

RESEARCH

Open Access



# Artificial Intelligence-driven image analysis for standardised programmed death-ligand 1 expression evaluation in non-small cell lung cancer

Chong Ge<sup>1,2†</sup>, Yi Shi<sup>1,3†</sup>, Wei Wang<sup>1,2,5</sup>, Anli Zhang<sup>1,2</sup>, Mengqi Huang<sup>4</sup>, Fang Zhao<sup>3</sup>, Ao Li<sup>3\*</sup>, Zhenzhong Feng<sup>4\*</sup>, Minghui Wang<sup>3\*</sup> and Haibo Wu<sup>1,2,5\*</sup>

## Abstract

**Background** Accurate assessment of programmed death-ligand 1 (PD-L1) immunohistochemical (IHC) expression is critical for immunotherapy in patients with non-small cell lung cancer (NSCLC). Yet, interpreting its staining is challenging, time-consuming, and causes inter-observer variability, potentially mis-stratifying patients. This necessitates the development of an artificial intelligence (AI) model to effectively quantify PD-L1 expression. Hence, we developed an AI-based deep-learning approach to automatically assess PD-L1 expression in NSCLC using IHC 22C3 assay-stained whole slide images (WSIs).

**Methods** A total of 706 patients with NSCLC were included in this study and 1212 WSIs were collected from three distinct study cohorts. We accurately matched the hematoxylin and eosin-stained images of the internal dataset with the IHC WSIs. Foreground regions containing tumor tissue were extracted from WSIs, and a multi-granular multiple-instance learning approach employing instance embeddings with coarse and fine granularities was implemented to extract patch-level morphological features. A multi-grained expression interpreter-based model aggregated these features to stratify PD-L1 expression status.

**Results** The model showed strong interpretive ability in all three cohorts and wide applicability to different specimen types. The macro-average area under the receiver operating characteristic curve (AUC) were 0.940/0.915/0.944 for surgical specimens, 0.955/0.844/0.865 for biopsy specimens, and 0.901/0.958/0.883 for metastases.

**Conclusion** This study emphasizes the potential benefits of deep learning in automatically, rapidly, and accurately inferring PD-L1 expression from complex IHC images. It also showcases how AI frameworks can improve routine digital pathology workflows in current PD-L1 detection methods.

<sup>†</sup>Chong Ge and Yi Shi contributed equally to this work.

\*Correspondence:

Ao Li  
aoli@ustc.edu.cn  
Zhenzhong Feng  
fzzapple1976@163.com  
Minghui Wang  
mhwang@ustc.edu.cn  
Haibo Wu  
wuhaibo@ustc.edu.cn

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

**Keywords** Programmed death-ligand 1, Whole slide imaging, Deep Learning, Artificial Intelligence, Non-small cell lung cancer, Immunohistochemistry

## Introduction

Lung cancer is one of the most prevalent and deadly malignancies worldwide, with ongoing advancements in its diagnosis and treatment garnering widespread attention [1]. Histologically, lung cancer is classified into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), with the latter accounting for approximately 80–85% of all cases and encompassing several subtypes, including adenocarcinoma, squamous cell carcinoma, and adenosquamous carcinoma [2]. Treatment strategies for NSCLC have evolved significantly over the years, reflecting our growing understanding of tumor biology and immune interactions. The integration of immunotherapy and targeted therapy in neoadjuvant or adjuvant settings for early-stage NSCLC, as well as in the treatment of advanced-stage disease, has markedly improved patient outcomes [3]. Immune checkpoint inhibitors (ICIs) targeting the programmed death 1 (PD-1) and programmed death-ligand 1 (PD-L1) pathway are at the forefront of this immunotherapeutic revolution. ICIs have become pivotal in immunotherapy, fundamentally altering the treatment paradigm for NSCLC [4–6]. The mechanism of action of these ICIs involves blocking the interaction between PD-1 on T cells and PD-L1 on tumor cells, thereby reactivating the immune system's ability to recognize and attack cancer cells [7]. This approach has yielded remarkable results, with the five-year survival rate for advanced NSCLC increasing from 5% in the chemotherapy era to 13.4–23.2% with immunotherapy [8].

PD-L1, a type I transmembrane protein belonging to the B7 family, is frequently overexpressed on the surface of tumor cells. It exerts immunosuppressive effects through binding to PD-1/PD-2 receptors, thereby inhibiting T cell activation and negatively regulating immune responses [9–11]. Moreover, PD-L1 has been demonstrated to induce T cell apoptosis and functional impairment, further compromising anti-tumor immunity [12–14]. As a clinically validated biomarker, PD-L1 expression guides ICI therapy selection. The U.S. Food and Drug Administration has approved the PD-L1 immunohistochemical assay (22C3; DAKO, Agilent Technologies) as a companion diagnostic for pembrolizumab treatment [15, 16]. In NSCLC, PD-L1 expression levels critically inform therapeutic decision-making. Extensive clinical studies have demonstrated that, compared with conventional chemotherapy, PD-1 monoclonal antibody monotherapy in patients with high PD-L1

expression exhibits superior progression-free survival and overall response rates [17]. For patients with low PD-L1 expression, first-line combination therapy with PD-1/PD-L1 inhibitors plus platinum-based doublet chemotherapy demonstrates superior clinical efficacy. Randomized controlled trials have demonstrated that this regimen significantly improves overall survival and enhances objective response rates alone [18].

Despite its clinical utility, the current scoring system for PD-L1 expression has several limitations. The assessment of PD-L1 expression relies heavily on the expertise and judgment of the pathologist, introducing an element of subjectivity that can lead to inter-observer variability, particularly in settings where pathologists have limited experience with PD-L1 interpretation [19, 20]. This subjectivity is further compounded by the inherent intratumoral heterogeneity of PD-L1 expression and its dynamic nature in response to various factors such as treatment modalities and the tumor microenvironment [21]. Moreover, the tissue used for analysis must accurately represent the tumor, a task complicated by the considerable heterogeneity of PD-L1 expression in NSCLC [22]. This heterogeneity increases the risk of false-positive and false-negative results depending on the representativeness of the tissue biopsy, potentially leading to suboptimal treatment decisions.

These challenges underscore the pressing need for more objective, standardized, and comprehensive assessment methods of PD-L1 expression. The rapid advancement of artificial intelligence (AI) and its application in computational pathology (CPATH) offers promising solutions to these challenges. CPATH is widely applied in various aspects of pathology, including cytology, histology, immunohistochemistry, molecular analysis, and tumor immune microenvironment (TIME) assessment [23, 24]. CPATH leverages whole-slide imaging (WSI) and AI algorithms to emulate the diagnostic reasoning of pathologists, enabling intelligent and automated analysis of PD-L1 expression [25–27]. This AI-powered quantitative approach has substantially enhanced the accuracy, reproducibility, and objectivity of PD-L1 scoring, offering a transformative paradigm for predicting immunotherapy response. Despite these advances, a critical challenge remains: the absence of standardized, clinically validated assessment models that can be seamlessly integrated into therapeutic decision-making. Addressing this gap is imperative to advance precision immunotherapy in NSCLC and optimize patient stratification [28, 29].

