Diagnostic Accuracy Study

Medicine®

OPEN

Dual-scale categorization based deep learning to evaluate programmed cell death ligand 1 expression in non-small cell lung cancer

Xiangyun Wang, MD[d], Peilin Chen, PhD[a], Guangtai Ding, PhD[b], Yishi Xing, MS[a], Rongrong Tang, MS[a], Chaolong Peng, MD[c], Yizhou Ye, PhD[a], Qiang Fu, MD[e,*]

Abstract

In precision oncology, immune check point blockade therapy has quickly emerged as novel strategy by its efficacy, where programmed death ligand 1 (PD-L1) expression is used as a clinically validated predictive biomarker of response for the therapy. Automating pathological image analysis and accelerating pathology evaluation is becoming an unmet need. Artificial Intelligence and deep learning tools in digital pathology have been studied in order to evaluate PD-L1 expression in PD-L1 immunohistochemistry image. We proposed a Dual-scale Categorization (DSC)-based deep learning method that employed 2 VGG16 neural networks, 1 network for 1 scale, to critically evaluate PD-L1 expression. The DSC-based deep learning method was tested in a cohort of 110 patients diagnosed as non-small cell lung cancer. This method showed a concordance of 88% with pathologist, which was higher than concordance of 83% of 1-scale categorization-based method. Our results show that the DSCbased method can empower the deep learning application in digital pathology and facilitate computer-aided diagnosis.

Abbreviations: IHC = immunohistochemistry, NSCLC = non-small cell lung cancer, PD-L1 = programmed cell death ligand 1, ROC = receiver-operating characteristic, TPS = Tumor Proportion Score, WSI = whole-slide images.

Keywords: deep learning, Dual-scale Categorization, patch classification, PD-L1 expression

1. Introduction

The application of classical machine learning methods to solve complex tasks in the field of medical images analysis has shown the ability of automatically learning pathologic classification features.[1–3] Compared to the classical machine learning methods, which rely heavily on the selection of hand-crafted features, deep learning methods are more effective in resolving more complicated tasks such as image recognition/classification.[4–7] Although lacking of high-quality annotation dataset and the uninterpretability of image features related to the concluded prediction limit the application of deep learning methods in clinical practice, deep learning methods have still been widely used in biomarker analytics such as ki67-index estimation and HER2 status estimation,[8–10] and cancer stage diagnosis especially in grading, classification, and metastasis detection.[11,12] From the training image sets, deep learning methods learn and discover the pathogenic features, among which many are not visually apparent or readable to pathologists. However those features may better represent the disease status than the other known features.[13,14]

Programmed cell death ligand 1 (PD-L1) is a major immune checkpoint biomarker for immunotherapies,[15] and higher PD-L1 expression on tumor cells has been associated with greater efficacy.[16–19] Therefore PD-L1 expression status was approved as companion diagnosis by FDA (eg, 22C3 for Pembrolizumab in non-small cell lung cancer [NSCLC]) along the drug approval.[20] Besides the approved companion diagnosis, multiple PD-L1 immunohistochemistry (IHC) assays (including 22C3, 28-8, SP263, and SP142) have been developed to effectively detect PD-L1 expression status in different cancer types. Given the nature of IHC assays, different system yields different evaluation scoring grades.[21] In clinical practice, calculation of Tumor Proportion Score (TPS) usually requires an experienced pathologist to count the PD-L1-positive tumor cells and PD-L1-negative tumor cells under a microscope, while a tumor tissue section often contains a population of many thousands of cells. The calculated result will provide valuable information for an oncology doctor to determine whether the patient will potentially benefit from immunotherapy. Moreover, factors regarding to intra-tumor,

Editor: Meliha Ekinci.

The authors report no conflicts of interest.

The datasets generated during and/or analyzed during the present study are available from the corresponding author on reasonable request.

[a] Department of Data Systems, 3D Medicines Inc., [b] School of Computer Engineering and Science, Shanghai University, [c] Department of Pathology, 3D Medicines Inc., [d] Department of Respiratory and Critical Care Medicine Changzheng Hospital, Naval Military Medical University, [e] Oncology Department, Changhai Hospital of Shanghai, Shanghai, China.

[*] Correspondence: Qiang Fu, Oncology Department, Changhai Hospital of Shanghai, No 168 Changhai Rd, Yangpu District, Shanghai 200433, China (e-mail: qiang.fu@vip.163.com).

How to cite this article: Wang X, Chen P, Ding G, Xing Y, Tang R, Peng C, Ye Y, Fu Q. Dual-scale categorization based deep learning to evaluate programmed cell death ligand 1 expression in non-small cell lung cancer. Medicine 2021;100:20(e25994).

