Peer Review File

Article information: http://dx.doi.org/10.21037/atm-19-4690

Responses to Reviewer #1

In this manuscript the Authors analyze the role of morphological CT scan parameteres and machine learning algorithm of artificial intelligence (Three-dimensional densely connected convolutional networks – 3D DenseNet) in the prediction of PD-L1 expression in 127 patients with stage III and IV lung adenocarcinoma. A correlation between CT characteristics and EGFR mutation was also assessed. The results of the study show that the 3D DenseNet algorithm had a higher predictive value of PD-L1 expression than CT scan morphological parameters. Despite some limitations as the relatively low number of patients for this type of trial and the fact that most histological specimens were obtained with FNA and not on the whole tumor, the topic is certainly of interest. The manuscript has however to be improved:

We thank the reviewer for the kind suggestions and valuable comments.

Comment. (1) In line 210, the number of patients distributed according to PD-L1 expression has to be corrected. In fact, since only 46 patients were PD-L1 positive, it is not clear how 89 of them could be weakly positive for PD-L1.

Response : In fact, in our article, cases were divided into positive and negative by PD-L1 TPS of 1%, weakly positive and strongly positive by 50%. So the weakly positive cohort included patients with PD-L1 TPS of 5%-50% and PD-L1 negative cases, so there were totally 89 cases in weakly positive group. We give the definition in our article (see line 102-105). Maybe the definition was not so appropriate, in order not to confuse the readers, we now directly use PD-L1-TPS to describe the group, so that the expression will be clearer.

Changes in the text: Line 12, 13, 15, 199-202, 210,212, 234

Comment. (2) The results of the study show that 3D DenseNet assessment had a better correlation with PD-L1 expression than CT scan morphological characteristics. Nevertheless it should be stated that the method was only moderately accurate (AUC 0.750), and a further improvement is needed to avoid PD-L1 analysis on the histological specimen.

Response : We strongly agree with the reviewer's opinions and make corresponding modifications to the article.

Changes in the text: Line 237-239

Comment. (3) ROC curves for CT scan morphologic characteristics should be shown.

Changes in the text: Figure 3, Line 215-217, 466-467

Comment. (4) In the paragraph starting with line 230, more details should be given concerning the results of PD-L1 expression prediction by the DL 3D DenseNet Model.

Changes in the text: Line 224-228

Comment. (5) In the discussion section other possible indications for the use of Deep Learning Models in the prediction of PD-L1 expression should be cited (See Sha et al. Multi-Field-of-View Deep Learning Model Predicts Non-small Cell Lung Cancer Programmed Death-Ligand 1 Status from Whole-Slide Hematoxylin and Eosin Images)

Response : We strongly agree with the reviewer's opinions and make corresponding modifications to the article.

Changes in the text: Line 292-294

Responses to Reviewer #2

The authors introduce a deep learning-based approach to investigate the association between PD-L1 expressing status and the characteristics of the CT images of advanced lung adenocarcinoma. The performance of the proposed approach compared against morphological-based features and outperformed them. The manuscript is interesting and well-written. I have some recommendations over the editorial natures as well as some major comments regarding the employed methodology. My major concern is that it seems the methodological details are provided in a hasty way. In particular:

We thank the reviewer for the critical suggestions and kind comments.

Comment. (1) Line 148: Does the manual segmentation perform only on the axial view (2D slice by slice)? Was it, then, refined in sagittal and coronal views?

Response : Yes. The labelling process is only manually segmenting the ROI slice by slice on the axial view.

Comment. (2) Did all the images from the two scanners acquire with the same spatial resolution, slice thickness, and spacing between the slices? If no, were the images resampled isotropically?

Response : In our hospital, the scanning thickness and interval of the two Canon CT are the same, CT images acquired by the two scanners are reconstructed with slice thickness and spacing of 1 mm and 0.8mm by using a high-spatial

frequency algorithm. So the images from the two scanners acquire with the same spatial resolution, slice thickness, and spacing between the slices.

Comment. (3) Which intensity windowing was applied to CT images before feeding them into the deep model? Also, what kind of intensity normalization methods used? (e.g., maximum value normalization, z-scoring, etc.)

Response : The windowing is dynamic and random around a range (50, 350) in model training phase, which is a method of data augmentation. This is to ensure that windowing will not affect the feature extraction performance, and meanwhile increase the robustness of deep model. In testing phase, there is a fixed window center and width (50, 350).

The normalization method have several steps. First, image pixel value (intensity) within the window center/ width range are reserved, otherwise set to zero. Second, further normalize the image to range [0, 255] by mapping the original range to new range using linear transformation.

Comment. (4) For both of the analyses, positive/negative, and weak/strong, the data set is quite imbalanced. What kind of strategy used to compensate that? Relating this imbalance issue, from the reported results in Figure2, it can be seen that there is almost 20 percent of performance difference between the training and validation sets. Does this large difference relate to class imbalance issues? It is recommended that authors include the accuracy-loss figures from the training and validation set, to assure the readers such a large gap between the training and validation performances are not caused by overfitting.