Here, we describe an AI-based platform for interpreting PD-L1 expression in digitized IHC slides of NSCLC specimens. We developed and validated a computational method that provides a rapid, accurate, and objective assessment of PD-L1 expression across different types of specimens, including radical surgical specimens, puncture or biopsy specimens, and metastases, thereby enhancing the precision of immunotherapy decision-making. This study aimed to contribute to the evolving landscape of digital pathology and precision oncology, potentially improving patient outcomes in NSCLC treatment. By addressing the current limitations of PD-L1 assessment, our study has the potential to optimize treatment selection, reduce inter-observer variability, and ultimately lead to more personalized and effective immunotherapy strategies for patients with NSCLC.

## Methods

We developed and validated an AI-based platform for interpreting PD-L1 expression in digitized IHC slides of NSCLC specimens. The AI algorithm was used to assess PD-L1 expression across distinct specimens, including radical surgical specimens, biopsy specimens, and metastases.

### Population and clinicopathological data

Patient selection criteria were as follows: (1) specimens obtained from the First Affiliated Hospital of the University of Science and Technology of China (USTC) cohort and the Second Affiliated Hospital of Anhui Medical University (AMU) cohort between October 2019 and December 2023; (2) diagnosis of NSCLC according to the diagnostic criteria of the 5th Edition of the WHO Classification of Thoracic Tumors [30]; (3) availability of clinicopathologic data, including age, gender, tumor stage, tumor site, pathological type, and PD-L1 expression IHC result.

All specimens were fixed with 10% neutral formalin, routinely dehydrated, embedded in paraffin, continuously sectioned with a thickness of 4 microns, and stained with hematoxylin and eosin (H&E). All PD-L1 staining was performed using the EnVision FLEX detection system, dyeing equipment (Autostainer Link48, DAKO, Agilent Technologies, Denmark), dyeing reagent (22C3; DAKO, Agilent Technologies, Denmark). All H&E and IHC-stained slides were digitized at 40X magnification using a Sqray Slide Scanning System SQS-600P (Shenzhen Shengqiang Technology Co., Ltd., Shenzhen, China) and converted to SVS format. Cases with suboptimal section quality or those that were unsuitable for image feature extraction were excluded.

All IHC interpretation results were re-evaluated by three senior pathologists, and cases with consistent

findings were retained. The assessment of PD-L1 expression status is primarily determined by evaluating the IHC tumor proportion score (TPS), with <1% positive staining of tumor cells considered negative expression,  $1\% \leq \text{TPS} < 50\%$  indicating low expression, and  $\geq 50\%$  signifying high expression [31].

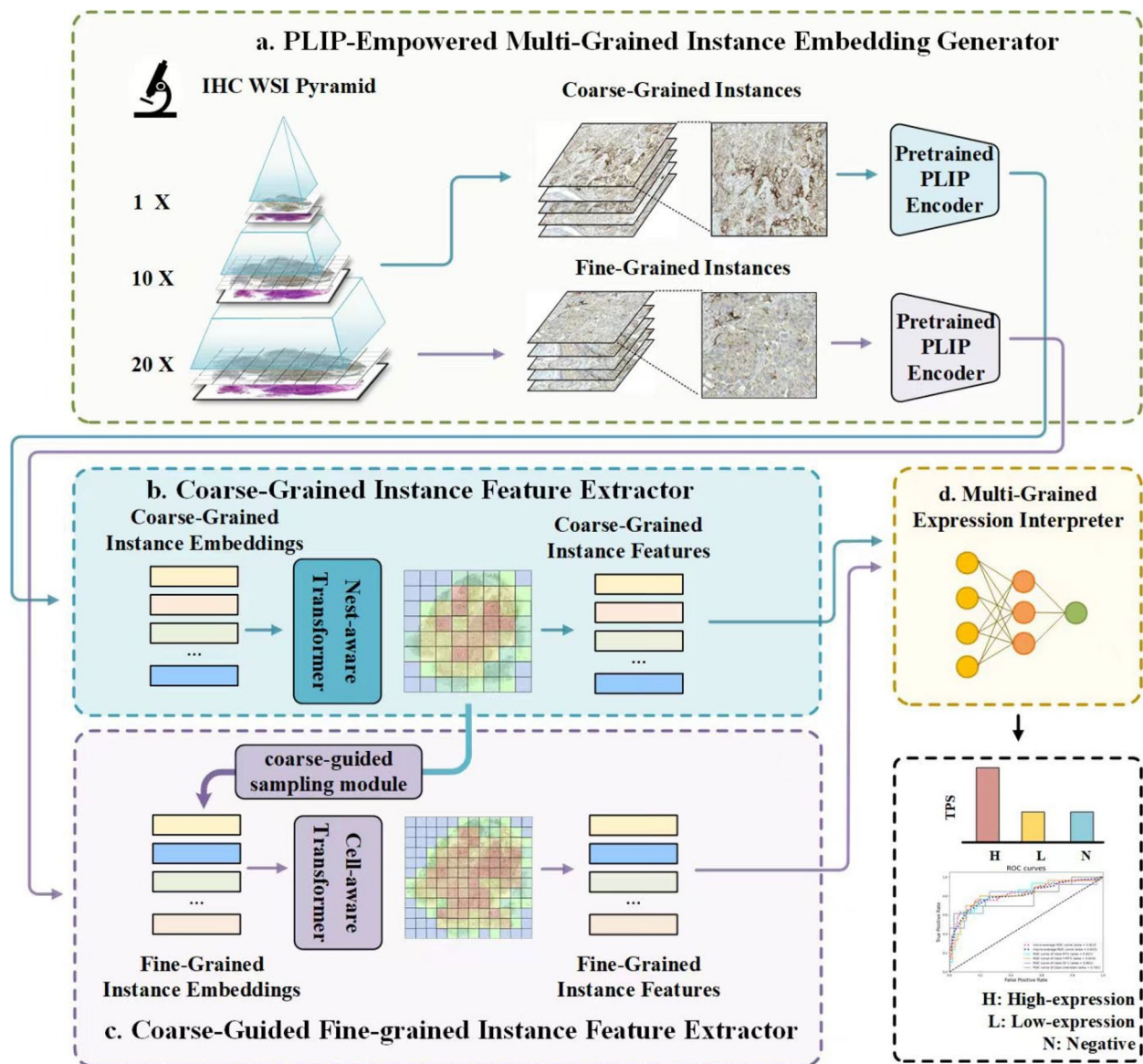
The USTC cohort comprised 294 eligible patients randomly divided into a training set ( $n=224$ ) and an internal validation set ( $n=70$ ) for model development and assessment. For external validation, 212 patients were included in the AMU cohort. Additionally, our model was tested on a publicly available Memorial Sloan Kettering-Multimodal Integration of Data (MSK-MIND) dataset containing 247 WSIs from 247 patients, and 200 samples were included in the final analysis after excluding blurry images [32].

