Deep Learning-Based Evaluation of PD-L1 Immunohistochemistry-Stained Slides for Robust Tumor Proportion Scoring Enables Better Categorization of Non-Small-Cell Lung Cancer Cases

Corinna Wolf[1], Aslihan Gerhold-Ay[1], Kenneth Bloom[2], Christian Ihling[1], Hans Juergen Grote[1], Avi Laniado[2], Amit Bart[2], Christoph Rohde[1], Andrea Harpe[1], Angela Manginelli[1], Yuval Shachaf[2], Yoav Blum[2], Ori Zelichov[2], Lotan Chorev[2] and Thomas Mrowiec[1]

[1]The healthcare business of Merck KGaA, Darmstadt, Germany
[2]Nucleai Ltd, Tel-Aviv, Israel

CONCLUSION

We successfully developed a robust solution for PD-L1 scoring in NSCLC, generalizing across heterogenous data and the two most widely used clones, 22C3 and SP263.

Our data demon-strate that utilization of a DL-based solution enables identification of otherwise mis-categorized cases, particularly around the clinically relevant cut-offs of 1% and 50%.

Given the importance of accurate PD-L1 scoring and the limited inter-observer reproducibility among pathologists, DL assistance may be clinically useful.

INTRODUCTION

- Accurate and reproducible assess-ment of PD-L1 status is essential for treatment selection in patients with NSCLC.
- Pathologists' interobserver repro-ducibility at the 1% and 50% tumor proportion score (TPS) cut-offs is limited.
- We developed a novel deep learning (DL)-based solution for objective, scalable and reproducible analysis of PD-L1 immunohistochemistry (IHC).
- Our solution performs robustly for the two most widely used clones, 22C3 (Dako) and SP263 (Ventana).
- We evaluated the ability of our solution to assist pathologists in categorizing cases around clinically relevant cut-offs.

METHODS

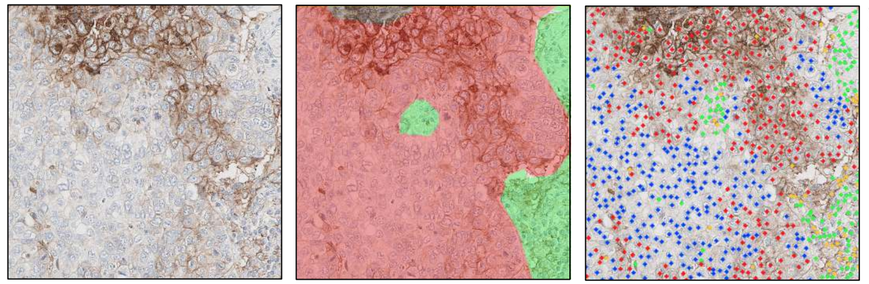
- We **trained a DL model for PD-L1 IHC scoring** of the tumor proportional score (digital DL TPS) on a heterogeneous data set of more than 100 whole slide images (Table 1) from procured **samples of various sources**, stained at **five different laboratories** with either **22C3 or SP263**, and scanned with **three different scanners**.
- The PD-L1 IHC image analysis workflow has been developed to serve an automated pipeline which includes automated tissue detection, exclusion of artefacts, tissue, region prediction for tumor, stroma and necrosis detection, as well as cell type detection and positivity prediction.
- To handle unforeseen, rare regions (e.g., lumina of adenocarcinoma) and for correction of miscategorized areas, a manual annotation option was included.
- Approximately 100,000 cells were annotated by expert pathologists to train the image analysis model (Figure 1). The model was validated with an unseen cohort of 107 slides and results were correlated with the assessment of two independent pathologists.

Table 1. Distribution of cases used for training and validating the area segmentation and cell identification models across two different clones

		22C3 (Dako)	SP263 (Ventana)	
Area model	Train	99	68	**167**
	Validation	16	19	**35**
Cell model	Train	30	102	**132**
	Validation	9	39	**48**

RESULTS

Figure 1 – Visualization of area prediction, cell identification and PD-L1 positivity by the DL model



- Our DL model identified positive and negative tumor cells as well as positive immune cells with high confidence (Figure 1). Cells not relevant for TPS or CPS scoring are identified as "other".

