-L1 negative tumor cells (negTC, yellow), and PD-L1 positive immune cells (posIC, purple). This augmented reality visualization of the Gastric PD-L1 CPS AI Model outputs could be toggled by the lead pathologist, allowing the 11 scoring pathologists to compare AI-assisted views with unassisted views of the original tissue samples. The pathologists' key experimental task was to view each of the predefined ROIs independently and produce PD-L1 CPS outputs (cell counts and CPS score) either manually on day 1 or with AI-assistance enabled by the ARM-AI system on day 2. Each of the 35 independent biopsy cases contained one pre-marked ROI. ROIs were photographed and the coordinates recorded so that scoring pathologists could view each of the ROIs on multiple occasions. Scoring pathologists were asked to independently assess all 35 ROIs and provide individual cell counts of posTC, negTC and posIC. The final CPS score for each ROI was calculated and presented to the pathologists for either agreement or disagreement. If pathologists did not agree with the derived CPS score, they could 'gut-check' the derived CPS and change the underlying cell counts, but not the CPS directly. Thereby, final CPS outputs, whether manual or AI-assisted, were governed individually by each of the 11 scoring pathologists.

Impact of assistance on pathologist performance

Overall, AI-assistance increased the pathologists' awareness of absolute cell counts most likely by improving their visual perception when looking at the FOV. posTC, negTC, and posIC cell counts revealed that pathologists were reporting more cells with AI-assistance on day 2 in comparison to manual counts on day 1 (Fig. 3a), with a statistically significant increase ($p < 0.01$) of approximately 2.5-fold for negTC, 1.5-fold for posTC, and 4.9-fold for posIC (Fig. 3b). Notably on day 2, pathologists observed qualitatively that



Mean cell counts by pathologists

ROI 1

	Manual View	AI-Assisted View
#posTC	0.4	0.2
#negTC	479.8	1092.1
#posIC	66.4	207.8

ROI 2

	Manual View	AI-Assisted View
#posTC	1.9	0.7
#negTC	507.1	965.5
#posIC	17.8	159.0

ROI 3

	Manual View	AI-Assisted View
#posTC	65.2	25.4
#negTC	169.1	242.5
#posIC	128.1	898.5

AI-assisted scoring prompted them to become aware of the actual number of viable tumor cells within each of the ROIs.

Intraclass correlation coefficient (ICC) was used to assess agreement among pathologists for manual scoring on day 1 vs ARM-AI-assisted scoring on day 2. Analysis of scores for CPS components revealed poor to fair agreement among pathologists when scoring manually but statistically significant improvement ($p < 0.01$) in agreement with AI assistance. ICCs summarized in Table 1 include 0.38 (95% CI, 0.31-0.45) manual vs 0.90

Fig. 2 | An ARM-AI framework designed to evaluate the 'trustworthiness' of AI algorithms for use in AI-assisted PD-L1 CPS scoring. a Overview of the ARM-AI framework and experimental design. An Olympus BX53 light microscope (Evident, Hunt Optics & Imaging Inc., Atlanta, GA) was retrofitted with an ARM unit (Augmentiqs, D.N. Misgav, Israel) that interfaced with the Gastric PD-L1 CPS AI Model (Mindpeak, Hamburg, Germany) and extended with 16 binocular heads for direct viewing of glass slides by scoring pathologists. The ARM-AI unit was connected to a computer workstation for ARM image processing and cloud-based AI analysis. Community viewing was available on a 75-inch monitor. Live feed from the auxiliary camera enabled stage tracking to locate ROI through x,y coordinates at 20x magnification. The lead pathologist maneuvered the ARM-AI unit and selected the ROIs but did not participate in the scoring. 11 scoring pathologists were tasked to submit numeric values for posTC, negTC and posIC through individual computers using a web-based data capture tool (Mindpeak, Hamburg, Germany); Scoring pathologists evaluated 35 individual ROIs on unique biopsy cases on day 1 using manual scoring and on day 2 using AI-assistance. On day 2, toggling between normal viewing and the ARM-AI overlay was controlled by the lead pathologist to ensure the scoring pathologists had equivalent time per sample. Each ROI was wholly contained within a single field of view with day 1 and day 2 x,y coordinate matching. **b** Description of data set. Imaging was performed at 20x (0.25 mpp) magnification and 1 z-plane. Glass slides of FFPE primary tumor GC/GEJC/EAC biopsies stained with H&E and PD-L1 IHC 28-8 pharmDx assay as previously executed and reported by Robert et al., were inspected for difficult to interpret ROIs based on a PD-L1 CPS range (zero, low, medium, high) and complex morphology. 35 biopsy cases were selected containing one ROI per case. Left, representative biopsy showing H&E staining; Middle, tissue fragment of choice showing case #34. Right, actual ROI, with areas of immune cell infiltration, evaluated manually on day 1 and with AI-assistance on day 2. **c** Representative ROIs as seen for manual scoring (day 1) vs AI-assisted scoring (day 2). For CPS, single cell detection for PD-L1 posTC (red), PD-L1 negTC (yellow) and PD-L1posIC (purple). For each ROI, average cell counts across 11 pathologists were calculated showing the potential impact of AI-assisted scoring on improving the estimation of cell populations (negTC and posIC) less familiar to pathologists and/or improving the estimation of cell populations in large numbers (posTC).

(95% CI, 0.85–0.95) AI-assisted for TC; 0.40 (95% CI, 0.31–0.47) manual vs 0.91 (95% CI, 0.85–0.95) AI-assisted for negTC; 0.27 (95% CI, 0.22-0.37) manual vs 0.65 (95% CI, 0.56-0.71) AI-assisted for posTC; and 0.48 (95% CI, 0.26–0.56) manual vs 0.70 (95% CI, 0.60–0.75) AI-assisted for posIC. Interobserver concordance at the clinically relevant PD-L1 CPS ≥ 5 cut-off was analyzed. Agreement among pathologists by Fleiss' Kappa showed overall improved agreement from 0.52 (0.40–0.65) with manual scoring on day 1 to 0.76 (0.62-0.87) with AI-assisted scoring on day 2 with statistical significance ($p < 0.01$, Bootstrapping of differences) (Fig. 3c). Ultimately, biomarkers such as PD-L1 are used to select patients and possibly predict response to checkpoint inhibitor treatment. Categorizing patients accurately and reproducibly is vital for optimizing treatment outcomes and minimizing treatment adverse effects. To this end, we investigated the calling of PD-L1 CPS ≥ 5 positivity for each ROI manually and with AI-assistance. The number of PD-L1 CPS ≥ 5 positive ROIs was assessed by using the median score among all 11 individual scores from each pathologist as consensus. AI-assisted scoring prompted pathologists to identify 11 additional cases as PD-L1 CPS ≥ 5 positive (Fig. 3d). Pathologists identified 15 cases as PD-L1 CPS ≥ 5 positive with manual scoring on day 1 compared to 26 cases as PD-L1 CPS ≥ 5 positive by AI-assisted scoring on day 2, corresponding to a 31% increase of positive cases found when using AI-assistance. While using AI assistance, some pathologists changed their assessment from PD-L1 positive to negative (see Fig. S1 in Supplementary Materials). In one case with significant discordance when manually scoring, AI assistance led to full consent among all pathologists to categorize the case as negative.