### Computational framework

To achieve IHC-based PD-L1 status interpretation, we developed a multi-grained, multi-instance learning model called MG-MIL. As shown in Fig. 1, the MS-MIL model comprises four main components: (a) a Pathology Language-Image Pretraining (PLIP)-empowered multi-grained instance embedding generator, (b) a coarse-grained instance feature extractor, (c) a coarse-guided fine-grained instance feature extractor, and (d) a multi-grain expression interpreter. First, we leverage the capabilities of a large medical foundation model to build a PLIP-empowered multi-grained instance embedding generator that generates instance embeddings with coarse and fine granularities. The PLIP is a large-scale pathology-specific vision-language model pretrained on the OpenPath dataset, which contains 200,000+ pathological images with clinical data. Its image encoder excels at recognizing pathological patterns. In our work, we leveraged PLIP's encoder to extract tile features for accurate PD-L1 assessment. Second, a coarse-grained instance feature extractor mines the tissue structure information from instance embeddings to produce a global attentive map and coarse-grained instance features. Third, we designed a coarse-guided fine-grained instance feature extractor that harnessed a global attentive map to instruct the generation of fine-grained instance features. Finally, a multi-grained expression interpreter integrated coarse- and fine-grained features, enabling the stratification of PD-L1 expression status.

### PLIP-empowered multi-grained instance embedding generator

Inspired by advances in medical foundation large models, such as PLIP [33] We introduced a PLIP-empowered multi-grained embedding generator to harvest the hierarchical information inherent in huge-size WSIs [34,



**Fig. 1** Computational framework. **a** PLIP-empowered multi-grained instance embedding generator, which generated instance embeddings with coarse and fine granularities. **b** A coarse-grained instance feature extractor mined tissue structure information from instance embeddings, producing a global attentive map and coarse-grained instance features. **c** A coarse-guided fine-grained instance feature extractor, which harnessed the global attentive map to instruct the generation of fine-grained instance features. **d** A multi-grained expression interpreter integrated coarse- and fine-grained features, enabling stratification of PD-L1 expression status

35]. This component processes IHC-stained pathological image tile instances at high or low magnification using the pre-trained image encoder of PLIP, deriving coarse- and fine-grained instance embeddings. The operation consisted of three procedures: 1) background removal using the popular HistoQC method [36], 2) instance tiling at different magnifications (10X and 20X), and 3) embedding generation using the pretrained image encoder of PLIP to process coarse- and fine-grained embedding sequences.

#### Coarse-grained instance feature extractor

Pathologists can observe cancer nest regions at low magnifications and recognize diverse types of cells at high magnifications. Therefore, to mimic zoom-in and zoom-out operations, we developed coarse- and fine-grained instance feature extractors based on Transformer [37], exploiting such hierarchical information across different granularities. The coarse-grained instance feature extractor consists of a transformer encoder with multi-head self-attention blocks [38], and the input of each block is

the output of the former block. In addition, we can obtain a global attentive map by summarizing attention matrices, which can denote informative areas such as cancer nests and TIL structures [39].

#### Coarse-guided fine-grained instance feature extractor

Except for coarse-grained features, fine-grained information is also important for precise molecular expression assessment, as the clinical interpretation of PD-L1 status is established for discriminating between various cells such as lymphocytes and tumor cells, which can be recognized with high magnification while being unseen at low magnification. Therefore, we propose a coarse-guided, fine-grained instance feature extractor that includes a coarse-guided sampling module and a transformer-based feature extractor. Specifically, the global attentive map denoted the most task-relevant regions, and the fine-grained branch was used only to analyze the most pre-dominant regions. To achieve this, we first sampled the most focused regions in the attentive map and projected their locations at 20X magnification. We then leveraged a transformer encoder similar to a coarse-grained branch to derive the fine-grained instance features.

#### Multi-grained PD-L1 expression interpreter

To integrate precious multi-grained instance features, we first combined coarse- and fine-grained instance features by concatenation to pool them effectively and employed a multilayer perceptron consisting of a sequence of fully connected layers to explore the pooled feature. Finally, we deployed softmax activation on the final three output neurons to calculate the class distribution probabilities for different PD-L1 expression levels, that is, negative, low expression, and high expression:

$$y = \text{softmax}(x_{MG}) = \left\{ \frac{\exp(x_{MGi})}{\sum_j^c (\exp(x_{MGj}))} \right\}_i^c$$

where  $y$  is the output probability of three PD-L1 status;  $\text{softmax}(\cdot)$  denotes the employed softmax normalisation layer;  $x_{MG}$  represents the processed pathological feature of multi-layer perceptron;  $c$  is the number of output classes.

#### Statistics

Categorical measures of clinicopathological variables were compared using the chi-squared test or Fisher's exact test. Statistical significance was set at  $p < 0.05$ . Analyses were performed using the SPSS software (version 25.0; IBM Corp., Armonk, NY, USA). In the testing phase, the results generated by the AI model include the probabilities of interpretation for each category

Among these probabilities, the category with the highest probability is designated as the final Interpret category of the AI model. The final interpretation results are then compared with the actual outcomes to assess the accuracy of the AI model. Additionally, the Sensitivity, Specificity, and accuracy (ACC) as important indicators of the model for quantitative assessment can sensitively measure the PD-L1 interpretation accuracy of the AI model. Furthermore, the area under the ROC curve and the AUC for each category are evaluated as another quantitative indicator for overall assessment. These metrics comprehensively reflect the classification performance of the model at different thresholds, providing an important benchmark for evaluating the generalization ability of the AI model.

## Results

### Clinicopathological and PD-L1 expression features of the study cohorts

The training and internal validation sets of the USTC cohort comprised 294 patients, of whom 59.5% were male, 40.5% were female. Most patients (77.2%) had lung adenocarcinoma, 22.1% had lung squamous cell carcinoma, and other histological types of NSCLC accounted for 0.7%. Of the 294 patients, 201 underwent surgical resection and 93 underwent biopsy (including 26 metastatic biopsies). For external validation, the AMU cohort included 212 patients (67.5% male, 32.5% female), of which 66% had lung adenocarcinoma, 33.5% had lung squamous cell carcinoma, and 0.5% had other NSCLC histological types. Surgical resection was performed in 105 patients, and biopsy was performed in 107 patients. Additionally, clinicopathological data and corresponding WSIs for 200 samples (41.5% male, 58.5% female) were obtained from the MSK-MIND dataset, of which 74% had lung adenocarcinoma, 14.5% had lung squamous cell carcinoma, and 11.5% had other NSCLC histological types. Surgical resection and biopsy were performed in 9 and 191 patients (including 111 metastatic), respectively. Age, clinical stage, tumor site, specimen type, histological type, and PD-L1 expression results of the three cohorts are shown in Table 1.