Received: 19 October 2020 / Received in final form: 13 April 2021 / Accepted: 29 April 2021

http://dx.doi.org/10.1097/MD.0000000000025994

inter-observer, and inter-assay heterogeneities, often led to challenges for a pathologist to evaluate an accurate TPS score.[21]

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer death.[22] NSCLC is the predominant histological type of lung cancer, and comprises approximately 80% to 85% of all lung cancers.[23] There are a few pilot studies about leveraging machine learning methods to evaluate PD-L1 expression status in NSCLC. Taylor et al presented an automated image analysis approach, which was based on a feedback machine learning method, to evaluate PD-L1 expression on both tumor and immune cells in NSCLC. The results of the automated method showed good concordance with the pathologists' scores.[24] However, the approach largely depended on hand-crafted features, such as statistics of cell shape and nuclear texture; those features were essential to accurately reflect cell morphology, but were often difficult to adapt to new data sets.[25] Kapil et al also tried a pixel-based deep learning method to predict the PD-L1 expression status in NSCLC. They used the auxiliary classifier generative adversarial networks to generate large amount of fake images to construct the prediction model, and the model's accuracy was concordant to some extents with visual scoring by pathologists. Their study indicated that the proposed method using the area instead of cell count to calculate TPS score was a feasible solution.[26]

Instead of counting cell number or calculating cell area on pixel-level, in this article we tried to use a patch-based method to evaluate the PD-L1 expression level in NSCLC stained by 22C3 clone. We constructed a novel deep learning framework, named as Dual-scale Categorization-based VGG16 (DSC-VGG16), which employed 2 VGG16 neural networks, one network for one scale, in order to obtain a more accurate TPS score. Using 1% or 50% as the TPS cutoff points, our results showed that the TPS scores of DSC-VGG16 model were highly consistent with that of pathologists, and DSC-VGG16 model performed better than the VGG16 network with one-scale categorization based method.

2. Materials and methods

2.1. Specimens

From January 2018 to January 2020, a Total of 300 NSCLC samples were collected in Changhai hospital and Changzheng hospital (Shanghai, China). All the samples were processed in the 3DMed Clinical Laboratory, which is accredited by College of American Pathologists and certified by Clinical Laboratory Improvement Amendments. PD-L1 IHC was performed with the PD-L1 IHC 22C3 pharmDx kit on the Dako Autostainer Link 48 platform according to the kit's manufacturer recommendations. All immune-stained slides were digitally scanned into whole-slide images (WSI) under a resolution of 0.5 μm/pixel (20× magnifications) using PRECICE 500 (manufactured by UNIC Tech Company, Beijing). No personalized health information was obtained and required, therefore only information of samples' tumor type and their corresponding PD-L1 expression data were kept in this study. This study was approved by the Changzheng Hospital Ethics Committee (2018-021-01).

2.2. Training, validation, and test sets

A total of 110 samples (61.82%, TPS <1%; 27.27%, TPS: 1%–49%; and 10.91%, TPS ≥50%) were selected for testing; and 190 samples were selected for training and validation. Four categorized regions, PD-L1-positive tumor cells (TP) regions, PD-L1-negative tumor cells (TN) regions, PD-L1-positive immune cells (IP, including macrophage and lymphocyte cells) regions, and the other (OT) regions, were manually labeled on WSIs of the training and validation samples by 2 in-house pathologists.

In the training and testing process, the annotated regions were divided into patches of 128 × 128 pixels size and also classified into 4 categories: TP, IP, TN, and OT, which were called the macro scale categories. There were 6800 patches in each category. Some of the TP patches, especially those obtained from samples with weak PD-L1 expression, were composed of both PD-L1-positive tumor cells and PD-L1 negative tumor cells. Furthermore, 1500 TP patches were further classified into three so called "micro categories": TP1 represents the patches that contain the maximum counts of PD-L1-positive tumor cell, TP2 represents 50% PD-L1-positive tumor cell of TP1, and TP3 represents 25% PD-L1 positive tumor cell of TP1 (Table 1).

2.3. Modeling

Our model was deployed on deep learning framework Keras version 2.2.4, on top of Tensorflow version 1.13.1,[27] and based on VGG16 model (from Keras) composed of 4 different types of layers: convolutional layers, max-pooling layers, fully connected layers, a soft-max layer. The VGG16 model was initialized with ImageNet pretrained weights,[28] and the last 7 layers were unfrozen for training. Flatten layer and dense layer were specifically added after convolutional layers. Rectified linear unit (ReLU) was chosen as our activation function in dense layers and dropout layer was added to avoid over fitting. Stochastic gradient descent was used as the optimizer, and the learning rate was 0.01. Input patch size was set as 128 × 128 pixels. Batch with 256 labeled patches was chosen. One Nvidia RTX 2080Ti was used to train "DSC-marco" (for classifying 4 macro categories) and "DSC-micro" model (for classifying 3 microcategories) (Fig. 1).

Table 1

List of training, validation, and testing sets for modeling in patch category.

Model	Patch category	Training	Validation	Testing	Total no.
DSC-macro	TP	5400	1000	400	6800
	TN	5400	1000	400	6800
	IP	5400	1000	400	6800
	OT	5400	1000	400	6800
DSC-micro	TP1	400	50	50	500
	TP2	400	50	50	500
	TP3	400	50	50	500

DSC = Dual-scale Categorization, IP = PD-L1 positive immune cells patch, OT = other regions patch, TN = PD-L1 negative tumor cells patch, TP = PD-L1 positive tumor cells patch, TP1: patches that contain the maximum counts of PD-L1 positive tumor cell, TP2 = patch that represent 50% PD-L1 positive tumor cell of TP1, TP3 = patch that represent 25% PD-L1 positive tumor cell of TP1.