As treatment outcomes were not available, we evaluated the significance of our findings using the Observers Needed to Evaluate Subjective Tests (ONEST)[30] plot analysis, which estimates the number of pathologists needed to ascertain a reliable concordance estimate. ONEST plots (Fig. 4) revealed that 8-9 pathologists were required to yield reliable PD-L1 CPS ≥ 5 estimates. Comparison of ONEST plots between manual scoring on day 1 and AI-assisted scoring on day 2 revealed that using manual scoring any two pathologists agree on 77% of cases, while with AI-assisted scoring any two pathologists agree on 91% of the cases, demonstrating a 14% improvement in agreement with AI-assistance. When comparing agreement among all 11 scoring pathologists, agreement was achieved in 43% of the cases with manual scoring in comparison to 69% with AI-assisted scoring, demonstrating a 26% improvement in agreement when using AI-assistance. In summary, higher agreement among pathologists was achieved with the Gastric PD-L1 CPS AI Model, even when viewing was restricted to one 20x FOV.

Discussion

AI-based diagnostic pathology platforms are delivering unparalleled accuracy in cancer detection and, when fully integrated into real world practice, promise to shorten time to diagnosis and increase accuracy[5,49]. Developing clinical grade diagnostic IHC AI models for biomarker detection, however, requires a more careful consideration of the parameters used to train the algorithms. IHC biomarker outputs in real-world practice lack quantitative accuracy and assurance that patient level outcomes remain aligned with population level responses observed in clinical trials. Furthermore, there is a gross underappreciation of the complexity pathologists face in routine practice[50–52]. Factors including poor training, high case volume, complex scoring methods, nuanced biomarker cut-offs, biomarker positivity on non-tumor cells, and variable validation approaches have added difficulty to an already arduous task, essentially leaving testing labs and pathologists to shoulder accountability unaided[18,20,53,54]. Fully aware of IHC discrepancies, pathologists are skeptical about automated AI models and seeking to understand their performance thresholds before fully embracing them as 'first read' solutions[36]. Mindful that future end-to-end AI models should be highly accurate and that human pathologists will retain legal and moral authority over clinical decisions, we sought to understand the effects of establishing trust between AI and pathologists.

In this study, through augmented reality capabilities, we demonstrated that active participation by pathologists in the training and deployment of a novel IHC PD-L1 CPS AI Model, unlike comparatively passive roles as annotators, resulted in mutual performance improvements and facilitated trust with pathologists. We observed that trust could be attributed to (1) allowing human agency, (2) preserving decision accountability, (3) provoking visual perception awareness and scoring behavior and (4) increasing decision confidence associated with difficult cases. Given the potential benefits of fully automated digital pathology systems, we believe that a pathologists-in-the-loop participatory role is a necessary intermediate step that will encourage increased adoption and improved performance of AI models.

Building on a multi-organ cell-based foundation model, our hybrid training approach used both self-supervised learning for recognition of general tissue structures and supervised learning to detect cellular and subcellular structures. Our choice of examining PD-L1 CPS ≥ 5 expression on gastroesophageal biopsies was based on our previous work[48] and other reports[55–57] describing the low concordance rates among pathologists when manually scoring PD-L1 CPS. Given the architectural complexity of gastroesophageal histology and the clinically relevant PD-L1 CPS ≥ 5 cut-off used to determine patient eligibility for checkpoint immunotherapy, we included ROIs with complex heterogeneity that would likely be ignored by both human pathologists and AI models. We first employed direct participation during the fine-tuning process by selecting expert pathologists, qualified as PD-L1 CPS trainers, to examine ROIs that resulted in ambiguous outputs by our PD-L1 CPS AI Model. After adjudication of the ROIs, we constructed a rules-based decision tree termed 'Gastric Cell Atlas' and provided it to additional pathologists employed as annotators (Fig. 1b). As a

