



OPEN

## Comparing deep learning and pathologist quantification of cell-level PD-L1 expression in non-small cell lung cancer whole-slide images

Leander van Eekelen<sup>1✉</sup>, Joey Spronck<sup>1</sup>, Monika Looijen-Salamon<sup>1</sup>, Shoko Vos<sup>1</sup>, Enrico Munari<sup>3</sup>, Ilaria Girolami<sup>4</sup>, Albino Eccher<sup>5</sup>, Balazs Acs<sup>6</sup>, Ceren Boyaci<sup>6</sup>, Gabriel Silva de Souza<sup>1</sup>, Muradije Demirel-Andishmand<sup>1</sup>, Luca Dulce Meesters<sup>1</sup>, Daan Zegers<sup>1</sup>, Lieke van der Woude<sup>1</sup>, Willemijn Theelen<sup>7</sup>, Michel van den Heuvel<sup>8</sup>, Katrien Grünberg<sup>1</sup>, Bram van Ginneken<sup>2</sup>, Jeroen van der Laak<sup>1</sup> & Francesco Ciompi<sup>1</sup>

Programmed death-ligand 1 (PD-L1) expression is currently used in the clinic to assess eligibility for immune-checkpoint inhibitors via the tumor proportion score (TPS), but its efficacy is limited by high interobserver variability. Multiple papers have presented systems for the automatic quantification of TPS, but none report on the task of determining cell-level PD-L1 expression and often reserve their evaluation to a single PD-L1 monoclonal antibody or clinical center. In this paper, we report on a deep learning algorithm for detecting PD-L1 negative and positive tumor cells at a cellular level and evaluate it on a cell-level reference standard established by six readers on a multi-centric, multi PD-L1 assay dataset. This reference standard also provides for the first time a benchmark for computer vision algorithms. In addition, in line with other papers, we also evaluate our algorithm at slide-level by measuring the agreement between the algorithm and six pathologists on TPS quantification. We find a moderately low interobserver agreement at cell-level level (mean reader-reader F1 score = 0.68) which our algorithm sits slightly under (mean reader-AI F1 score = 0.55), especially for cases from the clinical center not included in the training set. Despite this, we find good AI-pathologist agreement on quantifying TPS compared to the interobserver agreement (mean reader-reader Cohen's kappa = 0.54, 95% CI 0.26–0.81, mean reader-AI kappa = 0.49, 95% CI 0.27–0.72). In conclusion, our deep learning algorithm demonstrates promise in detecting PD-L1 expression at a cellular level and exhibits favorable agreement with pathologists in quantifying the tumor proportion score (TPS). We publicly release our models for use via the Grand-Challenge platform.

Immunotherapy with immune checkpoint inhibitors (ICIs) has made a substantial impact in recent years on the treatment of a subset of non-small cell lung cancer (NSCLC) patients<sup>1</sup>, with clinical trials even reporting up to a doubling of median overall survival of patients compared to chemotherapy<sup>2</sup>. However, only a minority of NSCLC patients respond to ICI and only a few have a durable response<sup>3</sup>.

Selecting patients eligible for ICI therapy targeting the interaction between programmed death-ligand 1 (PD-L1) and PD-1 is typically performed based on PD-L1 immunohistochemistry (IHC). PD-L1 expression is

<sup>1</sup>Department of Pathology, Radboud University Medical Center, P.O.Box 9101, 6500 HB Nijmegen, The Netherlands. <sup>2</sup>Department of Radiology, Radboud University Medical Center, Nijmegen, The Netherlands. <sup>3</sup>Pathology Unit, Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy. <sup>4</sup>Department of Pathology, Provincial Hospital of Bolzano (SABES-ASDAA), Bolzano-Bozen, Italy. <sup>5</sup>Department of Pathology and Diagnostics, University and Hospital Trust of Verona, Verona, Italy. <sup>6</sup>Department of Clinical Pathology and Cancer Diagnostics, Karolinska University Hospital, Stockholm, Sweden. <sup>7</sup>Department of Thoracic Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands. <sup>8</sup>Respiratory Diseases Department, Radboud University Medical Center, Nijmegen, The Netherlands. ✉email: leander.vaneekelen@radboudumc.nl

estimated visually by a pathologist and expressed as the tumor proportion score (TPS): the percentage of PD-L1 positive (viable) tumor cells as compared to the total number of tumor cells. TPS is currently the only approved biomarker to select NSCLC patients for ICI treatment<sup>3</sup>.

Despite its daily use in the clinic, IHC-based TPS suffers from several fundamental limitations, caused by variations in PD-L1 monoclonal antibodies and staining intensity, heterogeneous PD-L1 expression throughout tumors<sup>4,5</sup>, and the intrinsic limitation of the human visual capability to detect, classify and quantify thousands of tumor cells in histopathology slides<sup>6</sup>. As a result, studies have found substantial interobserver variability at TPS assessment<sup>7,8</sup>. Lastly, TPS has been shown to have an intrinsically limited predictive power as a biomarker: Lu et al.<sup>9</sup> found an AUC of 0.65 in a meta-analysis of studies assessing the accuracy of PD-L1 IHC predicting response to PD-L1 targeting ICIs.

TPS quantification in NSCLC can be viewed as a combination of three different tasks: (1) localization of cells within the image, (2) classification of individual cells into PD-L1 negative tumor cells, PD-L1 positive tumor cells or 'other' (i.e. 'to be excluded') and (3) finally calculating a ratio of the positive versus the total number of tumor cells. This combination of tasks makes TPS an excellent candidate for automation by deep learning (DL), as previous works have shown good performance of DL-based computer models for cell localization and classification<sup>10,11</sup>. In addition, addressing TPS quantification by automating localization and classification at cellular level can enable development of novel biomarkers based on spatial analysis of the tumor tissue.

There are a number of prerequisites for developing deep learning systems that detect PD-L1 expression at a cellular level. Firstly, an efficient and scalable way of generating manual annotations is required for training supervised deep learning models. Secondly, cases for the development and validation of the system should be reflective of data in clinical practice. The data should ideally cover a broad range of tissue morphologies, specimen types (biopsy and resections) and scanners. Particular focus should be put on including a wide selection of PD-L1 monoclonal antibodies used in the clinic, as the system should generalize across multiple antibodies and staining expressions. Most technical studies have only reported results within well-controlled, single-center/single-antibody environments<sup>12–14</sup>, which hampers their generalizability and applicability. Lastly, since cell-level DL methods are trained with supervision, a human benchmark based on the input of multiple readers is needed, to establish the theoretical upper bound performance of computer methods. Previous works have only correlated algorithms to TPS as visually estimated by pathologists<sup>12–17</sup>, but to the best of our knowledge, no work exists that quantifies the interobserver variability of pathologists at classifying cell-level PD-L1 expression and compares deep learning methods to cell-level annotations in a multi-reader setting.