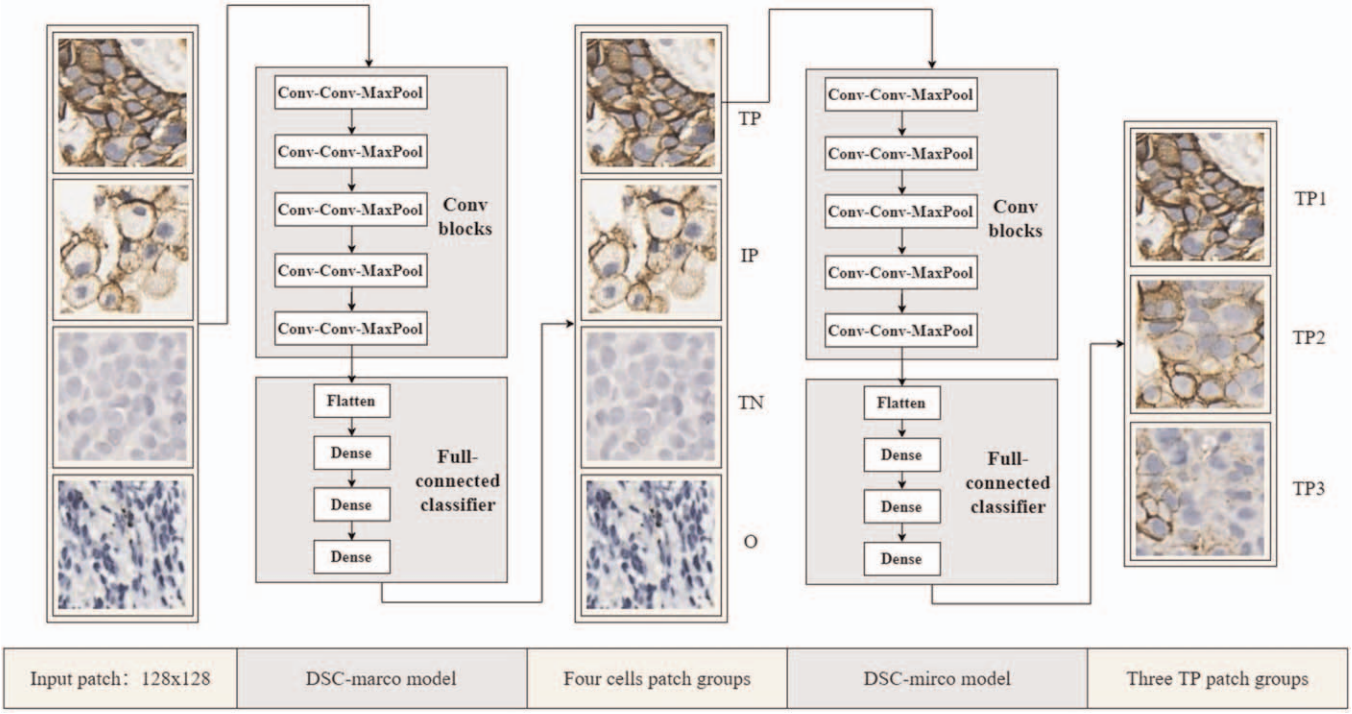

Figure 1. Proposed dual-scale categorization method based on VGG16 architecture presented in this study. The input patch size is 128 × 128 pixels. DSC-marco model was trained and used for classification of the 4 cells patch groups. DSC-mirco model was trained and used for classification of the three TP patch groups. The final classification of TN, TP1, TP2, TP3 were used for TPS calculation. (TP1 represent the patches contain the maximum counts of PD-L1 positive tumor cell, TP2 represent 50% PD-L1 positive tumor cell of TP1 and TP3 represent 25% PD-L1 positive tumor cell of TP1.). DSC = Dual-scale Categorization.

2.4. TPS evaluation

PD-L1 expression status is assessed based on tumor proportion score (TPS by PD-L1 IHC. Manual evaluation is performed by 2 certified pathologists, respectively.

TPS is scored as the percentage of viable tumor cells presenting partial or complete PD-L1 expression at the cell membrane relative to all viable tumor cells:

$$\mathrm{TPS} = \frac{\mathrm{PD-L1\ staining\ tumor\ cells}}{\mathrm{Totaltumor\ cells}} \tag{1}$$

In this article, $\mathrm{TPS}_{\mathrm{VGG16}}$ was defined as the percentage of PD-L1-positive tumor cells patches among all tumor cells patches. The numbers of TP patches and TN patches were evaluated by DSC-marco model, and the numbers of TP1 patches, TP2 patches, and TP3 patches were evaluated by DSC-micro model. Hence, $\mathrm{TPS}_{\mathrm{VGG16}}$ was defined as in the following:

$$\mathrm{TPS}_{\mathrm{VGG16}} \frac{\mathrm{TP1} + \mathrm{TP2}*50\% + \mathrm{TP3}*25\%}{\mathrm{TotalTP} + \mathrm{TN}} \tag{2}$$

2.5. Statistical analysis

Standard statistical testes were used to analyze the data, including Concordance Correlation Coefficient and Cohen κ. All statistical tests were two-sided, and statistical significance was considered where *P* values <.05.