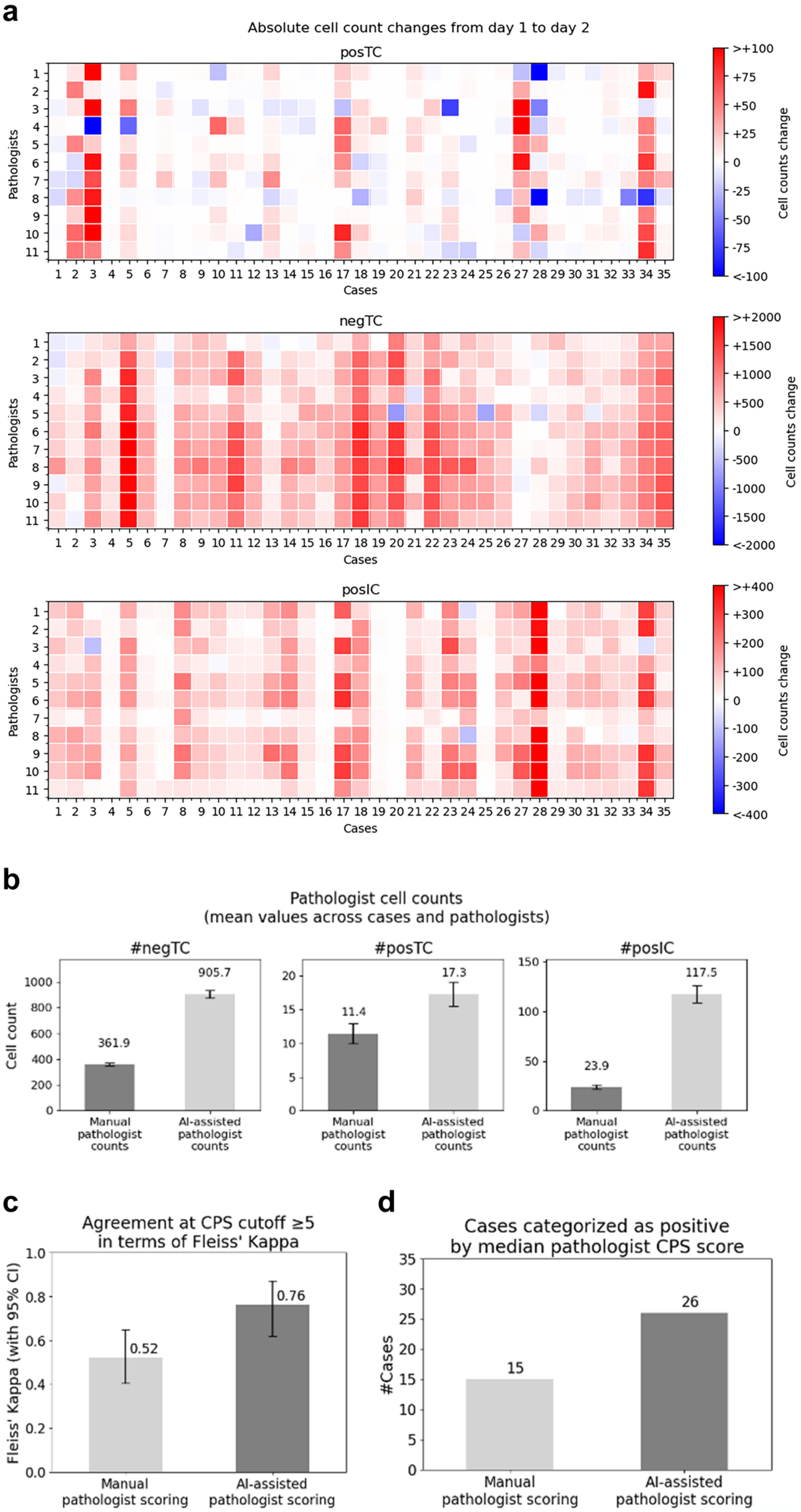


Fig. 3 | AI-assisted scoring dynamics indicate improved fidelity of visual perception and higher scoring agreement potentially leading to increased positive diagnoses. **a** Comparison of manual vs AI-assisted scoring dynamics among 11 pathologists by cell type. Each square represents the change in scoring from manual (day 1) to AI-assisted (day 2) by each of the 11 pathologists and color gradations represent absolute cell counts where blue/red indicates less/more cells on day 2. Heatmaps: Top, changes in posTC; Middle, changes in negTC; Bottom, changes in posIC. **b** AI-assisted scoring increased awareness of absolute cell counts and improved fidelity of visual perception among pathologists. Tally of absolute cell counts by cell type: Left, negTC manual 361.9 (349-375) AI-assisted 905.7 (878-933); Middle, posTC manual 11.4 (10-13) AI-assisted 17.3 (15-19), Right, posIC manual 23.9 (22-26) AI-assisted 117.5 (109-127). **c** Agreement at PD-L1CPS ≥ 5 as evaluated by Fleiss' kappa. Analysis of agreement among pathologists by Fleiss' Kappa on day 1 using manual scoring Fleiss' Kappa 0.52 (0.40-0.65) compared to day 2 using AI-assisted scoring Fleiss' Kappa 0.76 (0.62-0.87) showing overall improved agreement with AI-assistance. **d** AI-assisted scoring prompted pathologists to identify 11 additional cases as positive for PD-L1 CPS ≥ 5. Consensus CPS ≥ 5 was calculated as the median score from the 11 individual pathologists. Manual scoring (day 1) identifies 15 cases of PD-L1 CPS > 5 vs 26 cases of PD-L1 CPS ≥ 5 by AI-assisted scoring (day 2) resulting in a 31% increase of positive cases.

result, our PD-L1 CPS AI Model, trained in total on over 1.8 million annotations from 533 unique biopsies obtained from 16 labs, detected PD-L1 CPS expression on cell structures with 96.4% agreement against a pathologists' derived consensus score (Fig. 1c). To facilitate immersive participation for pathologists, we devised an ARM-AI framework that was operationalized through a multi-head light microscope system accommodating 11 pathologists who were tasked with scoring 35 ROIs on glass slides manually on day 1 and with AI-assistance on day 2 (Fig. 2a). The ARM-AI framework enabled real-time seamless integration of AI into routine workflows using glass slides and leveraged an environment where pathologists could experience AI-assistance to evaluate the trustworthiness of the PD-L1 CPS AI Model outputs.

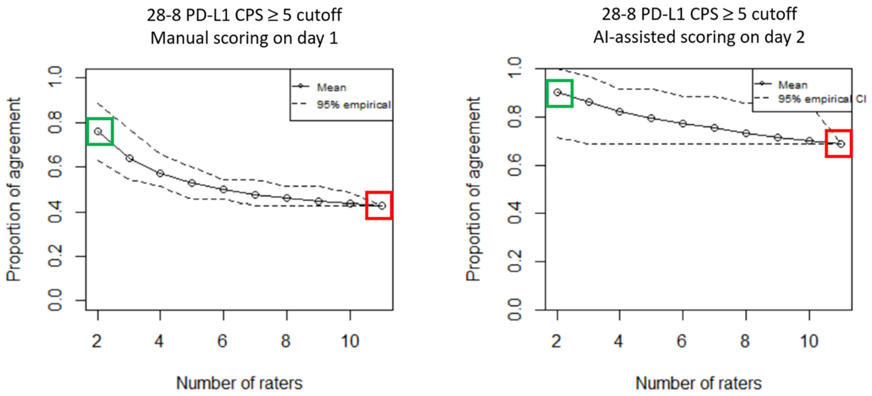


Fig. 4 | Observers needed to evaluate subjective tests (ONEST) plots. Overall percent agreement between pathologists on the y-axis vs the number of observers on the x-axis for manual scoring on day 1 (left panel) and AI-assisted scoring on day 2 (right panel). PD-L1 28-8 CPS ≥ 5 cut-off (solid line) and CIs (dotted lines). ONEST analyses indicate that AI-assisted scoring increased overall interobserver agreement. Agreement among any 2 raters using manual scoring is achieved in 77% of the cases (green box left panel) vs 91% of the cases with AI-assisted scoring (green box right panel) resulting in a 14% improvement. Agreement among 11 pathologists using manual scoring is achieved in 43% of the cases (red box left panel) vs 69% of the cases with AI-assisted scoring (red box right panel) resulting in a 26% improvement.

Table 1 | AI-assisted scoring results in increased agreement among pathologists across cell type categories

	Total viable TC	Total PD-L1 negTC	Total PD-L1 posTC	Total PD-L1 posIC
ICC Manual (95%CI)	0.38 (0.31–0.45)	0.40 (0.31–0.47)	0.27 (0.22–0.37)	0.48 (0.26–0.56)
ICC AI-Assisted (95%CI)	0.90 (0.85–0.95)	0.91 (0.85–0.95)	0.65 (0.56–0.71)	0.70 (0.60–0.75)

Interclass correlation coefficient (ICC) was used to assess scoring agreement among pathologists for manual (day 1) vs AI-assisted (day 2) categorical assessment for PD-L1 CPS ≥ 5. The ICC ranges from 0 to 1, with values close to 1 indicating higher agreement among pathologists.

We found that pathologists' participation not only improved AI fine-tuning but significantly improved their own performance, as a group and individually. Agreement, a measure of scoring reliability when ground truth is absent, increased among pathologists across PD-L1 CPS categories with AI-assistance, showing higher ICC for PD-L1 negTC, posTC and posIC (Table 1). The most revealing improvement, however, was in the category of total viable tumor cells. Even within a limited viewing area restricted to one 20x FOV, considered manageable compared to a full WSI, pathologists grossly underestimated the total number of viable tumor cells by at least 60%. Similarly, even though lower in number, they also underestimated the presence of PD-L1 posTC and PD-L1 posIC. Importantly, on day 2 using AI-assistance, pathologists became aware of their tendency to underestimate cell structures, something that they previously ignored or felt had negligible impact. Overall, AI-assistance improved agreement of PD-L1 CPS ≥ 5 (Fleiss' Kappa 0.52 vs 0.76, Fig. 3c), even though we had assumed that agreement would be much higher given that all the scoring was done within one 20x FOV. Of the 35 ROIs scored, 15 cases were deemed PD-L1 CPS ≥ 5 positive by manual scoring and 26 cases deemed positive with AI-assistance. By extrapolation, the impact of manual scoring variability errors indicates that pathologists may be under calling PD-L1 CPS positivity on gastric cancer biopsies.