Of the 294 cases in the USTC cohort, 104 were negative (35.4%), 129 showed low expression (43.9%), and 61 showed high expression (20.7%) for PD-L1. For external validation of the AMU cohort, negative results were detected in 49 patients (23.1%), low expression in 129 patients (60.9%), and high expression in 34 patients (16.0%). In addition, the MSK-MIND dataset, which included 69 (34.5%) negative cases, 36 (18.0%) low expression cases, and 95 (47.5%) high expression cases,

**Table 1** Baseline characteristics in the internal and external sets

| Variables          | USTC cohort (N = 294) (%) | AMU cohort (N = 212) (%) | MSK-Mind dataset (N = 200) (%) | P      |
|--------------------|---------------------------|--------------------------|--------------------------------|--------|
| Gender             |                           |                          |                                | < 0.01 |
| male               | 175 (59.5)                | 143 (67.5)               | 83 (41.5)                      |        |
| female             | 119 (40.5)                | 69 (32.5)                | 117 (58.5)                     |        |
| Age                |                           |                          |                                | 0.01   |
| ≤ 60               | 120 (40.8)                | 67 (31.6)                | 51 (25.5)                      |        |
| > 60               | 174 (59.2)                | 145 (68.4)               | 149 (74.5)                     |        |
| Clinical stage     |                           |                          |                                | 0.469  |
| I                  | 71 (24.1)                 | 42 (19.8)                | /                              |        |
| II                 | 101 (34.4)                | 85 (40.1)                | /                              |        |
| III                | 80 (27.2)                 | 59 (27.8)                | /                              |        |
| IV                 | 42 (14.3)                 | 26 (12.3)                | /                              |        |
| Tumor site         |                           |                          |                                | < 0.01 |
| Right Lung         | 179 (60.9)                | 121 (57.0)               | /                              |        |
| Left Lung          | 102 (34.7)                | 83 (39.2)                | /                              |        |
| Metastasis         | 13 (4.4)                  | 8 (3.8)                  | 111 (55.6)                     |        |
| Specimen type      |                           |                          |                                | < 0.01 |
| Surgical specimens | 201 (68.4)                | 105 (49.5)               | 9 (4.5)                        |        |
| Biopsy specimens   | 93 (31.6)                 | 107 (50.5)               | 191 (95.5)                     |        |
| Histological type  |                           |                          |                                | < 0.01 |
| Adenocarcinoma     | 227 (77.2)                | 140 (66)                 | 148 (74)                       |        |
| Squamous Carcinoma | 65 (22.1)                 | 71 (33.5)                | 29 (14.5)                      |        |
| Others             | 2 (0.7)                   | 1 (0.5)                  | 23 (11.5)                      |        |
| TPS of PD-L1       |                           |                          |                                | < 0.01 |
| < 1%               | 104 (35.4)                | 49 (23.1)                | 69 (34.5)                      |        |
| ≥ 1% and < 50%     | 129 (43.9)                | 129 (60.9)               | 36 (18.0)                      |        |
| ≥ 50%              | 61 (20.7)                 | 34 (16.0)                | 95 (47.5)                      |        |

**Table 2** Summary of PD-L1 detection by Immunohistochemical detection

| Dataset                        |                    | Total N (%) | Negative N (%) | Low expression N (%) | High Expression N (%) |
|--------------------------------|--------------------|-------------|----------------|----------------------|-----------------------|
| <b>Training set</b>            | Surgical specimens | 159 (71)    | 66 (29.5)      | 64 (28.6)            | 29 (12.9)             |
|                                | Biopsy specimens   | 65 (29)     | 15 (6.7)       | 34 (15.2)            | 16 (7.1)              |
| <b>Internal validation set</b> |                    |             |                |                      |                       |
| <b>USTC cohort</b>             | Surgical specimens | 40 (57.1)   | 17 (24.3)      | 15 (21.4)            | 8 (11.4)              |
|                                | Biopsy specimens   | 17 (24.3)   | 4 (5.7)        | 9 (12.9)             | 4 (5.7)               |
|                                | Metastasis         | 13 (18.6)   | 2 (2.9)        | 7 (10.0)             | 4 (5.7)               |
| <b>External validation set</b> |                    |             |                |                      |                       |
| <b>AMU cohort</b>              | Surgical specimens | 104 (49.1)  | 32 (15.1)      | 61 (28.8)            | 11 (5.2)              |
|                                | Biopsy specimens   | 100 (47.1)  | 15 (7.0)       | 64 (30.2)            | 21 (9.9)              |
|                                | Metastasis         | 8 (3.8)     | 2 (0.9)        | 4 (2.0)              | 2 (0.9)               |
| <b>MSK-Mind dataset</b>        | Surgical specimens | 9 (4.5)     | 3 (1.5)        | 2 (1.0)              | 4 (2.0)               |
|                                | Biopsy specimens   | 80 (40.0)   | 28 (14.0)      | 16 (8.0)             | 36 (18.0)             |
|                                | Metastasis         | 111 (55.5)  | 38 (19.0)      | 18 (9.0)             | 55 (27.5)             |