In this study, we developed a DL-based PD-L1 detector for NSCLC whole-slide images, trained on multi-centric, multi-PD-L1 assay data. First, we introduce a fast and scalable procedure to generate annotations by developing a nuclei detector that enables semi-automatic multi-class annotations. Second, we used the annotations to develop a multi-class cell detector in PD-L1 IHC. Third, we set up a reader study where a panel of pathologists established a cell-level reference standard. We used this to measure, for the first time in the literature, the interobserver variability of pathologists at determining cell-level PD-L1 expression, providing a benchmark for computer vision algorithms. Finally, in line with previous studies, we correlate the automated TPS derived from the output of our PD-L1 detector to the slide-level visual estimates of pathologists.

## Materials and methods

### Datasets

To capture the wide diversity of PD-L1 IHC in the clinic, we created an international, multi-centric, multi-scanner and multi PD-L1 assay collection of slides that included biopsies, resections and tissue microarrays (TMAs). In total 152 NSCLC patients were included. All patients had a single slide which originated either from the primary tumor or a metastasis to another site in the lungs. Patients originated from one of three clinical centers: (1) 73 patients from the Radboud University Medical Center, Nijmegen, the Netherlands, referred to as *RUMC*, (2) 54 patients from the Sacro Cuore Don Calabria Hospital of Negrar, Verona, Italy, referred to as *NEG* and (3) 25 patients from the Netherlands Cancer Institute–Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands, referred to as *NKI*. Material of patients in the *RUMC* and *NEG* cohorts (previously used and described in<sup>18</sup> and<sup>19</sup>) was gathered as part of routine clinical care. *NKI* patient material belongs to a subset of patients participating in the phase 2 PEMBRO-RT randomized control trial<sup>20</sup> and was collected pre-treatment. Slides were stained for PD-L1 and scanned by the laboratories of their respective clinical centers. Table 1 shows the scanners, collection period and monoclonal antibodies per dataset, in addition to the distribution of specimen types (see Supplementary materials & methods for additional details).

For the *RUMC* dataset, the need for informed consent was waived by the institutional review board of the Radboud UMC, Nijmegen, the Netherlands (reference 2018-4764). For the *NEG* dataset, the need for informed consent was waived by the Ethic Committee for Clinical Research of the Provinces of Verona and Rovigo (reference 25046). For the *NKI* dataset, the PEMBRO-RT clinical trial (reference 2014-005118-49) from which cases were taken was approved by the institutional review board of the Netherlands Cancer Institute–Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. The research was performed in accordance with the principles expressed in the Declaration of Helsinki.

The annotations used to train and validate the algorithms were made by trained research assistants (L.M, M.D.A) and one physician (G.S) under supervision of two pathologists (E.M and M.L.S). The pathologists selected regions of interest (ROIs), 3 per slide on average (sized 300 by 300  $\mu\text{m}$ ), deemed to be a representative sample of morphologies and cell types typically found in NSCLC histopathology, as well as with heterogeneous distribution of PD-L1 positivity. For the nuclei detector, all nuclei within the ROI were manually point annotated, while for the PD-L1 detector hand-drawn polygons were made for the semi-automatic annotation method (see next section). In total, 67,033 nuclei and 526,348 cells were annotated for the nuclei detector and the PD-L1

Dataset	Collection period	Scanner	PD-L1 monoclonal antibody	Specimen type	Total #unique patients	Nuclei detector			PD-L1 detector			TPS reader study	
						Training	Validation	Test	Training	Validation	Test		
RUMC	2017-2019	PANNORAMIC 1000 (3D Histech)	E1L3N	Biopsy	25							25	
		PANNORAMIC 250		Resection	48	19	12	12	19	4	12	12	
NEG	2003-2017	Ventana iScan HT scanner (Roche)	SP263 Ventana	Resection	38							9	38
				TMA	6				3	3			
				22C3 Dako	TMA	10	2	2		7	3		
NKI	2015-2018	PANNORAMIC 1000	22C3 (Custom NKI LDT)	Biopsy	25						9	25	
				<b>SUMS</b>	152	21	14	12	29	10	30	100	

**Table 1.** A general overview of the datasets and the number of cases used for each experiment.

detector respectively (see Supplementary Tables 1 and 2 for additional detail). Annotations were made using ASAP, an open-source in-house developed whole-slide image viewer.

### Deep learning models

The multi-class PD-L1 detector proposed in this paper was trained using fine-grained point annotations on a cell-per-cell basis. These are typically time intensive to produce, as an annotator must simultaneously determine both the cell's type (classification) and its approximate center (localization). Therefore, we decoupled the localization aspect from the annotation task to allow annotators to focus on the more knowledge-intensive task of classifying cell type. We propose an efficient two-step annotation pipeline: first, experts make free-hand annotations in pre-selected regions of interest that label regions of homogeneous cell types; second, we developed a nuclei detector for the detection of nuclei centers in IHC; detected nuclei are intersected with the free-hand drawn polygons and then given the corresponding label. We show a schematic representation of using the nuclei detector's output to generate a reference standard for the PD-L1 detector in Fig. 1A. The PD-L1 detector was developed to detect three cell types within NSCLC WSIs: PD-L1 positive tumor cells, PD-L1 negative tumor cells and an 'other cells' type. The 'other cells' type was used to make our algorithm applicable to WSIs, as these may contain many other structures than malignant cells.

We adopted the YOLOv5 convolutional neural network model architecture for both the nuclei and PD-L1 detectors, since it has strong computer vision performance with minimal required configuration. As YOLOv5 was originally designed for object detection with bounding boxes, we extend point annotations during training to bounding boxes of 20 by 20  $\mu\text{m}$ , as determined by measuring the average size of cells within the dataset and adding a safety margin to this for cells with skewed aspect ratios. During inference we consider the center of the predicted bounding box as the detection location. We preprocess slides with a tissue segmentation network<sup>21</sup> to limit analysis to tissue during whole-slide inference. Further details are given in the Supplementary materials and methods.