Agreement analysis using Observers Needed to Evaluate Subjective Tests (ONEST)[30] demonstrated that AI-assistance resulted in agreement between any two pathologists on 91% of the cases, and among all 11 pathologists on 69% of the cases. Agreement among pathologists as a measure of scoring reliability of subjective IHC staining interpretation is an imperfect proxy with low concordance rates underscoring the need for accurate AI models in routine pathology practice. Although we designed our experiment to achieve as much agreement as possible, by restricting cell counting to one 20x FOV, our data show that staining interpretation variability is multi-layered and not easily eliminated. The observed increased agreement among pathologists using AI-assistance is likely to be even more relevant when evaluating the whole tissue slide.

Regarding trust, we observed that pathologists needed hands-on familiarity through direct participation to become aware of their own baseline scoring behavior and were themselves surprised by the utility of AI-assistance in helping them accurately capture the contents in the FOV. Furthermore, the ARM-AI framework allowed us to pinpoint that human agency is a key component of decision accountability and that both aspects are critical for establishing the trustworthiness of IHC AI models. Although pathologists readily admit that they already work as a type of 'black box' themselves, often not able to systematically break down their decision steps, they are reluctant to accept AI tools based on published research without hands-on familiarity[42]. They want to understand how the AI model produces outputs and evaluate the AI model's usefulness in their practice[42]. We noted similar sentiments in our group of 11 scoring pathologists, with acceptance of the IHC AI model when appropriately validated and FDA approved but not on published data alone. Our pathologists unanimously agreed that evaluation of PD-L1 expression on non-tumor cells remains a difficult task, and that pathologists generally score PD-L1 expression with low confidence.

Our study has several limitations. First, the nature of the ARM-AI framework limited our analysis to one ROI per biopsy case. The ROIs were selected from 35 unique biopsy cases that were previously evaluated in a separate study where we documented poor concordance among pathologists when scoring PD-L1 CPS manually[48]. Since we could not pinpoint the root causes of variability through concordance approaches, even on seemingly straightforward tasks such as viable tumor cell count, we opted to design an experiment that would allow us to evaluate pathologists' scoring behavior against the PD-L1 CPS AI Model and simultaneously gauge the AI model's trustworthiness. By limiting our evaluation to one ROI per biopsy, we eliminated the multiple sources of variability present when pathologists score WSIs. When calling PD-L1 positivity, however, pathologists sign out cases based on their holistic evaluation of the entire biopsy section not just specific ROIs, no matter how relevant an individual ROI might appear. Second, our study relied on subjective concordance estimates comparing manual vs AI-assisted scoring outputs. While not optimal, concordance is still the key methodology by which IHC scoring reliability is evaluated since ground truth is not available. While most validation studies evaluate concordance between 1-3 pathologists, our study employed 11 pathologists to ensure we sufficiently captured the variability associated with subjective tests. Ultimately, validation against clinical responses may be needed to overcome the absence of absolute ground truth, however, formal validation of the PD-L1 CPS AI Model nor evaluation of its clinical utility were within the scope of our experiment. Instead, we set out to observe how pathologists interacted with AI and gain insight on the levers of trust and adoption. Finally, the classification of cells as tumor and non-tumor was based on morphological features and did not use specific IHC parameters such as keratin expression. Although this mimics clinical practice, the positivity of misclassified cells as tumor or non-tumor cannot be excluded, particularly challenging for cells with signet ring morphology.

New AI-assistance tools, including our PD-L1 CPS AI Model with integrated ARM, must undergo extensive clinical validation across relevant indications and all available cut-offs (e.g. PD-L1 CPS ≥ 1, ≥5, ≥10) to ensure robust accuracy in routine clinical practice applications. ARM specifically allows the pathologist to more directly interact with annotations, measurements, and AI overlays by significantly enhancing visualization capabilities. Such tools need to be seamlessly integrated within existing pathology infrastructures. Initially, pathologists may resist adoption because they might not be sufficiently familiar with the technology or because they carry a healthy skepticism toward AI models. Hands on experience is crucial for building trust and maturing the technology sufficiently to navigate regulatory standards. Ultimately, the lessons learned can be translated to applying the AI models directly on digital images.

In summary, our findings frame a potential roadmap for building trustworthy IHC AI models and lay out the initial design features of an on-demand digital pathology assistant that can be integrated into real-world pathology workflows. The impact of our work lies in two key areas. The first is to bridge fully automated and fully manual scoring using augmented reality approaches that can be deployed on systems familiar to pathologists. We found that this is a necessary intermediate step to increase trust and adoption as it builds the pathologists' experience with AI and simultaneously improves the performance of pathologists and IHC AI models. The second is to fuel a shift into increasingly quantitative IHC methods that can open the door to truly predictive biomarker discovery as precise clinical decision tools to match patients to therapies.

Methods

Gastric Cell Atlas for AI development

A single-cell annotation guideline was established to recognize and annotate both tumor cells (TC) and immune cells (IC) accurately. 31 ROIs were analyzed from 31 unique gastric cancer specimens using a 20x magnification. To avoid fatigue and ensure feasibility for pathologists, the number of ROIs was determined by what was assessable in terms of individual cells within a day. Tissues were procured from three different institutions to get diversity in clinical samples: Institute for Hematopathology Hamburg, Hamburg, Germany; ($n = 12$); Discovery Life Sciences, Kassel, Germany; ($n = 10$), and Ziekenhuis Netwerk Antwerpen, Antwerp, Belgium; ($n = 9$). These tissues were stained with the PD-L1 IHC 28-8 pharmDx assay according to the manufacturer's instructions at three different locations, namely at CellCarta, Antwerp, Belgium; ($n = 12$); Discovery Life Sciences, Kassel, Germany; ($n = 10$); and Ziekenhuis Netwerk Antwerpen, Antwerp, Belgium ($n = 9$). Slides were digitized using either an Aperio GT 450 (Leica Biosystems, Wetzlar, Germany), or a PANORAMIC 250 Flash (3DHISTECH, Budapest, Hungary) scanner. ROIs were chosen by S.B., K.D., and P.F. and contained at least 10–30 cells. The range of 10–30 cells was chosen to prevent fatigue. ROIs were selected as representative regions where single-cell level scoring was difficult due to one or more of the following five criteria: (1) interpretation of faint, shared, and/or granular staining; (2) inclusion of tumor-associated mononuclear inflammatory cells; (3) discrimination of tumor from non-tumor cells; (4) presence or absence of tumor invasiveness; (5) interpretation of cell numbers.