Response : 1. For the imbalance problem, we use oversampling strategy. In training, we oversample the samples of the class that has smaller percentage, making the two classes have nearly equally data. 2. The performance difference between training and validation data is not caused by data imbalance. The training and validation data have the same imbalance degree. The difference is because the model doesn't know the ground truth of the validation data, but know the ground truth of training data. So the performance in training data reflects the fitting performance of the model, and the performance in validation data reflects the true classification performance. 3. We choose the best performance of model for validation data, where the loss is the lowest and the accuracy is the highest. However, we are sorry that the data in the entire training process were not saved, so we currently cannot provide the accuracy-loss figure.

Comment. (5) The reported values in figure 2 for the averages should be followed by standard deviation values too.

Changes in the text: Figure 2-revised version

Comment. (6) Line 166: “900 steps”. It’s a little confusing. Does it refer to the number of epochs?

Response : It is different. In our experiment, an epoch includes many steps. The number of steps depends on the number of samples trained in each step (namely a batch).

Comment. (7) Line 168,169,170: It was mentioned that data augmentation techniques, transfer learning, and pre-trained networks were used to minimize the risk of overfitting. However, we still do not know which kind of data was used to pre-train the network. Also, in the fine-tuning phase, which layers were frozen and which layers were trained? What kind of data augmentation techniques employed? The authors are highly encouraged to elaborate on this section.

Response : 1. The pre-trained network gets parameters by training the ImageNet dataset. The original pretrained model is a 2D model, whose kernels is also 2D. We modifies the original model kernels to get a 3D model as our pretrained 3D model. 2. In fine-tuning phase, we allow all the layers to be trained so there are no frozen layers. 3. For data augmentation. First, we oversample a class’s samples that have smaller percentage, to make the two classes roughly have the same proportion. Then, data augmentation such as random image crop and flip are used during training. Thirdly, we also randomly change the window width and center as another type of augmentation.

Changes in the text: Line 153-155, 157-159

Comment. (8) Why the authors use 3D DenseNet model? In lines 297 and 298, it was mentioned that DenseNet yields significant improvement. However, no evidence was provided to validate the claim. The authors are encouraged to apply the same dataset with the same training parameters on other network architectures, e.g., 3D VGG or ResNet ... and compare the results to show the advantages of the DenseNet on the volumetric data.

Response : There should be more comparison models to test the performance among 3D models such as DenseNet, ResNet and VGG, etc. Here, we focuses on comparing the advances of “3D” property and compares it with traditional “2D” models, where traditional 2D model would analyze slices one by one or using a 2.5D approach but 3D would consider the information between slices. To this end, we choose a relatively new CNN model, which is DenseNet, considering it surpasses ResNet and VGG in the pasting competitions. We then hold the hypothesis that DenseNet in 3D version would also surpass old models, so we directly use it. The lack of 3D model comparison is a shortcoming of our present study, we will further confirm the superiority of the 3D DenseNet model in the future study.

Changes in the text: Line 295-296

Comment. (9) A major concern with all the deep learning classification models is that although the models result in high performance, it is not known which parts of the images the model focused on. One way to show the relevancy of the feature maps and the predicted class labels is to visualize the class activation maps. In this study, it will reveal where the model focused more on either the tumor edges, one certain region of tumors, or several disconnected areas. It is recommended that authors extend their research to visualize the class activation maps.

Response : This is a good suggestion. We also see some works have researched on revealing which parts of the image are really focused by the neural network. We think in medical imaging area, an explainable deep learning would be more effective and reliable. We thank you for your valuable recommendation and would considering working on this area in our next steps.

Comment. (10) In the discussion section, the authors referred to another study used Radiomics analysis. It is also highly recommended that authors perform the Radiomics study to first: objectively compare the results with the other study and second: compare the performance of deep features and hand-crafted features.

Response : In fact, in our previous work, both the radiomics and deep learning methods were used to try to predict the expression of PD-L1. We found that the prediction efficiency of deep learning was better, it was higher than that of radiomics. In order to highlight the key content of the article, the process of model comparison between radiomics and deep learning was not written in the present article.

Responses to editorial issues #3

Comment.

• Abstract: Abbreviations such as TPS and AUC were not defined.

Changes in the text: Line 12-13, 18, 98, 189-190

• Grammatical errors in lines 72,73.

Changes in the text: Line 67, 68

• Line 102: using capital letters for the abbreviations: Tumor Proportion Score (TPS)

Changes in the text: Line 98

• Too large and a little confusing sentence starts at line 86 and ends at 93. It would be better to split it. The same issue with lines 118 to 130.

Changes in the text: Line 81-89, 111-117

• Line 160: “200 samples” is very confusing when the data set includes 127 subjects.

Changes in the text: Line 144

• Line 271: “Military” does it refer to “Miliary”?

Changes in the text: Line 267

• Line 288: “interesting image features” is not a proper term. It may be replaced with “visually inspected image-based features”.

Changes in the text: Line 284