The primary outcome of the deep learning system is the patch class prediction, whereas the sensitivity and specificity of our receiver-operating characteristic (ROC) targeted were based on the whole slide prediction. This complicated the ROC analysis because the sensitivity/specificity of patches is not monotonically correlated to that of their residing whole slide, and more importantly we did not have a "criterion standard" of the patch classification from pathologists. To still perform the analysis, we used the predicted TPS as the input of the analysis. Youden index was calculated to find the optimal cutoff points with the higher sensitivity and specificity.

3. Results

3.1. Patch classification by DSC-macro and DSC-micro models

In DSC-macro model, patches in a WSI were classified into 4 patch groups: TN patch groups, TP patch groups, IP patch groups, and OT patch groups (Fig. 2). In DSC-micro model, TP patches were further classified into three subgroups. Two sets of testing patches were used to evaluate the performance of DSC-macro and DSC-micro models on patch classification. DSC-macro model performed well on classifying the four patch groups (Table 2), with the F1 scores for each classification >95%. In the testing sets, 7.5% PD-L1-positive tumor cell patches were wrongly classified as positive immune cell patches, and <1% positive immune cell patches were wrongly classified as positive tumor cell patches. Due to the relatively smaller size of training dataset, DSC-micro model could not perform as well as DSC-macro in classifying the sub-groups of TP patches (Table 2), with F1 scores for each class ranging from 59.18% to 84.21%.