To ensure accurate classification of individual cells, three pathologists with experience in PD-L1 staining interpretation (S.B., J.R., H.S.) convened at Mindpeak, Hamburg, Germany to establish gastric cancer PD-L1 CPS staining interpretation decision rules. T.L., F.F., P.F., G.L.K., M.K., and K.D. were present and participated in the staining interpretation discussion, with final decisions approved by S.B., J.R., and H.S. The study was conducted in two phases. During phase 1, the pathologists were individually presented with the 31 ROIs in consecutive, randomized order via a web-based annotation software (Mindpeak, Hamburg, Germany). The ROIs were independently reviewed in-person by the three pathologists and individual cells were classified into 5 categories (positive/negative TC, positive/negative IC, and unclear). Fibroblasts/endothelial cells were excluded. Pathologists used a recommended stepwise process for identifying single cells for AI scoring in GC, as previously described by Rüschoff et al.[58]. The three pathologists used identical equipment, under the same conditions. The pathologists had an overview of the whole biopsy slide and the regional tissue context in the tool. To simulate an AI algorithm analyzing the IHC slides, the pathologists were asked to classify cells based on IHC images only. During phase 2, the three pathologists gathered in front of a large computer monitor displaying their individual solutions for each ROI from phase 1 to discuss these. The solutions were presented in a randomized and anonymized manner to avoid bias. The pathologists discussed ambiguities in the existing annotation processes, analyzing all cells that were not unanimously classified in phase 1. Discrepancies were resolved by consensus, and output was used to inform the roadmap. With especially difficult cases, pathologists were allowed to request corresponding H&E-stained images to resolve queries, but this was only used in six instances to assess the extent of immune cell infiltration into the tumor. The pathologists jointly defined annotation principles and rules and decisions were noted by the study team. A consensus solution was developed for each ROI and was recorded separately by the study team under the pathologists' supervision. The pathologists reviewed the guidelines formulated in the discussion to ensure their general applicability and incorporation in the single-cell annotation roadmap presented here. The resulting Gastric Cell Atlas contained rules for the following aspects: (1) Positivity Interpretation: Scoring specimens with extensive granular (non-linear) membrane staining should be avoided. If most of the membrane staining is granular, then the specimen should be considered as non-assessable. In case of difficulty differentiating light brown from gray tumor membrane staining, clearly negative tumor cells in the surrounding area can be used as control/reference. For tumor and immune cells expressing both cytoplasmic and membrane staining, any strong linear membrane staining that can be distinguished from cytoplasmic background should be considered as positive. Immune cells with faint borderline membrane/cytoplasmic staining should be reassessed at lower magnification (equivalent to 20x). Not only positive and negative tumor cells, but also positive and negative immune cells should be annotated. If positivity or cell class cannot be determined unanimously, a cell should be marked as "unclear" and be excluded from assessment. The principle of assessing relevant cell classes could be generalized to all membrane biomarkers and all required cell classes. (2) Shared Positivity: When two or more tumor cells (or tumor and immune cells) share convincing partial linear membrane positivity, only the cell with staining following the contour of the cell membrane is to be considered positive. Stromal staining in rare foci can be overwhelming and can obscure the cellular outlines and characteristics. Tumor or inflammatory cells staining in these regions is not to be counted as it is not cell specific. (3) Tumor vs Non-Tumor Classification: Both cellular and nuclear morphology and architectural context enable a good distinction of invasive viable tumor cells from non-invasive/non-tumor cells in PD-L1 scoring. Nuclear size and shape are not always reliable criteria particularly when the tumor cells are small and have indistinct nucleoli. Plasmacytoid tumor cells can also be difficult to assess. In such cases, other morphological criteria and surrounding context should be considered. Additionally, pathologists can rely on the H&E stain to distinguish epithelial cells (normal form and tumor) as well as stromal cells (immune and non-immune). Since this may not be possible, however, we chose to mark a cell as "unclear" if its class could not be determined. Such cells can then be excluded in the training data for AI development, resulting in less ambiguous training and increased model stability. Particularly, “unclear” cells were excluded for the development of our Gastric PD-L1 CPS AI Model. (4) Invasiveness: Any epithelial cells that are part of carcinoma in situ (CIS), glandular non-invasive dysplasia as well as the associated mononuclear inflammatory cells (MICs) were excluded from the score. Histology criteria of invasive cancer include tubule / papillary formation, tumor budding, mitotic activity and nuclear pleomorphism. In regions with borderline criteria, context is to be assessed at a lower magnification and if it is cancer associated it must be considered invasive. 5) Cell Count: Epithelial cell maps are helpful for distinguishing individual cells. Care should be taken to ensure that multinucleated tumor and inflammatory cells (i.e. cells with a single contour) are counted as one cell. In the case of overlapping cells, those with one cell-contour (e.g., solitary membrane positivity) should be counted as one cell. Groups of cells with multiple contours that distinctively follow their respective borders are to be considered individual cells. If uncertain, the least number of overlapping cells should be counted. 6) Tumor associated inflammatory cells: Tumor-associated mononuclear inflammatory cells (MICs) are directly associated with tumor response and should be assessed using 20x magnification. Any pre-existing non-tumor related inflammations and lymphoid follicles must be excluded. In case of peritumoral retraction artefacts, the positive mononuclear inflammatory cells at both sides of the artefact should be considered (reconstructed) as tumor-associated and therefore included in the score.

PD-L1 CPS AI Model and development

AI software, based on deep learning with convolutional neural networks, for PD-L1 CPS quantification in GC/GEJC/EAC samples was developed at Mindpeak and termed Gastric PD-L1 CPS AI Model. It detects cells in IHC stained tissue, distinguishes CPS-relevant tumor and immune cells from other cells, and subsequently calculates cell counts and CPS scores. For a given image, the software analysis proceeds along a pipeline of image processing steps, including neural network models for tissue segmentation and cell detection. First, tissue is detected and distinguished from background using computer vision techniques based on image gradients and color distributions. Then, the tissue is segmented into invasive tumor, tumor-associated immune cell areas, and other non-CPS-relevant tissue areas, using a neural network. In the next step, using relevant tissue segments, cells are detected and classified into either one of the three cell classes relevant for CPS (posTC, negTC, posIC), or as other, irrelevant cell, using a second neural network. Cell detection is accomplished by predicting the probability of each pixel as part of a cell and then clustering neighboring pixel groups with high probability scores into distinct cell entities. Cell classification is achieved by predicting every pixel in a cell class (e.g. tumor and immune) and assigning detected cells after cell detection to classes. The employed neural network models consist of 25 convolutional layer blocks with rectified linear unit activation functions and batch normalization. The neural networks underlying the AI model had been trained in an AI learning phase before the respective experiments. During the experiments, neural networks models were locked and not adapted. The final Gastric PD-L1 CPS AI Model was developed by finetuning a multi-organ foundation model for PD-L1 IHC assessment. The underlying foundation model had been built prior to this study on a large multi-organ dataset, involving PD-L1 cases from gastric cancer, urothelial cancer, esophageal cancer, and non-small cell lung cancer, using 1.05 million image patches of PD-L1 stained tissue and 1.4 m manual annotations.