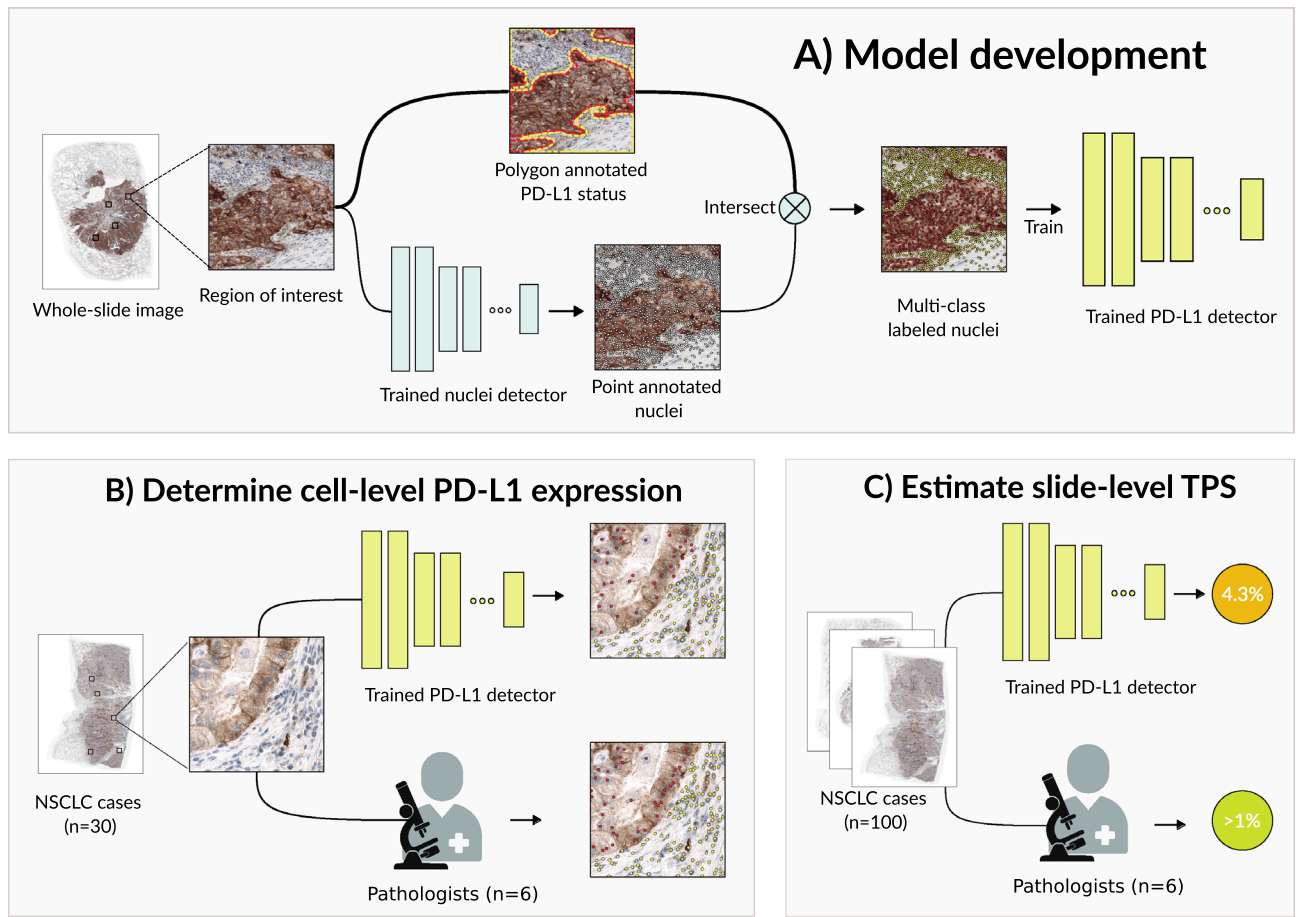
The nuclei detector and PD-L1 detector were developed using 35 and 39 slides respectively (taken from the RUMC and NEG datasets). The train/validation split for both algorithms is shown in Table 1 and was stratified in terms of NSCLC subtype and cell types, keeping the proportions identical between train and validation sets (Supplementary Table 3 shows the distribution of NSCLC subtypes per set). The validation set was used to monitor for overfitting during training (based on the classification, object and box loss of YOLOv5, see<sup>22</sup> for details). The model weights that performed best on the validation set were used to compare the models to human readers using holdout test sets (see next section).

### Benchmarking of the algorithms against readers

Interobserver variability was measured amongst research assistants for nuclei detection and amongst expert pathologists for slide-level (i.e. TPS) and cell-level (i.e. 'positive' or 'negative' cells) determination of PD-L1 expression. These reader studies served as holdout test sets for the nuclei and PD-L1 detectors respectively (all slides were previously unseen by the algorithms). This allowed us to determine whether the algorithms performed within interobserver variability (the theoretical upper bound performance of an algorithm trained on answers from a single reader).

The nuclei detector was tested on ROIs within 12 resections from the RUMC dataset. Each ROI was exhaustively annotated for nuclei by four readers: three trained student assistants (L.M, M.D.A and D.Z) and one trained physician (G.S), labelled  $R_i$  through  $R_{iv}$ .

For the PD-L1 detector, the interobserver variability for determining cell-level PD-L1 expression (Fig. 1B) was measured by inviting five expert pathologists and one trained physician to make manual annotations on ROIs within 30 cases (12 resections from the RUMC dataset, 9 resections from the NEG dataset and 9 biopsies from the NKI dataset). Each reader exhaustively point annotated ROIs for negative tumor cells, positive tumor and 'other cells'. ROIs were approximately 150 by 150  $\mu\text{m}$  (4–5 per case) and were selected by a pathologists (E.M.) The case series was read in two groups: RUMC cases were each annotated by 3 pathologists (A.E, E.M. and I.G,



**Figure 1.** Overview of study design. In (A), the model development process is shown. First, a nuclei detector is trained using point annotated nuclei. Secondly, polygons are drawn in regions of homogeneous cell types and intersected with detected nuclei. The nuclei are given the label of the corresponding region, resulting in multi-class labeled nuclei usable for training a multi-class PD-L1 detector. The PD-L1 detector is validated by (B) correlating slide-level tumor proportion score as determined by the system with the visual estimate of pathologists and (C) comparing its detection output to manual cell-level annotations of pathologists.