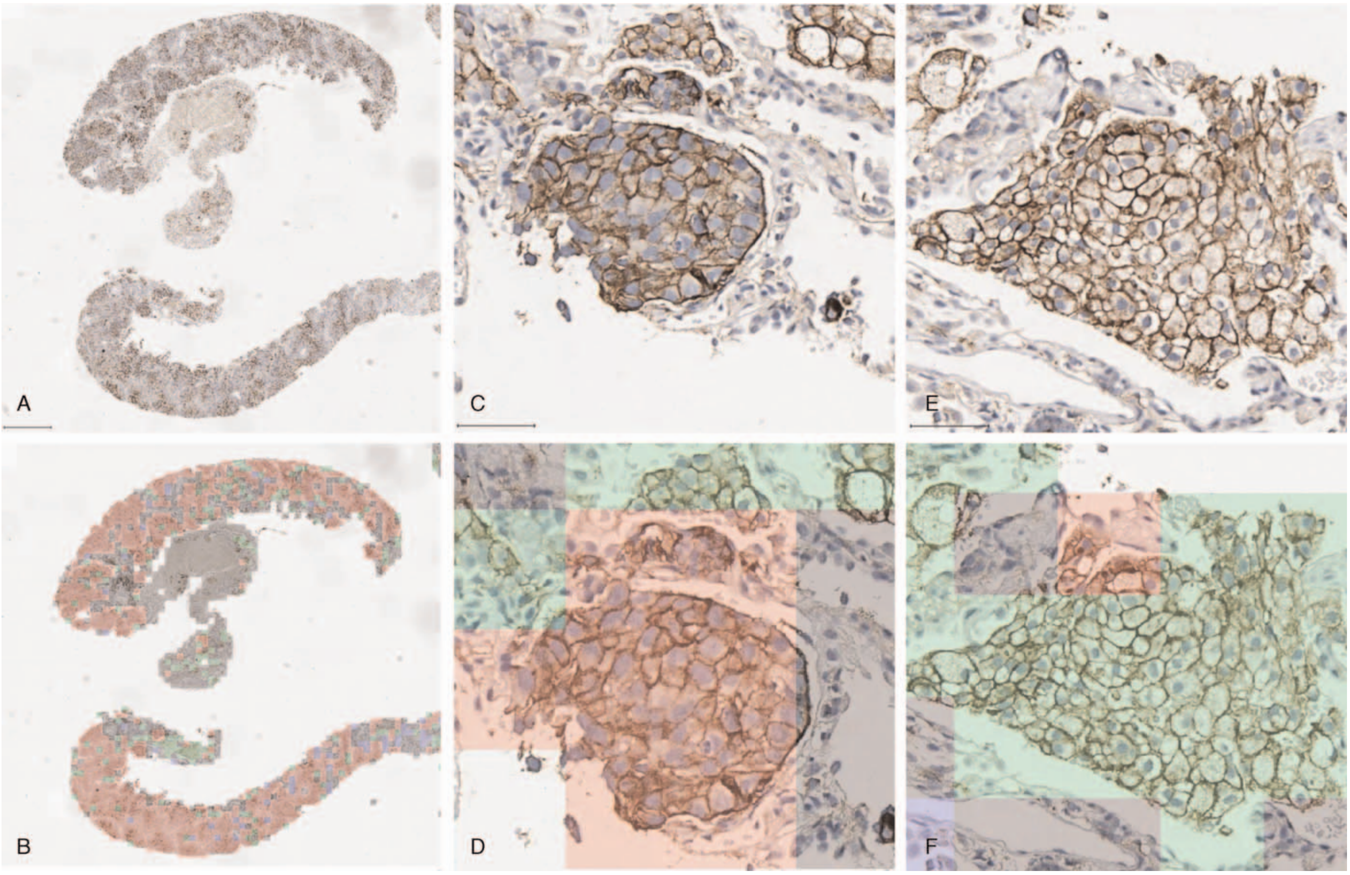

Figure 2. Example of classification result by Dual-scale Categorization-macro model. (A) Original programmed death ligand 1 (PD-L1) immunohistochemistry (IHC) images Scale bar: 0.5 mm. (B) Visualization of patch classification result, PD-L1 positive tumor cell patches are presented through red channel, PD-L1 positive immune cell patches are presented through green channel, PD-L1 negative tumor cell patches are presented through blue channel. (C) Original PD-L1 IHC images corresponding to the red box in 1A, Scale bar: 0.05 mm. (D) Visualization of the predicted PD-L1 positive tumor cell regions corresponding to the red box in 1A. (E) Original PD-L1 IHC images corresponding to the green box in 1A, Scale bar: 0.05 mm. (F) Visualization of the predicted PD-L1 positive immune cell regions corresponding to the green box in 1A.

3.2. TPS prediction by DSC-VGG16 and VGG16-macro model under different cutoff values

DSC-VGG16 model was tested on 110 independent NSCLC samples. In clinical practice, PD-L1 expression level in tumor cells is classified by TPS into 3 groups: TPS <1% (negative expression), TPS: 1%–49% (weakly positive expression), and ≥50% (highly positive expression).[29] We used 2 different cutoff points to evaluate the expression level of PD-L1 in NSCLC, and compared DSC-VGG16 model-based TPS score with pathologist-based TPS scores. Under the 1% cutoff point, the F1 score of DSC-VGG16 and VGG16-macro model for predicting positive expression groups were 90.24% and 87.23%, respectively. Under the 50% cutoff point, the F1 score of DSC-VGG16 and VGG16-macro model for predicting highly positive expression groups reached 81.82% and 73.33%, respectively. Under all different cutoff points, DSC-VGG16 model had higher specificity, but lower sensitivity, compared with those of VGG16-macro model (Table 3).

Table 2

Patch categorization with DSC-macro model in four cell patches groups and patch categorization with DSC-micro model in 3 TP patches groups.

Model	Patch category	Sensitivity	Specificity	F1 score
DSC-macro	TN	98.00%	99.17%	97.76%
	TP	91.50%	99.67%	95.06%
	IP	99.00%	96.92%	95.08%
	OT	96.00%	99.08%	96.60%
DSC-micro	TP1	80.00%	95.00%	84.21%
	TP2	80.00%	83.00%	74.77%
	TP3	58.00%	81.00%	59.18%

DSC = Dual-scale Categorization.

3.3. Comparison of dual scale model DSC-VGG16 and single macro scale model DSC-macro

The prevalence of PD-L1 expression level for DSC-VGG16 model and DSC-macro model were, respectively, 63.64% and 52.73% for negative expression, 27.27% and 30.91% for weekly positive expression, and 9.09% and 16.36% for highly positive expression (Table 4). The concordance between the DSC-VGG16 model and pathologists reached a higher agreement than the concordance between the DSC-macro model and pathologists (Cohen κ: 0.79 vs 0.68; Concordance Correlation Coefficient: 0.88 vs 0.83, Table 5).

TPS scores of 12 samples were inconsistent between DSC-VGG16 model and the pathologists (Table 4), and the discordant

Table 3

The overall sensitivity and specificity and F1 score using DSC-VGG16 and DSC-macro model for TPS prediction under different TPS cutoff points.

Cutoff points	Model	Sensitivity	Specificity	F1 score
1%	DSC-VGG16	88.10%	95.59%	90.24%
	VGG16-macro	97.62%	83.82%	87.23%
50%	DSC-VGG16	75.00%	98.98%	81.82%
	VGG16-macro	91.67%	92.86%	73.33%

DSC-VGG16 = Dual-scale Categorization-based VGG16, TPS = Tumor Proportion Score.

number of samples grew to 20 in the comparison between the scores of DSC-macro and pathologists. Of 68 negative (TPS < 1%) samples determined by pathologists, 11 was predicted as positive (TPS <1%) by DSC-macro model, and the number decreased to 3 by DSC-VGG model. Meanwhile, among the 30 pathologist-confirmed weak positive (TPS: 1%–49%) samples, the misclassification number was 8 by DSC-macro model and 6 by DSC-VGG model. However, the 8 misclassified samples by DSC-macro model consisted of 7 that fell into the category of strong positive (TPS > 50%), whereas only 1 of 6 was falsely predicted as strong positive (TPS > 50%) by DSC-VGG model. The negative-preferring nature of DSC-VGG model was also clear in the 12 samples that our pathologists believed to be strong positive (TPS > 50%), as 3 samples were predicted to be weak positive (TPS: 1%–49%) which was only 1 for DSC-macro model.

Overall, DSC-VGG16 model would generate a more accurate TPS score than DSC-macro model. However, the proposed DSC-VGG16 model was prone to predict a lower TPS scores than the pathologists' scores for a small percent of samples, which might be attributable to the low performance of DSC-micro model on subgroup patches classification.