The foundation model was finetuned to GC/GEJC/EAC using a large gastric cancer dataset. Data for AI development, both for pretraining and finetuning, did not overlap with the study data. To achieve robustness and consistency across changing preanalytical variables, this dataset involved data from 11 institutions, 7 scanners and 3 microscopes. This dataset was annotated by expert pathologists with cell and tissue annotations adhering to the rules of the previously developed PD-L1 Gastric Cell Atlas. A compendium for classifying individual cells in GC/GEJC/EAC biopsies was compiled by three expert pathologists, with guidelines for difficult assessment contexts. The resulting dataset with gastric-specific manual annotations used for AI development contained 406,867 manual cell annotations, resulting in a total multi-organ dataset with 1.8 m manual annotations. The dataset was randomly split into three partitions on a per case basis: training, validation, and test (70, 10, 20), stratified by institutions, scanners and microscopes. The primary metric for model development was the F1 score for cell detection and classification to balance positive predictive value and sensitivity. Model selection was performed by assessing this F1 score on the validation partition. Finetuning of the AI model to GC/GEJC/EAC was based on training neural networks with gradient descent using the Ranger optimization algorithm. To promote model robustness and generalization to unseen preanalytical contexts, a semi-supervised training approach was chosen including knowledge distillation via multiple neural network heads for cell detection and tissue segmentation.

ARM-AI framework

An Olympus BX53 light microscope (Evident, Hunt Optics & Imaging Inc., Atlanta, GA, USA) was retrofitted with an ARM unit (Augmentiqs, D.N. Misgav Israel). The ARM system was programmed to overlay Gastric PD-L1 CPS AI Model (Mindpeak, Hamburg, Germany) within the eyepiece and onto the current view of the sample in real-time (augmented reality). For full technical details on ARM, refer to Chen P.C., et al. [59] and the Augmentiqs home page: https://www.augmentiqs.com/digital-pathology-software-applications/. Multi-person viewing of the sample was made possible by attaching a 16-unit binocular head system to the main microscope (Fig. 2). The ARM-AI system was connected to a computer workstation for image processing and cloud-based AI analysis. Real-time images were displayed on a 75-inch monitor for community viewing and simultaneously available for multi-head binocular eyepieces for the scoring pathologists. Live feed from the auxiliary camera enabled stage tracking to locate ROI through x/y coordinates at 20x magnification. The lead pathologist maneuvered the ARM-AI system.

Experimental study design

GC/GEJC/EAC biopsy samples on glass slides were previously stained with H&E and with the PD-L1 IHC 28-8 pharmDx assay according to manufacturers’ instructions[48]. A single selected and pre-marked 20x FOV on each of 35 GC/GEJC/EAC biopsies on glass slides was used, with difficult to interpret ROIs including ambiguous identification of positively staining stromal cells, faint or variable intensity of staining, and difficulty in distinguishing membranous from cytoplasmic tumor staining. Each ROI was wholly contained within a single 20x FOV. The area was marked with a green Sharpie permanent marker. 11 pathologists, with a median of 10 years in clinical practice, were invited to perform PD-L1 CPS scoring on day 1 and day 2 of the experiment. 4 of 11 scoring pathologists (R.G., L.T., R.A., H.W.) had participated in a previous manual scoring PD-L1 CPS study[48]. The study proceeded in two phases. On day 1, pathologists manually scored the sample. Imaging was performed at 20x (0.25 microns per pixel) magnification in a 1 z-plane to capture a minimum of 100 viable tumor cells, following the CPS pharmDx interpretation manual (Agilent Technologies Santa Clara, CA, USA). Pathologists provided an exact numeric value for the three elements that comprise the CPS (PD-L1 posTC, negTC and posIC) into a web-based study software from individual computers (Mindpeak, Hamburg, Germany). The FOV from day 1 was captured by the study software. On day 2, the 35 cases were randomly presented in a different order from day 1, using the same FOVs as on day 1 by x/y coordinate matching. Pathologists examined cells using the same FOV as on day 1, but with AI-assistance. Cell counts and corresponding CPS scores were recorded in the study software. One pathologist (N.S.) noted that one biopsy possibly contained signet ring cells, difficult to visually identify and distinguish from macrophages, that were not verified by keratin staining. This circumstantial observation cannot entirely exclude the possibility that a small number of cells were misclassified. Toggling between normal viewing and the ARM-AI overlay was controlled by the lead pathologist to ensure the scoring pathologists had equivalent time per sample.

Ethics approval and consent to participate

The study was conducted in accordance and compliance with the BMS Bioethics Policy (https://www.bms.com/about-us/responsibility/position-on-key-issues/bioethics-policy-statement.html). The samples procured from the Mayo Clinic received IRB committee approval (#20-007665) in compliance with the institution's IRB review process (https://www.mayo.edu/research/institutional-review-board/overview) and conducted in accordance with the Declaration of Helsinki (purchased with funds provided by BMS). The samples procured from Discovery Life Sciences were processed by all applicable EU and US regulations as specified on the company's website (https://www.dls.com/resource-hub/faqs) and purchased by funds provided by BMS.

Statistical analysis

The intraclass correlation coefficient (ICC) was calculated to assess agreement among pathologists for cell counts for PD-L1 CPS relevant cell classes. The value of an ICC can range from 0 to 1, with 0 indicating no agreement among raters and 1 indicating perfect agreement among raters. Inter-pathologist CPS agreement was analyzed at a cutoff score of 5, representing the current clinically utilized cutoffs in gastric cancer, using agreement rates in percent as well as Fleiss' kappa. Fleiss' kappa is a measure for assessing the reliability of agreement between a fixed number of raters when assigning ratings to several items. It can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly. The concordance of the AI model with ground truth values in the validation before the study was measured by Cohen's kappa. Cohen's kappa is a statistic that measures the agreement between two raters beyond chance, accounting for the possibility of agreement occurring by chance. 95% confidence intervals for model performance were calculated using nonparametric bootstrapping with 1000 samples. Differences in means or medians for a continuous variable between two groups were assessed by a paired t-test. Categorical variables were compared using the McNemar test. 95% confidence intervals for concordance values were estimated by bootstrapping ($n = 1000$). The change in overall percent agreement (OPA) as a function of the number of observers was visualized using Observers Needed to Evaluate Subjective Tests (ONEST). Briefly, for any combination of pathologists, ONEST plots evaluate OPA using the proportion of tissue samples upon which all selected pathologists agreed. Calculation of OPA for all permutations of 12 pathologists resulted in 479,001,600 combinations, from which 100 were then randomly selected. The OPA was plotted against the number of pathologists, resulting in a graph that descends to a plateau. The plateau begins at the number of pathologists believed to be required to provide realistic concordance estimates. Data analyses and summaries were performed using R (version 3.6.1 under Windows 10) and Python 3.11. A list of R and Python packages used for analysis is available upon request.

Data availability

The human tissue samples used for these retrospective analyses remain anonymized. Tissue blocks, procured from the Mayo Clinic, were processed, sectioned, fixed on glass slides, scanned and converted to WSIs. Restrictions apply to data availability of anonymized patient data, and are thus not publicly available, owing to limitations imposed by data-sharing agreements with the data providers. In line with agreements and institutional policies, all data requests, whether for raw or processed data collected from glass slides or WSIs, should be made to the corresponding author (M.K.) and will be evaluated according to institutional and departmental policies to determine obligations to intellectual property or patient privacy compliance requirements.