labeled  $P_1$  through  $P_3$ ) and the NEG and NKI cases were each annotated by 2 pathologists and the physician (B.A, C.B and G.S, labeled  $P_4$  through  $P_6$ ). The interobserver variability for determining slide-level PD-L1 expression (Fig. 1C) was measured by inviting six expert pathologists (E.M, I.G, A.E, K.G, S.V. and M.L.S., labeled  $P_a$  through  $P_f$ ) to score TPS in 100 cases (12 resections and 25 biopsies from the RUMC dataset, 38 resections from the NEG dataset and 25 biopsies from the NKI dataset). TPS was estimated as the percentage of viable tumor cells showing partial/complete circumferential membranous staining above background level in five percent increments. Of all readers, only  $P_d$  has not previously taken courses for the reading of PD-L1 stained IHC. In both experiments, cases were selected to represent a wide diversity of PD-L1 IHC (Supplementary Table 3), encompassing three PD-L1 monoclonals (22C3 NKI, E1L3N and SP263). We refer to the RUMC and NEG cases as in-domain test cases, while the NKI cases served as out-of-domain test cases, meant to test the generalizability of the PD-L1 detector to a PD-L1 monoclonal antibody not included in its training set.

Cases were read via the online grand-challenge.org platform using the Reader Studies functionality (Supplementary Fig. 1). Whilst reading cases, readers could view the H&E and PD-L1 slide in parallel; whenever possible, this H&E slide was adjacent to the PD-L1 slide in terms of cutting order.

### Statistical analysis

To evaluate cell-level observer variability, we use a “hit criterion” based on the distance between predictions and a reference standard: a prediction is valid within an 8  $\mu\text{m}$  radius of a reference point. This chosen radius was identical for all cell types and corresponds to the average diameter of tumor and immune cell nuclei in the dataset. If multiple predictions fall within this radius, only the closest one to the reference standard is considered a hit. Predictions labeled as the correct cell type are true positives, while others are false positives. Reference standards without associated predictions are false negatives, and predictions that do not hit any ground truth points are false positives. Using this criterion, the F1 score was computed, defined as the harmonic mean of precision and recall:  $2 \cdot TP / (2 \cdot TP + FP + FN)$ . We report the mean and class specific F1 score for all pairings of reader and PD-L1 detector.



We measured the agreement on scoring TPS in pairwise combinations of pathologists and the PD-L1 detector in two ways: in a continuous fashion via the intra-class correlation (ICC) using a two-way random effect model with absolute agreement [ICC(2,1)] and in a discretized fashion taking into account the clinically relevant cutoffs (< 1%, 1–49% and  $\geq$  50% TPS) via linear Cohen's kappa.

95% confidence intervals were calculated by taking 1.96 standard deviations away from the mean for normally distributed metrics, as determined by a Shapiro-Wilks test. A significance level of 0.05 was chosen for all statistical tests.

## Results

### Agreement between readers and nuclei detector on localizing nuclei

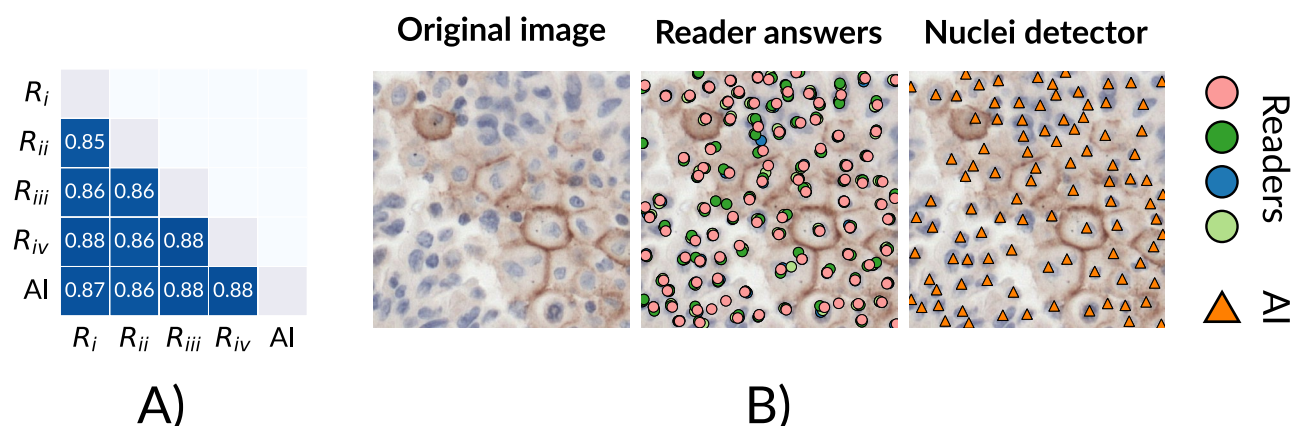
The F1 scores per reader-reader/reader-algorithm pairing are shown in Fig. 2A and we show examples of predictions of the nuclei detector together with predictions of the three readers in Fig. 2B. Overall, the average reader-reader and reader-algorithm F1 scores were identical at 0.87. The lowest F1 score was measured between reader i and reader ii (0.85), while the highest F1 score was tied between i-iv, iii-iv, iii-AI and iv-AI (0.88). Visual inspection of the algorithmic predictions indicated that the nuclei detector performs equally well on PD-L1 negative and positive cells, as well as other cell types.