3.4. Finding the optimal cutoff points of DSC-VGG166 model to predict TPS

ROC curve was used to find the optimal cutoff points of DSC-VGG166 model to predict TPS with the higher sensitivity and specificity. Under 1% cutoff point, the area under curve for distinguishing between positive samples versus negative samples was 0.97 (Fig. 3), and we found 0.7% was the optimal cutoff point of DSC-VGG166 model with sensitivity and specificity were 0.976 and 0.926, respectively. Under 50% cutoff point, the area under curve for distinguishing between strong positive samples versus negative and week positive samples was 0.99 (Figure 3), and we found 24.8% was the optimal cutoff point of DSC-VGG166 model with sensitivity and specificity were 1.0 and 0.939, respectively.

Table 4

The distribution of TPSs calculation by DSC-VGG16, DSC-macro model and pathologists at different TPS range.

		Pathologist-based TPS scores		
		<1%	1%–49%	≥50%
DSC-VGG16 based TPS scores	<1%	65	5	0
	1%–49%	3	24	3
	≥50%	0	1	9
DSC-macro-based TPS scores	<1%	57	1	0
	1%–49%	11	22	1
	≥50%	0	7	11

DSC-VGG16 = Dual-scale Categorization-based VGG16, TPS = Tumor Proportion Score.

Table 5

The concordance analysis between deep learning method and pathologists.

	Pathologist	
	Cohen κ	LCC
DSC-VGG16	0.79 (95 CI: 0.68–0.90)	0.88 (95 CI: 0.83–0.92)
DSC-macro	0.68 (95 CI: 0.56–0.80)	0.83 (95 CI: 0.76–0.88)

CI = confidence interval, DSC-VGG16 = Dual-scale Categorization-based VGG16, LCC = Concordance Correlation Coefficient.

4. Discussion

PD-L1 has gained widespread acceptance as a predictive biomarker in NSCLC to identify potential treatment responders.[29,30] In clinical practice, it is important to accurately evaluate the PD-L1 expression level. However, it is often impossible to manually count tumor cells used for calculating the TPS due to the huge number of cells in a tissue. In this condition, the pathologist will prefer to select the representative region and semi-quantitatively evaluate the TPS value by evaluating the target area. However, the process is tedious, time consuming and also subjective.[31,32] Moreover, it is known that tumor tissue is very complex and heterogeneous, and PD-L1 may also be expressed by lymphocytes, macrophages, or dendritic cells.[33] Such factors may lead to the quality and reliability being affected due to interobserver variation between pathologists.

Classical machine learning methods have been developed to overcome these challenges. These methods try to mimic experts' cognition of cells using highly generalized hand crafted features, and calculate the TPS by counting the tumor cell.[24,34] However, numerous statistical parameters need to be adjusted to accurately capture the cell morphology, and these parameters are difficult to adapt to new data sets, making these methods difficult to be automated.[35] A Rule-based methods combined with machine learning was propose to automated to detect CD8+ and PD-L1+ cells in single marker images, and demonstrated that a combined signatures of PD-L1+ tumor cells and CD8+ tumor infiltrating lymphocytes may allow better identification of responders to durvalumab monotherapy compared with manual PD-L1 scoring alone.[36] Deep learning methods have been developed to resolve this problem. Kapil et al established a deep learning method that could detect target cells at pixel level, and they took the ratio between the pixel counts of the detected positive tumor cell regions to the pixel count of all detected tumor cell regions as TPS score. Their results showed promise of deep learning method in solving this problem, and indicated that using the area instead of cell count to calculate TPS score was feasible.[26] Many biomarker evaluations need cell count to be quantitatively evaluated, it is reasonable to use patch numbers to represent the cells it contains. More importantly, a patch may have more features information than that of a single cell contains. Pitkäaho et al proposed a patched-based deep learning model to calculate her2 expression scores. The model calculated the her2-positive tumor cell patches instead of tumor cell counts to evaluate the her2 expression status, and the accuracy of the model in patches classification reached 97.7%.[37] The results in this study also indicated the patched-based deep learning method had the potential to resolve the cell counting issue.