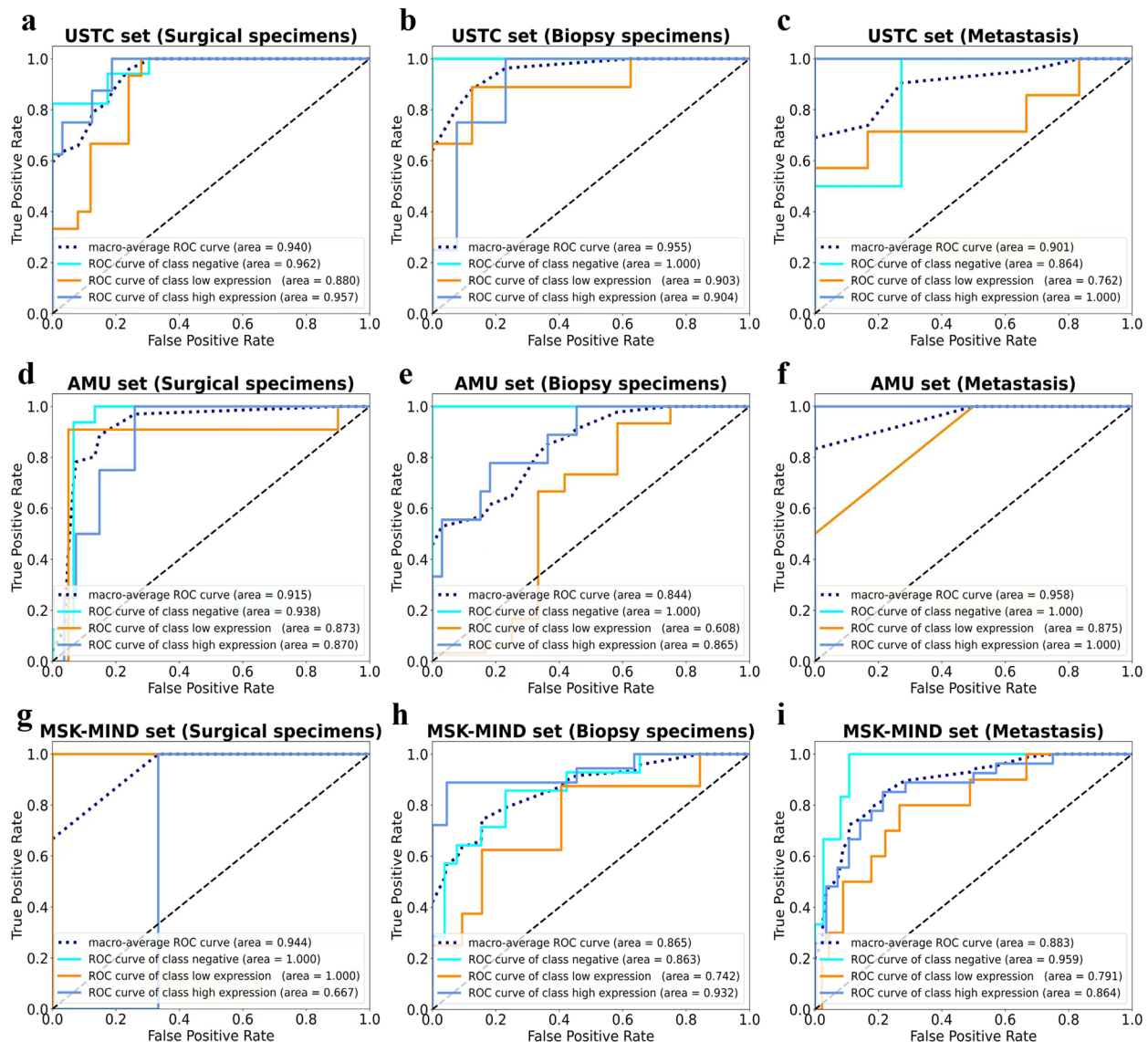
was used to validate the model. The patient distribution and TPS results are summarized in Table 2.

### Expression interpretation of validation experiments

For internal validation, the performance of the established model was evaluated on 70 unseen WSIs comprising three categories: negative (17 surgical specimens, 4 biopsy specimens, and 2 metastases), low expression (15 surgical specimens, 9 biopsy specimens, and 7 metastases), and high expression (8 surgical specimens, 4 biopsy

specimens, and 4 metastases). The model achieved patient-level areas under the receiver operating characteristic (ROC) curve (AUCs) of 0.940 for surgically resected samples, 0.955 for biopsies, and 0.901 for metastases (Fig. 2a–c). The sensitivity and specificity were 76.40% and 87.70% for surgical specimens, 88.80% and 93.30% for biopsies, and 73.8% and 88.4% for metastases, respectively (Table 3).

For external validation, 212 cases (108 surgically resected samples, 106 biopsies, and 8 metastases) were



**Fig. 2** ROC curve of PD-L1 interpretation in each cohorts and different specimens. Across the three validation sets, receiver operating characteristic (ROC) curves demonstrate the interpretation performance (quantified by area under the curve, AUC) for PD-L1 expression stratification across different specimen types. The top panel (a–c) illustrates AUC values for surgical resections (a), biopsy samples (b), and metastatic lesions (c) in the USTC cohort. The middle panel (d–f) presents corresponding AUC data for surgical (d), biopsy (e), and metastatic (f) specimens from the AMU cohort. Finally, the bottom panel (g–i) displays the AUC performance for surgical (g), biopsy (h), and metastatic (i) samples in the MSK-Mind dataset

**Table 3** Classification results on internal and external test sets

| Dataset                        |                    | AUC   | ACC   | Sensitivity | Specificity |
|--------------------------------|--------------------|-------|-------|-------------|-------------|
| <b>Internal validation set</b> |                    |       |       |             |             |
| <b>USTC cohort</b>             | Surgical specimens | 0.940 | 0.775 | 0.764       | 0.877       |
|                                | Biopsy specimens   | 0.955 | 0.882 | 0.88        | 0.933       |
|                                | Metastasis         | 0.901 | 0.769 | 0.738       | 0.884       |
| <b>External validation set</b> |                    |       |       |             |             |
| <b>AMU cohort</b>              | Surgical specimens | 0.915 | 0.806 | 0.756       | 0.878       |
|                                | Biopsy specimens   | 0.844 | 0.786 | 0.767       | 0.750       |
|                                | Metastasis         | 0.958 | 0.778 | 0.750       | 0.867       |
| <b>MSK-Mind dataset</b>        | Surgical specimens | 0.944 | 0.782 | 0.767       | 0.806       |
|                                | Biopsy specimens   | 0.865 | 0.775 | 0.777       | 0.866       |
|                                | Metastasis         | 0.883 | 0.787 | 0.774       | 0.886       |

recruited in the AMU cohort. The model achieved AUCs of 0.958 for surgical specimens, 0.844 for biopsies, and 0.894 for metastases (Fig. 2d–f). Consequently, the overall sensitivity and specificity were 79.2% and 91.7%, respectively, for the external validation set; 75.6% and 87.8%, respectively, for surgical specimens; 76.70% and 75.0%, respectively, for biopsy samples; and 75.0% and 86.7%, respectively, for metastases (Table 3). External validation of independent cohorts from the AMU cohort, predominantly including surgical biopsy samples and metastases, demonstrated the generalizability of this model.

To conduct more extensive external validation, the model was evaluated using a diverse and independent MSK-MIND dataset, which encompassed a wide range of histopathological features and expression results for PD-L1 (Table 1). The model achieved an AUC of 0.944, 0.865, and 0.883 for surgical specimens/biopsy samples/metastasis and reached patient-level sensitivity/specificity of 76.7/80.6%, 77.7/86.6%, and 77.7/88.6% for different sample types, respectively (Fig. 2g–i and Table 3).