Code availability

The deep learning framework used for AI development was PyTorch, which is available at https://www.pytorch.org/. The software used for basic image processing (OpenCV) is available at https://opencv.org/. The R software for data analysis is available at http://www.r-project.org, and the Python libraries used for data analysis, computation and plotting of the performance metrics (SciPy, Pandas, NumPy and MatPlotLib) are available at https://www.scipy.org/, https://pandas.pydata.org/, http://www.numpy.org/ and https://matplotlib.org/, respectively. The PD-L1 CPS AI Model used in this study is proprietary software protected by IP and owned by Mindpeak GmbH. The software operating the ARM system used in this study is proprietary, protected by IP and owned by Augmentiqs (D.N. Misgav, Israel).

Received: 13 January 2025; Accepted: 2 April 2025;
Published online: 12 May 2025

References

1. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
2. Malon, C. D. & Cosatto, E. Classification of mitotic figures with convolutional neural networks and seeded blob features. *J. Pathol. Inf.* **4**, 9 (2013).
3. Wang, H. et al. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *J. Med. Imaging* (Bellingham) **1**, 034003 (2014).
4. Paeng, K., Hwang, S., Park, S. & Kim, M. A Unified Framework for Tumor Proliferation Score Prediction in Breast Histopathology. In: Cardoso, M. et al. (eds) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS. Lecture Notes in Computer Science, vol 10553, (Springer, 2017).
5. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
6. Ehteshami Bejnordi, B. et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **318**, 2199–2210 (2017).
7. Beck, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108ra113 (2011).
8. Yuan, Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *J. R. Soc. Interface* **12**, 20141153 (2015).
9. Saltz, J. et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep.* **23**, 181–193 e187 (2018).
10. Corredor, G. et al. Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. *Clin. Cancer Res.* **25**, 1526–1534 (2019).
11. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
12. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
13. Kim, C. et al. Transparent medical image AI via an image-text foundation model grounded in medical literature. *Nat. Med.* **30**, 1154–1165 (2024).
14. Vaidya, A. et al. Demographic bias in misdiagnosis by computational pathology models. *Nat. Med.* **30**, 1174–1190 (2024).
15. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
16. Parikh, R. B., Teeple, S. & Navathe, A. S. Addressing Bias in Artificial Intelligence in Health Care. *JAMA* **322**, 2377–2378 (2019).
17. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med.* **5**, 48 (2022).
18. Magnani, B. & Taylor, C. R. Immunohistochemistry Should Be Regulated as an Assay. *Arch. Pathol. Lab Med.* **147**, 1229–1231 (2023).

19. Kujan, O. et al. Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: an attempt to understand the sources of variation. *Oral. Oncol.* **43**, 224–231 (2007).
20. Bogen, S. A. A Root Cause Analysis Into the High Error Rate in Clinical Immunohistochemistry. *Appl. Immunohistochem. Mol. Morphol.* **27**, 329–338 (2019).
21. Dabbs, D. J. et al. In Support of Magnani and Taylor. *Arch. Pathol. Lab Med.* **148**, 11 (2024).
22. 22C3, A. PD-L1 IHC 22C3 pharmDx. (https://www.agilent.com/en-us/product/pharmdx/pd-l1-ihc-22c3-pharmdx-overview, 2024).
23. 28-8, A. Dako PD-L1 IHC 28-8 pharmDx. (https://www.agilent.com/en-us/product/pharmdx/pd-l1-ihc-28-8-overview, 2024).
24. US-FDA, R.D.S. Ventana PD-L1 (SP263) Assay. (https://diagnostics.roche.com/global/en/products/lab/pd-l1-sp263-assay-ventana-rtd001235.html, 2024).
25. CE-IVD, R.D.S. Ventana PD-L1 (SP263) Assay. (https://diagnostics.roche.com/global/en/products/lab/pd-l1-sp263-ce-ivd-us-export-ventana-rtd001234.html, 2024).
26. US-FDA, R.D.S. Ventanaa PD-L1 (SP142) Assay. (https://diagnostics.roche.com/global/en/products/lab/pd-l1-sp142-assay-ventana-rtd001231.html, 2024).
27. CE-IVD, R.D.S.-. Ventana PD-L1 (SP142) Assay. (https://diagnostics.roche.com/global/en/products/lab/pd-l1-sp142-assay-us-export-ventana-rtd001232.html, 2024).
28. Hirsch, F. R. et al. PD-L1 Immunohistochemistry Assays for Lung Cancer: Results from Phase 1 of the Blueprint PD-L1 IHC Assay Comparison Project. *J. Thorac. Oncol.* **12**, 208–222 (2017).
29. Tsao, M. S. et al. PD-L1 Immunohistochemistry Comparability Study in Real-Life Clinical Samples: Results of Blueprint Phase 2 Project. *J. Thorac. Oncol.* **13**, 1302–1311 (2018).
30. Rimm, D. L. et al. A Prospective, Multi-institutional, Pathologist-Based Assessment of 4 Immunohistochemistry Assays for PD-L1 Expression in Non-Small Cell Lung Cancer. *JAMA Oncol.* **3**, 1051–1058 (2017).
31. Wang, X. et al. Concordance of assessments of four PD-L1 immunohistochemical assays in esophageal squamous cell carcinoma (ESCC). *J. Cancer Res. Clin. Oncol.* **150**, 43 (2024).
32. Tretiakova, M. et al. Concordance study of PD-L1 expression in primary and metastatic bladder carcinomas: comparison of four commonly used antibodies and RNA expression. *Mod. Pathol.* **31**, 623–632 (2018).
33. Prince, E. A., Sanzari, J. K., Pandya, D., Huron, D. & Edwards, R. Analytical Concordance of PD-L1 Assays Utilizing Antibodies From FDA-Approved Diagnostics in Advanced Cancers: A Systematic Literature Review. *JCO Precis Oncol.* **5**, 953–973 (2021).
34. Krigsfeld, G. S. et al. Analysis of real-world PD-L1 IHC 28-8 and 22C3 pharmDx assay utilisation, turnaround times and analytical concordance across multiple tumour types. *J. Clin. Pathol.* **73**, 656–664 (2020).
35. Keppens, C. et al. PD-L1 immunohistochemistry in non-small-cell lung cancer: unraveling differences in staining concordance and interpretation. *Virchows Arch.* **478**, 827–839 (2021).
36. Bogen, S. A. et al. A Consortium for Analytic Standardization in Immunohistochemistry. *Arch. Pathol. Lab Med.* **147**, 584–590 (2022).
37. Vani, K. et al. The Importance of Epitope Density in Selecting a Sensitive Positive IHC Control. *J. Histochem Cytochem* **65**, 463–477 (2017).
38. Sompuram, S. R., Vani, K., Schaedle, A. K., Balasubramanian, A. & Bogen, S. A. Quantitative Assessment of Immunohistochemistry Laboratory Performance by Measuring Analytic Response Curves and Limits of Detection. *Arch. Pathol. Lab Med.* **142**, 851–862 (2018).
39. Sompuram, S. R., Vani, K., Schaedle, A. K., Balasubramanian, A. & Bogen, S. A. Selecting an Optimal Positive IHC Control for Verifying Antigen Retrieval. *J. Histochem Cytochem* **67**, 275–289 (2019).
40. Materials, B.C.S.I.R. IHControls and IHCalibrators. (https://bostoncellstandards.com/products/, 2024).
41. McFadden, B. R., Reynolds, M. & Inglis, T. J. J. Developing machine learning systems worthy of trust for infection science: a requirement for future implementation into clinical practice. *Front. Digit Health* **5**, 1260602 (2023).
42. King, H., Wright, J., Treanor, D., Williams, B. & Randell, R. What Works Where and How for Uptake and Impact of Artificial Intelligence in Pathology: Review of Theories for a Realist Evaluation. *J. Med. Internet Res.* **25**, e38039 (2023).
43. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
44. Kang, M., Park, S. H., Yoo, S. & Pereira, D. S. Benchmarking Self-Supervised Learning on Diverse Pathology Datasets. *arxiv*, (2023).
45. Filiot A. et al. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. *medRxiv* https://doi.org/10.1101/2023.07.21.23292757 (2023).
46. Campanella G, K. R., et al. Computational Pathology at Health System Scale -- Self-Supervised Foundation Models from Three Billion Images. *arXiv* (2023).
47. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
48. Robert, M. E. et al. High Interobserver Variability Among Pathologists Using Combined Positive Score to Evaluate PD-L1 Expression in Gastric, Gastroesophageal Junction, and Esophageal Adenocarcinoma. *Mod. Pathol.* **36**, 100154 (2023).
49. Du, X. et al. Effectiveness and Cost-effectiveness of Artificial Intelligence-assisted Pathology for Prostate Cancer Diagnosis in Sweden: A Microsimulation Study. *Eur. Urol. Oncol.* **8**, 80–86 (2025).
50. B, F. Pathology Under Pressure: Unraveling the Exodus. in *The Pathologist* (2024).
51. Cohen, M. B. et al. Features of burnout amongst pathologists: A reassessment. *Acad. Pathol.* **9**, 100052 (2022).
52. Smith, S. M. et al. Burnout and Disengagement in Pathology: A Prepandemic Survey of Pathologists and Laboratory Professionals. *Arch. Pathol. Lab Med.* **147**, 808–816 (2023).
53. Food and Drug Administration, H. Medical Devices; Laboratory Developed Tests 21 CFR Part 809. (ed. Services, D.o.H.a.H.) (2024).
54. Diagnostics, F.a.D.A.A.C. List of Cleared or Approved Companion Diagnostic Devices (In Vitro and Imaging Tools). (2024).
55. Fernandez, A. I. et al. Multi-Institutional Study of Pathologist Reading of the Programmed Cell Death Ligand-1 Combined Positive Score Immunohistochemistry Assay for Gastric or Gastroesophageal Junction Cancer. *Mod. Pathol.* **36**, 100128 (2023).
56. Liu, D. H. W., Grabsch, H. I., Gloor, B., Langer, R. & Dislich, B. Programmed death-ligand 1 (PD-L1) expression in primary gastric adenocarcinoma and matched metastases. *J. Cancer Res Clin. Oncol.* **149**, 13345–13352 (2023).
57. Peixoto, R. D. et al. PD-L1 testing in advanced gastric cancer-what physicians who treat this disease must know-a literature review. *J. Gastrointest. Oncol.* **14**, 1560–1575 (2023).
58. Ruschoff, J. et al. HER2 diagnostics in gastric cancer-guideline validation and development of standardized immunohistochemical testing. *Virchows Arch.* **457**, 299–307 (2010).
59. Chen, P. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).