### Agreement between pathologists and PD-L1 detector on determining cell-level PD-L1 expression status

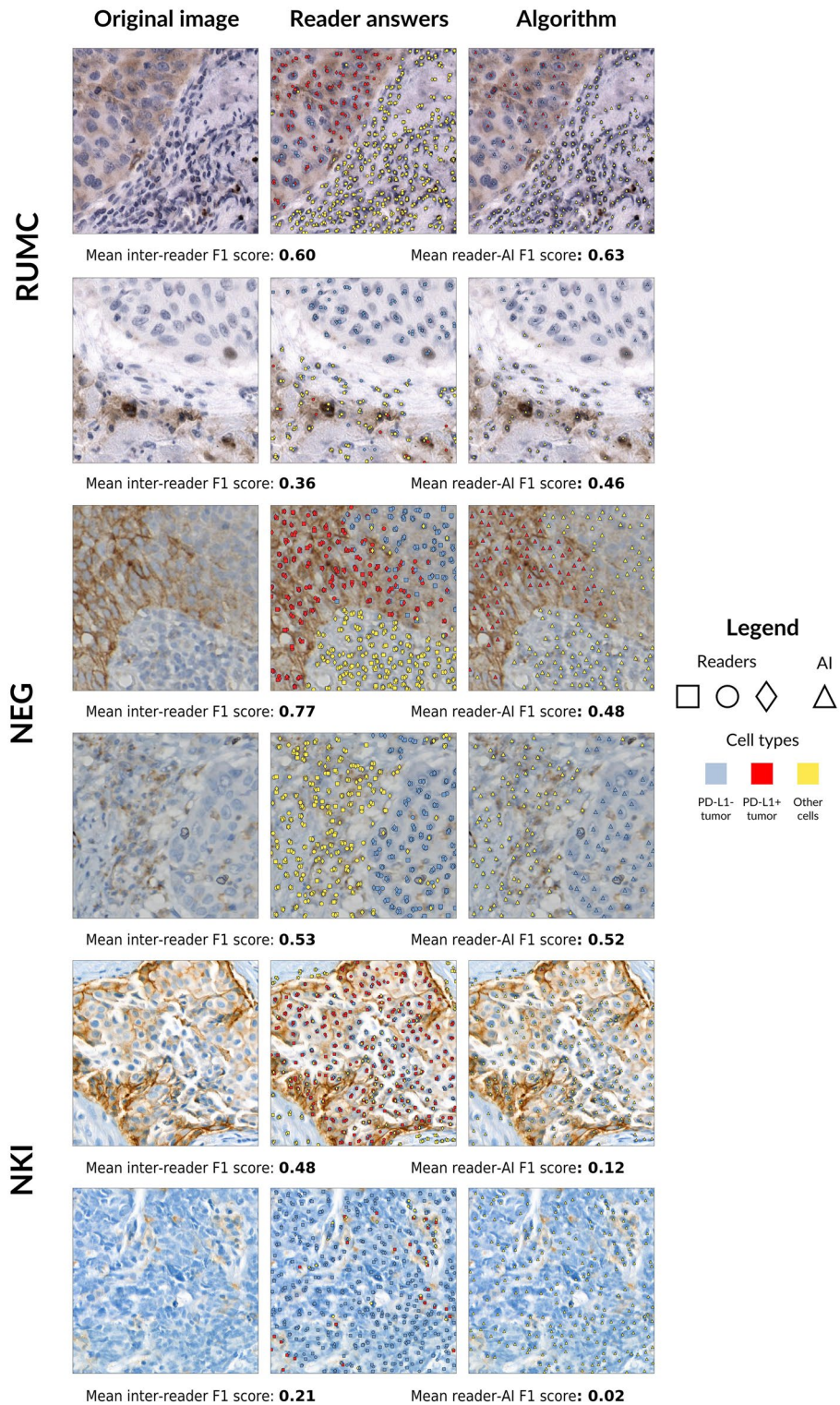
We show examples of ROIs of all clinical centers alongside predictions of the six readers and the PD-L1 detector in Fig. 3. The F1 scores of all pairings are shown in Fig. 4, split per clinical center (we show F1 scores split per cell type in Supplementary Fig. 2). When comparing the mean reader-reader and reader-algorithm F1 scores, we find an overall F1 score of 0.68 versus 0.55. Considering each clinical center separately, we see similar performance on RUMC cases (0.64 versus 0.62), a mild performance drop for the PD-L1 detector on NEG cases (0.75 versus 0.59) and a substantial performance drop on the out-of-domain NKI cases (0.65 versus 0.42). The performance drop of the algorithm on NKI cases is most noticeable on the PD-L1 positive tumor cells: the mean F1 score for this cell type is 0.11. We show NKI cases in the last two rows of Fig. 3, where we can see that the PD-L1 detector falsely considers most cells as 'other cells'.

### Agreement between pathologists and PD-L1 detector on determining slide-level TPS

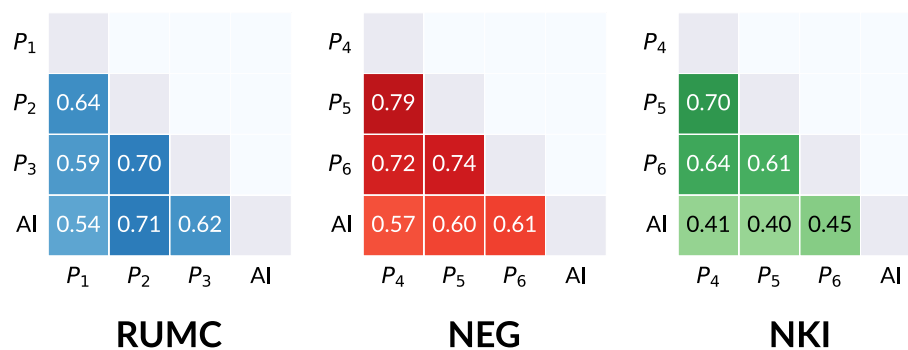
When measured continuously, the mean reader-reader and reader-algorithm agreements on determining slide-level TPS were both good (ICC = 0.852 [95% CI 0.768–0.936] versus ICC = 0.796 [95% CI 0.694–0.898]). We show the ICC for all pairings in Supplementary Fig. 3. The mean difference between the median reader TPS and the PD-L1 detector TPS was  $-4.70$  (95% CI  $-43.50$  to  $34.10$ ), indicating that the PD-L1 detector systematically measured a lower TPS than the median score of the pathologists. We show the agreement between the median reader TPS and the PD-L1 detector for all cases in a Bland Altman plot in Supplementary Fig. 4 and show the mean difference per clinical center in Supplementary Table 4. When measured in a discretized fashion, the mean reader-reader kappa was 0.535 (95% CI 0.258–0.812) and the mean reader-AI kappa was 0.494 (95% CI 0.272–0.715) for all cutoffs together (< 1%, 1–49% and  $\geq$  50%). The mean reader-AI kappas per clinical center for all cutoffs together are reported in Supplementary table 4. We show the distribution of kappa values per reader in Fig. 5 for the 1% cutoff and 50% cutoff alone and also plot the agreement versus the majority vote of the six readers (the kappa score of each reader compared to the majority vote was calculated excluding their own ratings). There, we see that for the difficult 1% cutoff, the PD-L1 detector has an agreement with the majority vote of 0.72, better than four of the six readers, while the maximum agreement with the majority vote was reached



