Peer Review File

Article information: http://dx.doi.org/10.21037/atm-21-25

Reviewer Comments

Comment 1: While the use of an external validation set is important, none of the models performed better than random chance in this data set (Table 1 and Table 2). The authors note that this independent data set was small (18 individuals with 9 recurrences), but further discussion of the limitations of applying the proposed technique to independent datasets is needed. Furthermore, the statements that 'The DLR-A model performed optimally on both internal and external data sets' and 'the difference between the DLR-A and the random model was statistically significant in both the internal and external groups' do not appear to be supported by the results (p values of Table 1 and Table 2 are not significant). The inability to predict recurrence in the external dataset should be included in the abstract.

Reply: Thank you very much for your valuable suggestions.

1) We acknowledge that in an external validation set, none of the models performed was better than random chance in this data set. There may be heterogeneity in the imaging data due to the different parameters in the scanning process at different centers, which can reduce the generalization ability of the prediction model. As indicated in some recent studies (Zhu X, et al. MICCAI 2017; Yang J, et al. MICCAI 2019; Zhang Y, et al. IEEE Transactions on Image Processing, 2020), domain adaptive technology based on deep learning may be applied to reduce the difference in data distribution to improve the generalization ability of the method in further studies. Some other factors, such as surgeons of different experience in different hospitals, and postoperative monitoring frequency, etc., can be influential, and require prospective studies to verify. We also recognized the independent data set of our study was small We have added this as one of the limitations in the discussion section. It needs a larger external dataset to further prove the robustness of the model in the future.

Reference:

Zhu X, et al. Maximum mean discrepancy based multiple kernel learning for incomplete multimodality neuroimaging data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2017. p. 72-80.

Yang J, et al. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2019. p. 255-63.

Zhang Y, et al. Collaborative unsupervised domain adaptation for medical image diagnosis[J]. IEEE Transactions on Image Processing, 2020, 29: 7834-7844.

2) We have corrected the mistake in the revised manuscript as advised, and the inability to predict recurrence in the external dataset has been added in the abstract as suggested.

Changes in the text:

For the first question, the changes in the text are shown in Page 19, Line 427-434.

For the second question, the changes in the text are shown in Page 3, Line 70-71; Page 16, Line 344-345; Page 20, Line 442-444 and Line 445-447.

Comment 2: How do the radiologist metrics (Supplementary Table S2) perform in the external data set? Given the limitations of radiomic and deep learning approaches in the external data set, it is of interest to compare radiologist metrics in the external data set.

Reply: Thanks for your valuable suggestions. We have supplemented the CT findings assessment to the external dataset. Among the examined CT findings, only the CT ratio and the relatively enhanced rate of the primary lesion in the arterial phase and the venous phase were significantly different between the recurrence and recurrence-free groups ($p < 0.05$). The AUC was 0.52. The accuracy, sensitivity and specificity were 0.50, 0.11 and 0.89, respectively. The performance of the recurrence risk prediction model established based on CT findings in the external independent data was shown in supplement figure S4. The results have been added in the manuscript. The limitation of the external sample size may make it difficult to find the statistical analysis results consistent with the internal data. These CT findings might be useful to some extent, but their performance in predicting the recurrence risk of pNENs after radical resection was unsatisfactory.

Changes in the text:

The results have been added in Page 3, Line 69-70; Page 13-14, Line 290-294.

The ROC was added in the supplement figure S4.

Comment 3: It is not clear why the DLR-V&A model is inferior to the DRL-A model. The authors suggest that feature redundancy may contribute to this, but further explanation is needed.

Reply: Thank you for your advice. We have added the further explanation in the manuscript.

The features extracted in DLR-V&A model may have a high degree of collinearity. Theoretically, high collinearity can lead to poor model prediction performance (Garg A, et al. International Journal of Modelling, Identification and Control, 2013). The same situation occurred in our previous study (Luo Y, et al. Neuroendocrinology, 2020). To elucidate this, we performed a collinearity analysis on the DLR features of arterial phase and venous phase, and the results showed that most of the two features have a high degree of collinearity (Supplementary Figure 5). Therefore, the redundant information brought by the highly collinear features should be the reason why the DLR-V&A model is inferior to the DRL-A model.