Acknowledgements

We thank members of Mindpeak and Augmentiqs who participated in discussions and especially Gabe Siegel and Dan Regelman for strategic direction and assistance in integrating the ARM hardware. We thank

Michelle H. Choi, John Marchlenski, Charles Reichley and the technical experts at Hunt Optics and Imaging for assembling the multi-head microscope. We thank members of the BMS IT department for providing laptop computers. We thank Kald Abdallah for supportive discussions and insights. Funding was provided by BMS. We thank and especially appreciate the patients who provided their tissue for scientific analysis.

Author contributions

G.L.K., S.B., F.F., T.L., and M.K., conceptualized the study and designed the experiments. For work performed on the Gastric Cell Atlas, S.B., K.D., and P.F. selected the ROIs, J.R., H.S., and S.B. interpreted the PD-L1 staining, and M.K., G.L.K., T.L., F.F., P.F., and K.D participated in the discussion and adjudication of the decision rules and J.R. devised the decision-tree. For work performed on day 1 and day 2 of the ARM-AI experiment, S.B. and G.L.K. selected and pre-marked the ROIs on 35 unique biopsies. T.L., P.F., F.F., M.P., and K.D. performed model development, and collected, cleaned and analyzed the data. J.P. performed the preliminary data analysis on day 1 and day 2 of the experiment. E.P. was responsible for integrating the ARM hardware and overseeing adjustments. R.A., D.C., R.S.G., R.P.G., A.M.K., X.L., A.Q., R.S., N.S., L.T., H.L.W., performed the PD-L1 CPS staining interpretation on day 1 and day 2 of the experiment. S.B., R.A., D.C., R.S.G., R.P.G., A.M.K., X.L., A.Q., R.S., N.S., L.T., H.L.W received hourly compensation for their participation on day 1 and day 2 of the experiment; pathologists' compensation was approved by M.K. and provided by BMS as part of research and development activities. T.L., G.L.K., S.B., and M.K. interpreted the experimental results; M.K. and T.L. prepared the paper with input from S.B. and G.L.K. and perspectives from all authors. M.K. supervised the research on behalf of BMS.

Competing interests

S.B., R.A., D.C., R.S.G., R.P.G., A.M.K., X.L., A.Q., R.M., N.S., L.T., H.L.W., received hourly payment for participation on day 1 and day 2. S.B., J.R., H.S., received hourly payment for work on Gastric Cell Atlas. J.R. is a consultant for DLS (co-founder of Targos now part of DLS), Astellas, AstraZeneca, BMS, Daiichi Sankyo, GSK, Merck Sharp&Dohme, Merck KGaA and QUip and co-founder of Gnothis Inc. R.A. is a consultant for BMS, Merck SD, AstraZeneca and Jazz pharmaceuticals. S.B. is a scientific advisor to Mindpeak, an ad hoc advisor to AstraZeneca, Daiichi Sankyo, Ventana-Roche, a speaker for AstraZeneca, Daiichi Sankyo, Agilent (Dako), Ventana-Roche, Merck, BMS, a research funding recipient of NCI R01CA121932, Eli Lilly and Agilent (Dako), is a Susan G. Komen Scholar, and the Director of ICGA Foundation, India. R.A. is the recipient of research funding from BMS, RAPT pharmaceuticals, StandUp2Cancer, Breakthrough Cancer and the NIH. G.K., M.K., were employees and shareholders of BMS, when the study was conducted. M.K. is a scientific advisor to DLS, ReviveMed, Picture-Health and a Board Member of FT3. J.P. is an employee and shareholder of BMS. T.L., P.F., M.P., K.D., F.F. are employees and shareholders of Mindpeak. E.P. is an employee and shareholder of Augmentiqs Medical.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-00899-5.

Correspondence and requests for materials should be addressed to Sunil Badve, George L. Kumar, Tobias Lang or Maria Karasarides.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.