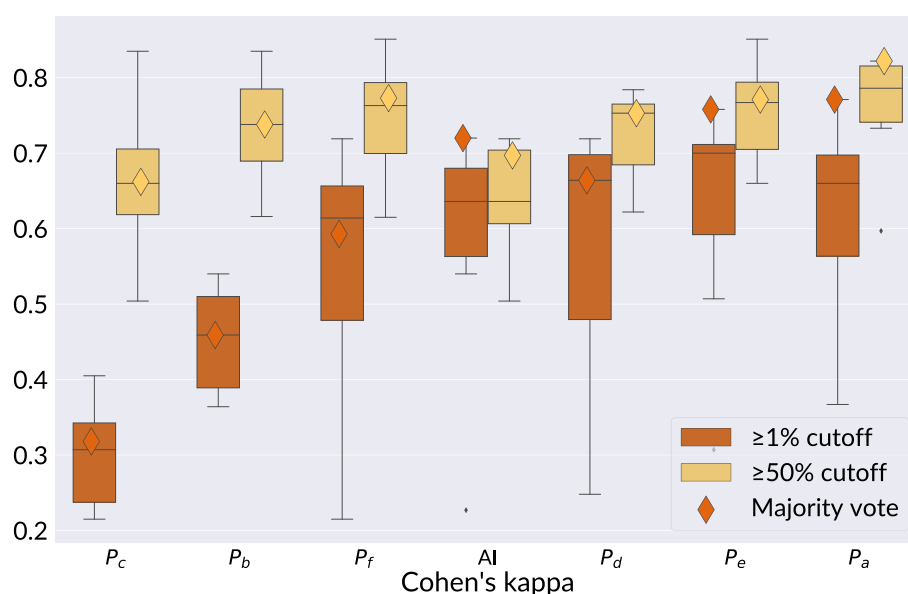
**Figure 2.** Validation of the nuclei detector. In (A), pairwise F1 scores are shown over cell predictions in the RUMC nuclei detection test set made by the nuclei detector and human readers. In (B), one ROI from the test set is shown with annotations by the readers and predictions from the nuclei detector overlaid on the PD-L1 immunohistochemistry. Annotations from the readers are shown as pink, light yellow, green and blue dots, while the AI predictions are shown as orange triangles.



**Figure 3.** An overview of PD-L1 detector results as compared to the cell-level predictions of the readers. Each reader in the middle column has their own unique symbol (diamond, square or circle), while the PD-L1 detector results are always shown as triangles in the last column. The color of the symbol indicates the predicted class: blue for PD-L1 negative tumor cells, red for PD-L1 positive tumor cells, yellow for other cells. The mean F1 scores for all reader-reader (‘inter-reader’) and reader-AI pairs are shown per ROI underneath each row.



**Figure 4.** Pairwise F1 scores over the cases from the cell-level reader study, measuring the cell-level interobserver agreement of pathologists and AI on determining PD-L1 expression. The F1 scores are shown per clinical (*RUMC*, *NEG* and *NKI*) and are averaged over the three classes (PD-L1 negative/positive tumor cells and ‘other cells’). The F1 scores per class are reported in Supplementary Fig. 2.



**Figure 5.** The distribution of the linear Cohen's kappa agreement of readers with other readers for predicting the tumor proportion score (TPS) on  $n=100$  test set cases. For this analysis, agreement on TPS was considered solely at the 1% and at the 50% cutoff. The diamonds in each boxplot visualize the agreement of the reader with the majority vote of human readers for that particular cutoff. The boxplots of each reader are sorted on the x-axis according to the median agreement with the majority vote over both cutoffs.

by  $P_a$  (0.77). For the 50% cutoff, the PD-L1 detector reaches an agreement with the majority vote of 0.697, only better than one pathologist, but it should be noted that all majority vote agreements are consistently good (0.662 to 0.822). Lastly, the PD-L1 has the lowest discrepancy between kappas on the 1% and 50% cutoffs.

## Discussion

In this work, we reported on a deep learning system designed for the automatic detection of PD-L1 positive and negative tumor cells in digital NSCLC histopathology images. These detections can be used subsequently for the automatic quantification of the tumor proportion score (TPS). The method consists of a single-stage PD-L1 cell detector, which was trained using a novel combination of a nuclei detector and polygon annotations to accelerate collection of the reference standard. We evaluate our system at (1) at cell-level by measuring the agreement with pathologists on estimating PD-L1 expression in individual cells and (2) at slide-level by measuring the agreement with pathologists on the clinically relevant task of scoring TPS.

The nuclei detector was validated against four human readers on an independent test set, where each reader annotated roughly 10,000 nuclei. There, the nuclei detector performed on par with the average reader (with an F1 score of 0.87), indicating the system performs within interobserver variability on tested data. The largest source of disagreement between readers and the nuclei detector arose in giant, multinucleated cells, where the nuclei



detector predicted each distinct lobe as a nucleus. In addition, cells with a faint hematoxylin staining (possibly deeper within the z-stack of the tissue) were inconsistently annotated by both the nuclei detector and the readers.

In our cell-level reader study, the PD-L1 detector performs on par or slightly below the interobserver variability on the in-domain RUMC and NEG datasets, but underperforms compared to interobserver variability on the out-of-domain dataset of the NKI. Here, the F1 score on PD-L1 positive cells drops notably due to overprediction of 'other cells'. We attribute this poor agreement to the fact NKI cases appear substantially different from the in-domain test cases, partially due to the custom 22C3 laboratory-developed test, which takes on an atypical red-brown color. Moreover, the tissue preparation of the NKI cases results in different tissue morphology, leading to further disparity between the in-domain and out-of-domain test sets. Lastly, there were no biopsies included in the training set of the PD-L1 detector, which may have withheld biopsy-unique morphological information from the PD-L1 detector.

Dealing with the domain-gap between the training set and test set of a model is an active research field in machine learning and there are multiple possible avenues to overcome the limited performance on the out-of-domain NKI dataset. First and foremost, the training set could be expanded with more multi-centric data, including biopsies (to account for possible morphological artifacts) and PD-L1 staining using the 22C3 antibody, optionally only including data scanned with the same scanner used at the NKI, to make the test set more comparable with the training set. Secondly, heavier data augmentation with augmentation techniques specifically tailored for IHC could be used. Moreover, style-transfer methods<sup>23,24</sup> have been shown to close the domain gap. Lastly, stain normalization, although heavily used for applications using standard H&E WSI, remains underused in immunohistochemistry applications and could be explored.