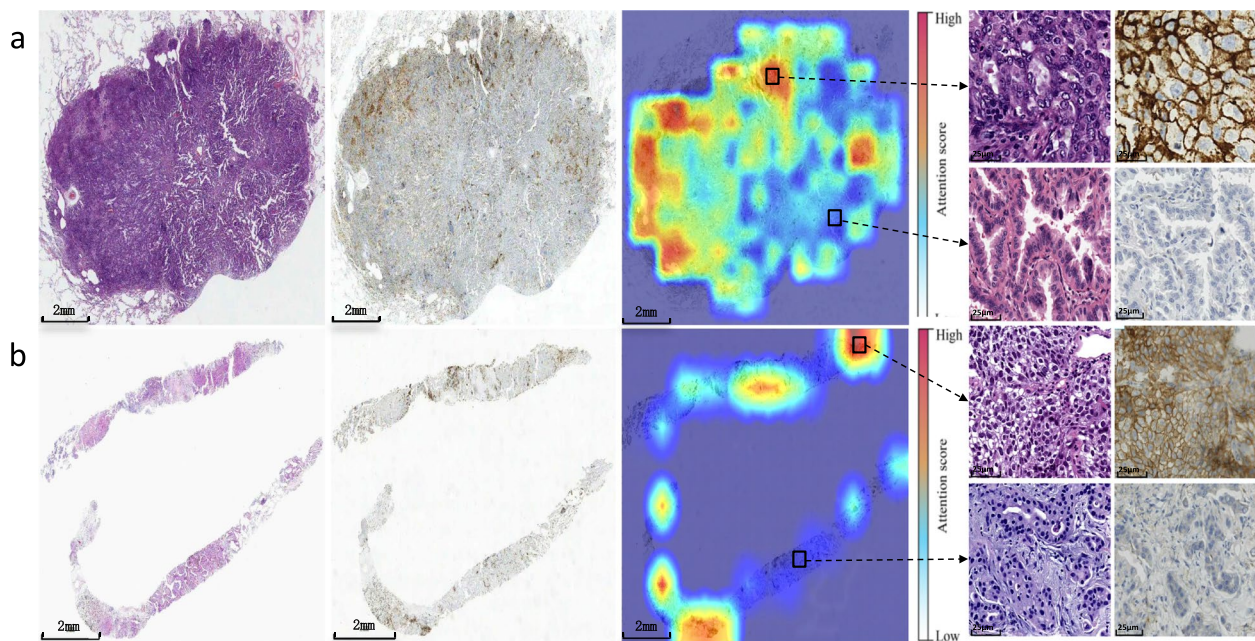
**Model interpretability**

To prove the interpretability of this computational model and to address potential deformations between different stains that could affect knowledge transfer from IHC to H&E, we performed delicate WSI registration between different stains of the internal dataset USTC. Attention heat maps were generated for each WSI, which served as visual representations of the model’s output attention scores. These heat maps assisted in distinguishing areas indicative of negative and positive PD-L1 expression. As shown in Fig. 3, the WSI heatmaps of H&E and IHC visually indicate that the model identified regions with positive PD-L1 expression in tumor regions. Our model leverages advanced neural network architectures and

image processing techniques, through the precise pairing of HE and IHC’s WSI, to effectively extract and identify PD-L1 positive expression regions from complex IHC images. As depicted in Fig. 3, the regions of high-attention areas denote PD-L1 positive expression of tumor cells, which are mainly manifested by any visible and well-defined partial or complete membrane linear staining of tumor cells stronger than the background. In low-attention regions, PD-L1 is negatively expressed in tumor cells. In addition, attention heat maps derived by our model showed that in the WSI generated from different specimens, the visualization results can identify different expression regions of PD-L1, implying the wide applicability and particular interpretability of the proposed AI model.

**Discussion**

In this study, we developed an AI-based multiple-instance learning (MIL) model called MG-MIL to qualitatively stratify expression in digital PD-L1 IHC slides of samples from patients with NSCLC. The performance of tumor region recognition by the embedded attention-based approach was consistent with the IHC images with different PD-L1 expression levels, which were able to accurately identify the tumor area and precisely distinguish PD-L1 membrane-positive tumor cells, showing a high level of agreement with actual expression levels. These rigorous results suggest that MG-MIL is comparable to pathologist diagnosis in assessing PD-L1 expression and is more economical, convenient, and efficient. It also shows strong interpretative ability across different specimen types and adapts well to the diversity of clinical specimens. In addition, this system is useful for different specimen types and shows high consistency in the classification of PD-L1 results for surgical specimens, biopsy samples, as well as metastases.



**Fig. 3** Heatmaps for PD-L1 interpretation. The attention heatmap analysis from our AI model, for surgical (a) and biopsy (b) specimens, we provide a systematic comparison from left to right: low-magnification overview of H&E-stained sections, overview thumbnails of the corresponding IHC slides, model-generated attention heatmap overlay, representative high-power fields showing randomly selected regions of H&E staining and matching high-power fields of PD-L1 IHC expression

The PD-L1 IHC assay (22C3) has been approved as a companion diagnostic tool for ICIs targeted therapy; however, the guidelines for 22C3 staining assessment are subjective, and scoring variability among pathologists is an ongoing concern. The traditional manual interpretation of PD-L1 results has many limitations. First, an accurate assessment of PD-L1 IHC expression requires pathologists to have professional training and expertise, as well as continuous practice over a long period, which results in a long training cycle and consumes considerable manpower and material resources. In addition, pathologists still do not have a perfect method to address the problem of accuracy and interpretation differences in PD-L1 TPS scoring, thus requiring repeated interpretations by multiple pathologists or experts to reach a consensus on a comprehensive interpretation, which increases the overall workload. Finally, in some local hospitals, the lack of experienced pathologists makes it difficult to ensure the accuracy of the results. These limitations affect PD-L1 expression assessment and directly influence the precision of treatment for patients with NSCLC. Therefore, a novel approach is required to address these problems.

Deep learning-based convolutional neural networks (CNNs) have been developed for pathological image analysis and processing [40]. Some studies have established different DL models for evaluating or predicting PD-L1

and have shown strong explanatory and predictive power. For instance, Shamaï et al. [41]. used manual data annotation and a fully supervised learning strategy to generate a dataset of 3376 patients to evaluate the feasibility of H&E for predicting PD-L1 in breast cancer, resulting in a high AUC of 0.91–0.93. In another study, Sha et al. [42]. developed and trained a deep learning model to recognize HE imaging patterns for tumor PD-L1 status prediction in NSCLC and presented over 145,000 examples to learn HE images. The trained model reliably predicted the PD-L1 status in the “unseen” test cohort (AUC=0.80). These models with enhanced recognition compared with manual labelling of the target area by pathologists showed excellent performance, comparable to the interpretation results of multiple pathologists. However, the ability of CNNs to process pathological images is challenged by tumor heterogeneity and the limited availability of reference data, which often requires cumbersome labelled data to provide accurate references. Due to the labor-intensive nature of annotating such data, it is not suitable for practical work or multi-center collaboration. Another representative study by Ansh Kapil et al. [43]. used a generative adversarial networks model to conduct a systematic analysis of the functional and morphological characteristics of epithelial regions in PD-L1 antibody-stained digital histopathological WSIs, successfully stratifying the TPS, and the results showed high agreement

with the pathologist score of 0.96. This method does not require manual annotation by a pathologist, making it feasible for clinical applications. However, the training set is small and contains only single-center data, which lacks the diversity of multicenter samples, and the results lack extensive verification.