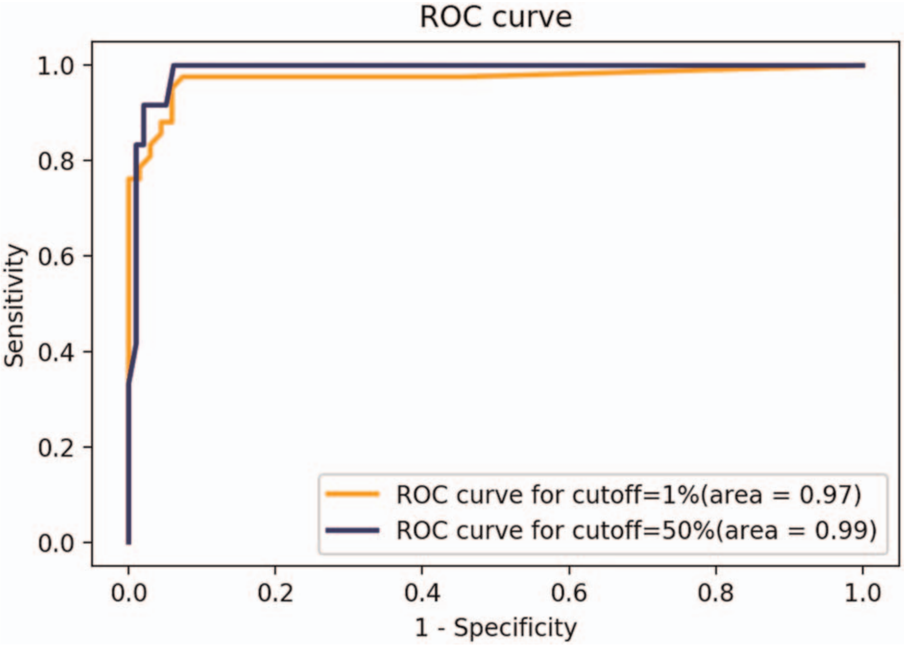

Figure 3. ROC curve and area under the curve for Tumor Proportion Score prediction under 1% and 50% cutoff points of Dual-scale Categorization-based VGG16 model. ROC = recesiver-operating characteristic.

PD-L1 can be expressed in the membrane or cytoplasm of normal tissue cells, including various immune cells, and necrotic cells, which should be excluded from evaluation in the 22C3 assay.[38] PD-L1 IHC false positivity will be brought in generally due PD-L1 positive immune cells and histiocytes lie between PD-L1 negative tumor cells which may be misinterpreted as positive.[39] In our established patch-based model, positive tumor cell patches were easily predicted as positive immune cell patches by the DSC-macro model. The deep learning method proposed by Kapil et al also showed that PD-L1-positive macrophage regions and PD-L1-positive tumor cell regions were easily to be mistaken as each other. Organizational structure was an important feature for cell of classification,[35] and this information could be obtained from a patch contains dozens of cells or more, but not from a single cell. We speculated that a patch with a size large than 128 × 128 pixels might provide more information to discriminate these PD-L1-positive tumors cells from other PD-L1-positive cells. Input size with 256x256 pixels is worth trying in future study, and a cell counting method may need to be incorporated into the patch-based model to improve the accuracy.

Considering the heterogeneity of PD-L1 expression in tumor tissue, the positive and negative of the tumor cells may coexist in the same patch, especially those patches obtained from samples with weak PD-L1 expression. In this situation, among all the tumor cells patches (TP and TN), treating all the TP patches as the same level of PD-L1 expression unit may introduce bias in the calculation of TPS. To resolve this problem, we employed a second classification model to further classify the PD-L1 positive patches to get a finer result. It turned out this extra step made our model more accurate. In this study, we divided the TP patches by PD-L1-positive tumor cell count into 3 categories, we believed that the accuracy of the DSC-VGG16 model could be further improved if we increase the categories of the TP subgroups and the materials used for training DSC-micro model. We found 0.7% was the optimal cutoff point of DSC-VGG166 model for distinguishing between positive samples versus negative samples. This remind us that in our model, predicted TPS with 0.7% to 1% need carefully checked as 5-week positive samples were fell into this interval. In view of the critically unbalanced classes (Table 4), more positive samples are needed for further ROC curve analysis under 50% cutoff point.

Before these patched-based deep learning methods are applied into clinical practice to evaluate quantitative index, there are 2 points we suggest to be considered. First, patch size needs to be selected carefully according to the specific problem we need to solve. As shown in our results, patch with 128 × 128 pixels may be suitable in the evaluation of PD-L1 expression status. However, patches with 128 × 128 pixels may loss some tissue structure features, and these features may be useful for discriminating the subtle differences between the normal tissue and highly differentiated tumor region or between the PD-L1-positive immune cell and PD-L1-positive cancer cell. Moreover, it is inevitable that more cells will be cut in half in small patches, and a patch containing these kind of cells is difficult to classify. Although a larger size is worth trying, it should be cautioned that a patch with large size will contain more than one cell type which may cause trouble for classification. Second, when the tissue is too small, it must be aware that using patch to present the target cell may cause large calculation deviations. In the situation of evaluation of a PD-L1-negative expression sample with small tissue size (such as needle biopsies), false-positive sample may occur if a PD-L1-negative tumor cell patch is predicted as a positive tumor cell patch. We have tried to produce more patches from images by allowing patch overlap but had little effect on the final results.

In summary, this study provides evidence that the patch-based dual-scale categorization method is cost-effective and accurate in evaluation of PD-L1 expression status and has great clinical application value. However, we need more data to demonstrate the practicality of our proposed method, especially considering a wide range of commercially available PD-L1 IHC kits which utilize different antibodies, different manufacturers, and different cutoff scores to detect or quantify tumor PD-L1 expression.

Author contributions

Conceptualization: Peilin Chen, Yizhou Ye.
Data curation: Xiangyun Wang, Chaolong Peng.
Formal analysis: Yishi Xing, Rongrong Tang.
Methodology: Peilin Chen, Guangtai Ding, Yizhou Ye.
Supervision: Xiangyun Wang, Guangtai Ding, Yizhou Ye, Qiang Fu.
Validation: Xiangyun Wang, Qiang Fu.
Writing – original draft: Peilin Chen.
Writing – review & editing: Yizhou Ye, Qiang Fu.