We show the confusion matrices that underlie the pairwise F1 scores of Fig. 4 in Supplementary Fig. 5. The most prevalent source of disagreement between readers was between PD-L1- tumor cells and 'other cells', most likely due to uncertainty on whether the cell is benign or malignant. The PD-L1 detector shows a similar pattern, but also frequently disagreed with readers on positive tumor cells, instead considering these as 'other cells', most likely due those cells being considered alveolar macrophages. It should be noted that the readers had access to the H&E stain while reading PD-L1 cases, while the PD-L1 detector did not, possibly making the morphological discrimination of macrophages easier. In the case that a reader or the PD-L1 detector predicted 'other cells' (such as stromal or inflammatory immune cells), there was almost always agreement with the other reader. Of note is the amount of times readers disagree amongst each other on the presence/absence of a cell (labeled 'background' in Supplementary Fig. 5), accounting for 22% of the disagreements on average. The PD-L1 detector has a significantly lower ( $p < 0.01$ ) mean 'background' disagreement of 15% when comparing its predictions to annotations by readers, reflecting its tendency to localize cells more accurately. We associate this to the fact that when producing point annotation, the simultaneous focus on localization as well as classification of cells may overload readers, which highlights the efficacy of our semi-automatic annotation style, eliminating the localization aspect of annotating.

In our slide-level reader study, we find that the mean kappa of paired readers sits very close to the mean kappa of readers versus AI (0.535 versus 0.494 respectively, with overlapping confidence intervals), indicating the PD-L1 detector is able to quantify tumor proportion score at a level within interobserver variability. When considering the cutoffs separately, we find that the PD-L1 detector is more consistent around the cut offs than the readers, implying the algorithm can be valuable for cases that are difficult to interpret. This is further substantiated by the fact that the PD-L1 detector ranks third in terms of its kappa agreement with the majority vote on the 1% TPS cutoff.

We find lower kappa values compared to previous studies that measure interobserver variability of pathologists and the agreement of computer models and pathologists for (visually) estimating TPS (0.535 for interobserver and 0.494 for AI-pathologist agreement). Wu et al.<sup>17</sup> measured the agreement between the majority vote of six pathologists and a deep learning system on 100 out-of-domain SP263 stained slides and found a Cohen's kappa of 0.65 for TPS below 50% and 0.926 for TPS above 50%. Hondelink et al.<sup>12</sup> correlated three pathologists to a proprietary algorithm on 139 in-domain 22C3 stained slides and found an overall Cohen's kappa of 0.68. Choi et al.<sup>15</sup> measured on 479 out-of-domain 22C3 stained cases, finding a mean Cohen's kappa between 3 pathologists and their PD-L1 analyser of 0.67. All three studies also found comparable interobserver variability of visually estimating TPS by pathologists ( $K = 0.7-0.89$ ,  $K = 0.61$  and  $K = 0.739-0.871$  respectively). We hypothesize that the unique multi-center, multi-stain setup of our reader study lowered both the agreement amongst pathologists and with the AI.

To our best knowledge, we provide for the first time a cellular benchmark for PD-L1 computer algorithms by measuring the interobserver variability of pathologists on determining cell-level PD-L1 expression. This precludes direct comparison to literature values. Only Choi et al.<sup>15</sup> reports cellular F1 scores, but they do so only on PD-L1 negative and positive tumor cells within neoplastic regions (0.723/0.722, comparable to ours).

Our results are noteworthy because they indicate substantial interobserver variability, showing that the reference standard for PD-L1 related algorithms must be established very carefully, preferably using multiple readers fused via majority vote to promote robustness. Careful consideration must go into instructing readers on annotating and selecting readers that have sufficient experience in reading PD-L1 stains. The effect size of these variables could be the topic of further study, similar to Butter et al.<sup>7</sup>

This study has some limitations. Firstly, we did not quantitatively measure the time gain of our semi-automatic annotation technique over other techniques. While the added value of the nuclei detector is most pronounced in areas of homogeneous cell types, time savings tend to be diminished in areas of heterogeneous cell types. Secondly, for the cell-level F1 score metric, we cannot distinguish between instances where readers overlooked a cell versus cases where they believed that a particular shape did not represent a cell, leading to potentially lower mean F1 scores, especially for the readers as their localization accuracy was noticeably worse than that of the PD-L1 detector. Thirdly, we did not measure the intraobserver variability of readers for the cell-level reader



study. Fourthly, there was a substantial class imbalance for the PD-L1 detector training set, as PD-L1 positive tumor cells only accounted for 8% of the annotations. This may have contributed to the reader-AI agreement being consistently lower for positive tumor cells. Finally, pathologists had access to morphological information from H&E slides in the TPS reader studies, which was not available to our AI algorithm. Although not required for AI to address the task of multi-class cell detection, this factor might play a role in some of the differences observed in TPS scores between experts and AI.

In conclusion, we developed a deep learning system for the automatic detection of PD-L1 positive and negative tumor cells. While doing so, we performed reader studies and provided the first interobserver cell-level analysis on PD-L1, which we used to measure and report the agreement amongst readers and AI-reader agreement on a cellular basis and at TPS level. Future work should focus on improving the model to generalize better to unseen antibodies and analyze the predictive value of the automatically quantified TPS for ICI treatment response as compared to pathologists (similar to Kapil et al.<sup>25</sup>). Lastly, the model's cell-level predictions could be used to develop (spatial) biomarkers that go beyond the predictive value of TPS.

## Data availability

The RUMC dataset is available from the corresponding author on reasonable request. The NEG and NKI datasets are not available as these were shared with the authors by third parties under licenses that preclude further dissemination. The authors publicly release the nuclei detector and the PD-L1 for model inference on Grand Challenge: <https://grand-challenge.org/algorithms/tumor-proportion-score-in-non-small-cell-lung-canc/>. <https://grand-challenge.org/algorithms/nuclei-detection-in-immunohistochemistry/>.