The above methods can overcome the limitations of manual interpretation of PD-L1 and show good interpretation or prediction ability. In recent years, the development of the MIL method and its advantages in processing high-pixel images have provided a boost for WSI analysis of pathological sections [44–46]. We developed a MIL model that demonstrated high accuracy in PD-L1 interpretation, achieving an AUC of 0.955 for internal validation and 0.958 for external validation. Furthermore, our model offers three significant advantages while maintaining strong interpretative performance: (1) traditional PD-L1 assays require complex manual labelling at the regional level, which is time-consuming, labor-intensive, and costly in cross-center practices, thereby limiting their clinical applicability. In contrast, our method requires only simple slide-level labelling for training, providing substantial advantages for application generalization; (2) Several of these methods utilize multistage and multimodal approaches, resulting in high complexity and heterogeneity of AI models, which can complicate application deployment and training procedures. In contrast, our approach employs an end-to-end framework that eliminates the need for a complex multistage process, aligning more closely with the clinical demands for rapid diagnosis; (3) Different cohort datasets and specimen types ensure a considerable amount of data, increasing the wide applicability of the model and allowing it to be well adapted to clinical diagnosis and treatment.

This study has several limitations. First, although we encompassed two institutes in this study, the recruited cohort involved was still relatively small, and more data are required from different institutions to increase the robustness of the AI model. Secondly, the WSI analysis model based on multi-instance learning lacks the ability to discriminate cell-level semantic information, which may lead to certain deviations in the interpretation of PD-L1 expression levels. Finally, the AI system was only tested on 22C3 specimens and should be further validated using more antibodies.

## Conclusion

In conclusion, we successfully developed an AI algorithm that can classify the results of PD-L1 from the WSI of IHC, which can assist pathologists in interpretation by reducing interpretation errors and improving interpretation accuracy, thus benefitting the individualized and precise treatment of patients with NSCLC. Extending

the model to other IHC examinations may enable a wider range of precise pathology applications. Overall, this study highlights the potential of CPATH techniques for extracting discriminative information from routine slides, thereby enhancing traditional pathology to guide personalized treatments for patients with cancer.

## Abbreviations

|       |                                      |
|-------|--------------------------------------|
| CNNs  | Convolutional neural networks        |
| CPATH | Computational pathology,             |
| DL    | Deep learning,                       |
| ICIs  | Immune checkpoint inhibitors         |
| ML    | Machine learning                     |
| MIL   | Multiple-instance learning           |
| NSCLC | Non-small cell lung cancer           |
| PD-1  | Programmed death 1                   |
| PD-L1 | Programmed death-ligand 1            |
| SCLC  | Small cell lung cancer               |
| TIME  | Tumour immune microenvironment       |
| TPS   | Tumour proportion score              |
| ROC   | Receiver Operating Characteristic    |
| PLIP  | Pathology Language–Image Pretraining |

## Acknowledgements

School of Information Science and Technology, University of Science and Technology of China: Special thanks to Professor Li Ao, Shi Yi and Zhao Fang for the design and research of computer algorithms. The Second Affiliated Hospital of Anhui Medical University: Thank Feng Zhengzhong and Huang Mengqi for providing the cases and sections.

## Authors' contributions

CG: Writing – original draft, review, & editing, Visualization, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. YS and WW: Writing-original draft, Resources, Methodology, Investigation, Data curation. FZ: Visualization, Methodology, Investigation, Formal analysis, Data curation. MH and AZ: Resources, Data curation, Investigation. ZF: Supervision, Resources. AL: Validation, Supervision, Resources, Project administration. MW: Visualization, Validation, Supervision, Software, Project administration. HW: Writing-review & editing, Visualization, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. All authors read and approved the final manuscript.

## Funding

This work was supported by the Joint Fund for Medical Artificial Intelligence under Grants MAI2023C014; Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM under Grants No. 2023IHM01043; National Natural Science Foundation of China under Grants 61871361, 62371435 and 62272325; Anhui Provincial Natural Science Foundation under Grant 2308085MF191.

## Data availability

No datasets were generated or analysed during the current study.

## Declarations

### Ethics approval and consent to participate

The study has been reviewed by the First Affiliated Hospital of USTC, was approved by ethics commission (No. 2024-RE-198).

### Consent for publication

Not Applicable.

### Competing interests

The authors declare no competing interests.

## Author details

<sup>1</sup>Department of Pathology, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230036, China. <sup>2</sup>Intelligent Pathology Institute, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230036, China. <sup>3</sup>The School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China. <sup>4</sup>Department of Pathology, The Second Affiliated Hospital of Anhui Medical University, Hefei, Anhui 230601, China. <sup>5</sup>Department of Pathology, Centre for Leading Medicine and Advanced Technologies of IHM, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230001, China.