- A **CE-marked algorithm** for the analysis of PD-L1 (SP263) stained lung cancer specimen achieved a Spearman correlation coefficient **(R^2) of 0.81** with the average TPS of two pathologists (Figure 2, n=48).
- In comparison, **our digital DL TPS** achieved a higher **R^2 of 0.91** with the mean pathologist TPS in the heterogeneous validation cohort(Figure 3, n=107).

Figure 2. Correlation between CE-marked algorithm and mean pathologist TPS

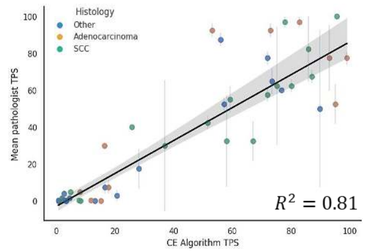
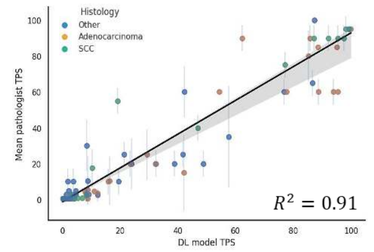


Figure 3. Correlation between digital DL TPS and mean pathologist TPS on validation cohort



- Correlation analysis for both clones individually reached an **R^2 of 0.80 for 22C3** (Figure 4) and **0.95 for SP263** (Figure 5), suggesting a better performance of the model for SP263.

Figure 4. Correlation between DL and manual assessment of TPS scores for clone 22C3

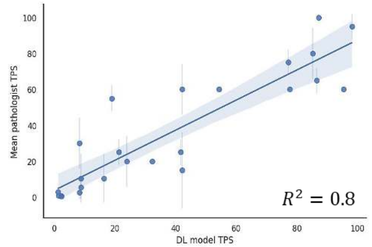


Figure 5. Correlation between DL and manual assessment of TPS scores for clone SP263

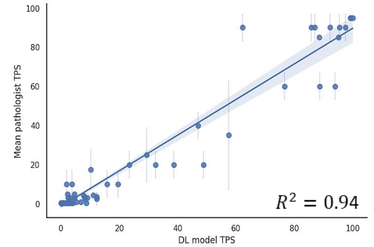
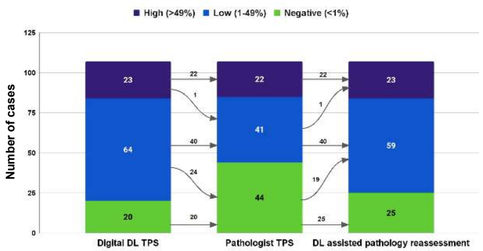


Figure 6. Categorization into clinically relevant groups by DL, independent and DL assisted pathologists



- Clinically relevant decisions depend on the categorization of cases into PD-L1 high (≥ 50%), low (1-49%) and negative (<1%).
- For PD-L1 high cases, independent categorization by pathologists had a 96% (22/23) agreement with the categorization according by DL.
- A higher discrepancy between DL and pathologists was observed around the 1% cut-off:
- While all 20 cases categorized negative by the DL were called negative by the pathologists as well,
- At least one pathologist called 24 cases negative that were categorized PD-L1 low by the DL (Figure 6).
- Discrepant cases were reassessed by the pathologists, taking advantage of the DL model's quantitative data output.
- **DL assisted reassessment led to re-categorization of 18% (19) of the cases into the PD-L1 positive, treatment-eligible group, which were originally categorized PD-L1 negative and, therefore, treatment-ineligible** (Figure 6).

Disclosure statement: Corinna Wolf, Hans Juergen Grote and Thomas Mrowiec are employees and stock owners of the healthcare business of Merck KGaA, Darmstadt, Germany. Aslihan Gerhold-Ay, Christian Ihling, Christoph Rohde, Andrea Harpe and Angela Manginelli are employees of the healthcare business of Merck KGaA, Darmstadt, Germany. Kenneth Bloom, Avi Laniado, Amit Bart, Yuval Shachaf, Yoav Blum, Ori Zelichov and Lotan Chorev are employees of Nucleai Ltd, Tel-Aviv, Israel.

Presented at the USCAP 2023 Annual Meeting name | March 11-16, 2023 | New Orleans, LA