Received: 17 November 2023; Accepted: 14 March 2024

Published online: 26 March 2024

## References

1. Onoi, K. et al. Immune checkpoint inhibitors for lung cancer treatment: A review. *J. Clin. Med.* **9**, 1362 (2020).
2. Nasser, N. J., Gorenberg, M. & Agbarya, A. First line immunotherapy for non-small cell lung cancer. *Pharmaceuticals* **13**, 373 (2020).
3. Prelaj, A. et al. Predictive biomarkers of response for immune checkpoint inhibitors in non-small-cell lung cancer. *Eur. J. Cancer* **106**, 144–159 (2019).
4. Paolino, G. et al. PD-L1 evaluation in head and neck squamous cell carcinoma: Insights regarding specimens, heterogeneity and therapy. *Pathol. Res. Pract.* **226**, 153605 (2021).
5. Uruga, H. & Mino-Kenudson, M. Predictive biomarkers for response to immune checkpoint inhibitors in lung cancer: PD-L1 and beyond. *Virchows Arch.* **478**, 31–44 (2021).
6. Girolami, I. et al. Challenges facing pathologists evaluating PD-L1 in head & neck squamous cell carcinoma. *J. Oral Pathol.* **50**, 864–873 (2021).
7. Butter, R. et al. The impact of a pathologist's personality on the interobserver variability and diagnostic accuracy of predictive PD-L1 immunohistochemistry in lung cancer. *Lung Cancer* **166**, 143–149 (2022).
8. Cooper, W. A. et al. Intra- and interobserver reproducibility assessment of PD-L1 biomarker in non-small cell lung cancer. *Clin. Cancer Res.* **23**, 4569–4577 (2017).
9. Lu, S. et al. Comparison of biomarker modalities for predicting response to PD-1/PD-L1 checkpoint blockade: A systematic review and meta-analysis. *JAMA Oncol.* **5**, 1195–1204 (2019).
10. Graham, S. et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
11. Verma, R. et al. MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Trans. Med. Imaging* **40**, 3413–3423 (2021).
12. Hondelink, L. M. et al. Development and validation of a supervised deep learning algorithm for automated whole-slide programmed death-ligand 1 tumour proportion score assessment in non-small cell lung cancer. *Histopathology* **80**, 635–647 (2022).
13. Liu, J. et al. Automated tumor proportion score analysis for PD-L1 (22C3) expression in lung squamous cell carcinoma. *Sci. Rep.* **11**, 1–9 (2021).
14. Taylor, C. R. et al. A multi-institutional study to evaluate automated whole slide scoring of immunohistochemistry for assessment of programmed death-ligand 1 (PD-L1) expression in non-small cell lung cancer. *Appl. Immunohistochem. Mol. Morphol.* **27**, 263 (2019).
15. Choi, S. et al. Artificial intelligence-powered programmed death ligand 1 analyser reduces interobserver variation in tumour proportion score for non-small cell lung cancer with better prediction of immunotherapy response. *Eur. J. Cancer* **170**, 17–26 (2022).
16. Widmaier, M. et al. Comparison of continuous measures across diagnostic PD-L1 assays in non-small cell lung cancer using automated image analysis. *Mod. Pathol.* **33**, 380–390 (2020).
17. Wu, J. et al. Artificial intelligence-assisted system for precision diagnosis of PD-L1 expression in non-small cell lung cancer. *Mod. Pathol.* **35**, 403–411 (2021).
18. Munari, E. et al. PD-L1 assays 22C3 and SP263 are not interchangeable in non-small cell lung cancer when considering clinically relevant cutoffs. *Am. J. Surg. Pathol.* **42**, 1384–1389 (2018).
19. Munari, E. et al. Impact of PD-L1 and PD-1 expression on the prognostic significance of CD8+ tumor-infiltrating lymphocytes in non-small cell lung cancer. *Front. Immunol.* **1**, 1971 (2021).
20. Theelen, W. S. M. E. et al. Effect of pembrolizumab after stereotactic body radiotherapy vs pembrolizumab alone on tumor response in patients with advanced non-small cell lung cancer: Results of the PEMBRO-RT phase 2 randomized clinical trial. *JAMA Oncol.* **5**, 1276–1282 (2019).
21. Bándi, P., Balkenhol, M., Van Ginneken, B., Van Der Laak, J. & Litjens, G. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. *PeerJ* **2019**, 7 (2019).
22. Jocher, G. et al. ultralytics/yolov5: v7.0—YOLOv5 SOTA realtime instance segmentation. <https://doi.org/10.5281/zenodo.7347926> (2022).
23. Tellez, D. et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (2019).
24. De-Bel, T. et al. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Med. Image Anal.* **70**, 102004 (2021).

25. Kapil, A. *et al.* Domain adaptation-based deep learning for automated tumor cell (TC) scoring and survival analysis on PD-L1 stained tissue images. *IEEE Trans. Med. Imaging* **40**, 2513–2523 (2021).

## Acknowledgements

We would like to acknowledge the NKI-AVL Core Facility Molecular Pathology & Biobanking (CFMPB) for supplying NKI-AVL biobank material. We would like to thank Gabriel Silva de Souza, Muradije Demirel-Andishmand, Luca Dulce Meesters and Daan Zegers for their help with data collection, scanning and making manual annotations. This project was funded by a NWO-VIDI grant (number 18388).

## Author contributions

LvE conceived the study, selected data, set up and coordinated manual annotations and reader studies, performed the experiments, analyzed the results and wrote the manuscript. GS, MDA, LM and DZ. made manual annotations, supervised by MLS and EM. MLS, SV, EM, IG, AE, KG, BA, CB and GS participated and annotated cases for the reader studies. LvdW, MvdH and WT were involved in collecting, setting up and running the PEMBRO-RT study and making cases available for this study, EM was involved in collecting NEG cases and MLS and SV were involved in collecting RUMC cases. FC conceived the study, co-designed experiments and supervised the work. BvG and JvdL co-supervised the work. All authors reviewed the manuscript and agreed with its contents.

## Competing interests

IG is a member of the advisory board for Roche Diagnostics. WT reports research funding from Astrazeneca, MSD and Sanofi/Regeneron. MvdH reports sponsorship or research funding from Amgen, Astrazeneca, BMS, Janssen Pharmaceutica, Stichting Treatmeds, Merck, MSD, Novartis, Pamgene, Pfizer, Roche and Roche diagnostics. BvG is a shareholder of Thirona BV, the Netherlands. JvdL was a member of the advisory boards of Philips, the Netherlands and ContextVision, Sweden, and received research funding from Philips, the Netherlands, ContextVision, Sweden, and Sectra, Sweden in the last five years. He is chief scientific officer (CSO) and shareholder of Aiosyn BV, the Netherlands. FC was chair of the Scientific and Medical Advisory Board of TRIBVN Healthcare, France, and received advisory board fees from TRIBVN Healthcare, France in the last five years. He is a shareholder of Aiosyn BV, the Netherlands. All other authors declare no potential conflict of interest.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-57067-1>.

**Correspondence** and requests for materials should be addressed to L.v.E.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024