Received: 16 February 2025 Accepted: 15 August 2025

Published online: 26 September 2025

## References

- Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2022. *CA Cancer Clin.* 2022;72(1):7–33.
- Hendriks LEL, Remon J, Faivre-Finn C, et al. Non-small-cell lung cancer. *Nat Rev Dis Primers.* 2024;10(1):71.
- Liu SM, Zheng MM, Pan Y, et al. Emerging evidence and treatment paradigm of non-small cell lung cancer. *J Hematol Oncol.* 2023;16(1):40.
- Rousseau B, Foote MB, Maron SB. The spectrum of benefit from checkpoint blockade in hypermutated tumours. *N Engl J Med.* 2021;384(12):1168–70.
- Reck M, Remon J, Hellmann MD. First-line immunotherapy for non-small-cell lung cancer. *J Clin Oncol.* 2022. <https://doi.org/10.1200/JCO.21.01497>.
- Conroy M, Forde PM. Advancing neoadjuvant immunotherapy for lung cancer. *Nat Med.* 2023;29(3):533–4.
- Chu X, Tian W, Wang Z, et al. Co-inhibition of TIGIT and PD-1/PD-L1 in cancer immunotherapy: mechanisms and clinical trials. *Mol Cancer.* 2023;22(1):93.
- Lahiri A, Maji A, Potdar PD, et al. Lung cancer immunotherapy: progress, pitfalls, and promises. *Mol Cancer.* 2023;22(1):40.
- Akinleye A, Rasool Z. Immune checkpoint inhibitors of PD-L1 as cancer therapeutics. *J Hematol Oncol.* 2019;12(1):92.
- Mandai M, Hamanishi J, Abiko K, et al. Dual faces of IFN $\gamma$  in cancer progression: a role of PD-L1 induction in the determination of pro- and antitumour immunity. *Clin Cancer Res.* 2016;22(10):2329–34.
- Dai X, Gao Y, Wei W. Post-translational regulations of PD-L1 and PD-1: mechanisms and opportunities for combined immunotherapy. *Semin Cancer Biol.* 2022;85:246–52.
- Sun C, Mezzadra R, Schumacher TN. Regulation and function of the PD-L1 checkpoint. *Immunity.* 2018;48(3):434–52.
- Yi M, Niu M, Xu L, et al. Regulation of PD-L1 expression in the tumour microenvironment. *J Hematol Oncol.* 2021;14(1):10.
- Lee D, Cho M, Kim E, et al. PD-L1: From cancer immunotherapy to therapeutic implications in multiple disorders. *Molecular Therapy.* 2024;32(12):4235–55.
- Büttner R, Gosney JR, Skov BG, et al. Programmed death-ligand 1 immunohistochemistry testing: a review of analytical assays and clinical implementation in non-small-cell lung cancer. *J Clin Oncol.* 2017;35(34):3867–76.
- Maule JG, Clinton LK, Graf RP, et al. Comparison of PD-L1 tumour cell expression with 22C3, 28–8, and SP142 IHC assays across multiple tumour types. *J Immunother Cancer.* 2022;10(10):e005573.
- Pérol M, Felipe E, Dafni U, et al. Effectiveness of PD-(L)1 inhibitors alone or in combination with platinum-doublet chemotherapy in first-line (1L) non-squamous non-small-cell lung cancer (Nsq-NSCLC) with PD-L1-high expression using real-world data. *Ann Oncol.* 2022;33(5):511–21.
- Ren S, Wang X, Han BH, et al. First-line treatment with camrelizumab plus farnitinib in advanced or metastatic NSCLC patients with PD-L1 TPS  $\geq 1\%$ : results from a multicenter, open-label, phase 2 trial. *J Immunother Cancer.* 2024;12(2):e007227.
- McLaughlin J, Han G, Schalper KA. Quantitative assessment of the heterogeneity of PD-L1 expression in non-small-cell lung cancer [published correction appears in *JAMA Oncol* 2016;2(1):46–54.
- Cooper WA, Russell PA, Cherian M, et al. Intra- and interobserver reproducibility assessment of PD-L1 biomarker in non-small cell lung cancer. *Clin Cancer Res.* 2017;23(16):4569–77.
- Brunnström H, Johansson A, Westbom-Fremer S, et al. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: inter-pathologist variability is higher than assay variability. *Mod Pathol.* 2017;30(10):1411–21.
- Gniadek TJ, Li QK, Tully E, et al. Heterogeneous expression of PD-L1 in pulmonary squamous cell carcinoma and adenocarcinoma: implications for assessment by small biopsy. *Mod Pathol.* 2017;30(4):530–8.
- Bera K, Schalper KA, Rimm DL, et al. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol.* 2019;16(11):703–15.
- Jiang Y, Yang M, Wang S, et al. Emerging role of deep learning-based artificial intelligence in tumour pathology. *Cancer Commun (Lond).* 2020;40(4):154–66.
- Baxi V, Edwards R, Montalto M, et al. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Mod Pathol.* 2022;35(1):23–32.
- Wang X, Zhao J, Marostica E, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature.* 2024;634(8035):970–8.
- Li X, Eastham J, Giltner JM, et al. Automated tumour immunophenotyping predicts clinical benefit from anti-PD-L1 immunotherapy. *J Pathol.* 2024;263(2):190–202.
- Yolchuyeva S, Giacomazzi E, Tonneau M, et al. Radiomics approaches to predict PD-L1 and PFS in advanced non-small cell lung patients treated with immunotherapy: a multi-institutional study. *Sci Rep.* 2023;13(1):11065.
- Ben Dori S, Aizic A, Sabo E, et al. Spatial heterogeneity of PD-L1 expression and the risk for misclassification of PD-L1 immunohistochemistry in non-small cell lung cancer. *Lung Cancer.* 2020;147:91–8.
- Tsao MS, Nicholson AG, Maleszewski JJ, et al. Introduction to 2021 WHO classification of thoracic tumours. *J Thorac Oncol.* 2022;17(1):e1–4.
- Prelaj A, Tay R, Ferrara R, et al. Predictive biomarkers of response for immune checkpoint inhibitors in non-small-cell lung cancer. *Eur J Cancer.* 2019;106:144–59.
- Vanguri RS, Luo J, Aukerman AT, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat Cancer.* 2022;3(10):1151–64.
- Huang Z, Bianchi F, Yuksekogonul M, et al. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med.* 2023;29(9):2307–16.
- Tong L, Sha Y, Wang MD. Improving classification of breast cancer by utilizing the image pyramids of whole-slide imaging and multi-scale convolutional neural networks. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) 2019; Vol 1: 696–703.
- Liu S, Zhu C, Xu F. BCI: Breast cancer immunohistochemical image generation through pyramid Pix2pix. In: the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 2022:1815–1824.
- Pocock J, Graham S, Vu QD, et al. Tiatoolbox as an end-to-end library for advanced tissue image analytics. *Commun Med.* 2022;2(1):1–14.
- Janowczyk A, Zuo R, Gilmore H, et al. HistoQC: An open-source quality control tool for digital pathology slides. *JCO Clin Cancer Inform.* 2019;3:1–7.
- Vaswani A, Shazeer N, Parmar N. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems* 2017:6000–6010.
- Chen RJ, Chen C, Li Y. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE; 2022:16123–16134.
- Jin D, Liang S, Shmatko A. Teacher-student collaborated multiple instance learning for pan-cancer PDL1 expression prediction from histopathology slides. *Nat Commun* 2024;15(1):3063. Published 2024 Apr 9.
- Shamai G, Livne A, Polónia A. Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. *Nat Commun* 2022;13(1):6753. Published 2022 Nov 8.
- Sha L, Osinski BL, Ho IY, et al. Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. *J Pathol Inform.* 2019;10:24. [https://doi.org/10.4103/jpijpi\\_24\\_19](https://doi.org/10.4103/jpijpi_24_19).

43. Kapil A, Meier A, Steele K, et al. Domain adaptation-based deep learning for automated tumour cell (TC) scoring and survival analysis on PD-L1 stained tissue images. *IEEE Trans Med Imaging*. 2021;40(9):2513–23.
44. Cao L, Wang J, Zhang Y, et al. E2EFP-MIL: end-to-end and high-generalizability weakly supervised deep convolutional network for lung cancer classification from whole slide image. *Med Image Anal*. 2023;88:102837.
45. Shao Z, Bian H, Chan Y. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Proc. Advances in Neural Information Processing Systems*. 2021;34:2136–2147.
46. Zhang H, Meng Y, Zhao Y. DTFT-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: *Proc. IEEE Conf Comput Vis Pattern Recognit 2022*;18802–18812.

## Publisher's Note

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