References

[1] Anwar SM, Majid M, Qayyum A, et al. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018;42:226.

[2] Carin L, Pencina MJ. On deep learning for medical image analysis. JAMA 2018;320:1192–3.

[3] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. J Med Syst 2017;42:60–88.

[4] Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15:

[5] Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67.

[6] Komura D, Ishikawa S. Machine learning approaches for pathologic diagnosis. Virchows Arch 2019;475:131–8.

[7] Yu KH, Zhang C, Berry GJ, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7:12474.

[8] Saha M, Chakraborty C, Arun I, et al. An advanced deep learning approach for Ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. Sci Rep 2017;7:3213.

[9] Saha M, Chakraborty C, Racoceanu D. Efficient deep learning model for mitosis detection using breast histopathology images. Comput Med Imaging Graph 2018;64:29–40.

[10] Vandenberghe ME, Scott ML, Scorer PW, et al. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Sci Rep 2017;7:45938.

[11] Goncalves WGE, Dos Santos MHP, Lobato FMF, et al. Deep learning in gastric tissue diseases: a systematic review. BMJ Open Gastroenterol 2020;7:e000371.

[12] Wang S, Yang DM, Rong R, et al. Artificial intelligence in lung cancer pathology image analysis. Cancers 2019;11.

[13] Couture HD, Williams LA, Geradts J, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 2018;4:30.

[14] Litjens G, Sanchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.

[15] Martinez P, Peters S, Stammers T, et al. Immunotherapy for the first-line treatment of patients with metastatic non-small cell lung cancer. Clin Cancer Res 2019;25:2691–8.

[16] Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015;373:1627–39.

[17] Garon EB, Rizvi NA, Hui R, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med 2015;372:2018–28.

[18] Herbst RS, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet 2016;387: 1540–50.

[19] Reck M, Rodriguez-Abreu D, Robinson AG, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. N Engl J Med 2016;375:1823–33.

[20] Vaddepally RK, Kharel P, Pandey R, et al. Review of indications of FDA-approved immune checkpoint inhibitors per NCCN guidelines with the level of evidence. Cancers 2020;12:

[21] Ancevski Hunter K, Socinski MA, Villaruz LC. PD-L1 testing in guiding patient selection for PD-1/PD-L1 inhibitor therapy in lung cancer. Mol Diagn Ther 2018;22:1–10.

[22] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.

[23] Osmani L, Askin F, Gabrielson E, et al. Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy. Semin Cancer Biol 2018;52:103–9.

[24] Taylor CR, Jadhav AP, Gholap A, 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 2019;27:263–9.

[25] Beck AH, Sangoi AR, Leung S, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 2011;3:108ra13.

[26] Kapil A, Meier A, Zuraw A, et al. Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies. Sci Rep 2018;8:17343.

[27] Chollet F. Keras. *GitHub* 2015; available at: https://github.com/fchollet/keras.

[28] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In CVPR 2009;248–55.

[29] Incorvaia L, Fanale D, Badalamenti G, et al. Programmed death ligand 1 (PD-L1) as a predictive biomarker for pembrolizumab therapy in patients with advanced non-small-cell lung cancer (NSCLC). Adv Ther 2019;36:2600–17.

[30] Arora S, Velichinskii R, Lesh RW, et al. Existing and emerging biomarkers for immune checkpoint immunotherapy in solid tumors. Adv Ther 2019;36:2638–78.

[31] Smits AJ, Kummer JA, de Bruin PC, et al. The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. Mod Pathol 2014;27:168–74.

[32] McLaughlin J, Han G, Schalper KA, et al. Quantitative assessment of the heterogeneity of PD-L1 expression in non-small-cell lung cancer. JAMA Oncol 2016;2:46–54.

[33] Loke P, Allison JP. PD-L1 and PD-L2 are differentially regulated by Th1 and Th2 cells. Proc Natl Acad Sci U S A 2003;100:5336–41.

[34] Koelzer VH, Gisler A, Hanhart JC, et al. Digital image analysis improves precision of PD-L1 scoring in cutaneous melanoma. Histopathology 2018;73:397–406.

[35] Rodenacker K, Bengtsson E. A feature set for cytometry on digitized microscopic images. Anal Cell Pathol 2003;25:1–36.

[36] Althammer S, Tan TH, Andreas Spitzmüller, et al. Automated image analysis of NSCLC biopsies to predict response to anti-PD-L1 therapy. J Immunother Cancer 2019;7:121.

[37] Pitkäaho T, Lehtimäki TM, McDonald J, et al. Classifying HER2 breast cancer cell samples using deep learning. Irish Machine Vision and Image Processing Conference: Irish Pattern Recognition and Classification Society. In Proc Irish Mach Vis Image Process Conf 2016;1–104.

[38] Verocq C, Decaestecker C, Rocq L, et al. The daily practice reality of PD-L1 (CD274) evaluation in non-small cell lung cancer: a retrospective study. Oncol Lett 2020;19:3400–10.

[39] Thunnissen E, Allen TC, Adam J, et al. Immunohistochemistry of pulmonary biomarkers: a perspective from members of the Pulmonary Pathology Society. Arch Pathol Lab Med 2018;142:408–41.