Reference:

Garg A, Tai K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity[J]. International Journal of Modelling, Identification and Control, 2013, 18(4): 295-312.

Luo Y, Chen X, Chen J, et al. Preoperative prediction of pancreatic neuroendocrine neoplasms grading based on enhanced computed tomography imaging: Validation of deep learning with a convolutional neural network[J]. Neuroendocrinology, 2020, 110(5): 338-350.

Changes in the text:

The further explanation was added in Page 18-19, Line 408-414.

Comment 4: The study groups local recurrence and distal metastasis together, but it would be interesting to examine each outcome separately. If the study is powered to predict these two outcomes separately it may be an interesting addition.

Reply: We fully agree with the reviewer's comments. However, in our present study, we cannot predict these two outcomes (local recurrence and distal metastasis) separately. The reasons are listed as follows: In our internal data, there were a total of 5 patients with local recurrence and 5 patients with distant metastasis. In the external data, there were only 2 patients with local recurrence and 7 patients with distant metastasis. If we set the local recurrence and distant metastasis as two separate labels in modeling, this would lead to further imbalance in the distribution of sample categories, probably resulting in decreased model performance. Moreover, both of them are progressive outcome of the disease. There are quite a few previous studies that combine the two outcomes for analysis, such as in the treatment study of melanoma (Ascierto Paolo A, Lancet Oncol, 2020) and the prognosis study of pNENs (Jie Hua, Annals of surgery, 2020; Daisuke Asano, Annals of surgery, 2020). We will continue to collect data in the future in order to collect enough data to build a model to predict these two outcomes separately. We have added this as one of the limitations in the discussion section.

Reference:

Ascierto Paolo A, et al. Adjuvant nivolumab versus ipilimumab in resected stage IIIB-C and stage IV melanoma (CheckMate 238): 4-year results from a multicentre, double-blind, randomised, controlled, phase 3 trial.[J] .Lancet Oncol, 2020, 21(11): 1465-1477.

Hua Jie, et al. Expression Patterns and Prognostic Value of DNA Damage Repair Proteins in Resected Pancreatic Neuroendocrine Neoplasms.[J] .Ann Surg, 2020, Advance online publication.

Asano Daisuke, et al. Curative Surgery and Ki-67 Value Rather than Tumor Differentiation Predict the Survival of Patients with High-grade Neuroendocrine

Neoplasms.[J] .Ann Surg, 2020, Advance online publication.

Changes in the text:

The limitation was added in Page 20, Line 449-451.

Comment 5: I've never seen traditional radiomics abbreviated 'TR' before. Unless this is standard notation, please refer to these techniques as radiomics. In the imaging literature 'TR' is commonly used to refer to 'repetition time' in MRI. Using 'TR' in a non-conventional manner in this manuscript was confusing.

Reply: Thank you for your advice. We have deleted the abbreviated 'TR' and keep the expression "radiomics".

Changes in the text:

The expression was changed in the manuscript, also in figures and tables.

Comment 6: Reference style is inconsistent. References are superscripted in the introduction but not superscripted in the discussion section.

Reply: The reference style has been modified according to the journal's requirement.

Changes in the text:

The reference style was changed into required format in all places in the manuscript.

Comment 7: The authors note the ease of automatic segmentation for the DLR method, but do not quantify the difference in time required compared to the manual segmentation approach. If the time required for automatic vs. manual segmentation was recorded it would be helpful to include this information.

Reply: Thank you very much for your valuable comment.

In the DLR method, we used the segmentation ground truth for training the network. We only need to put the CT images into the pre-trained model, then get the feature vector. However, we still need radiologists to assist in locating the tumor. We annotated (by Song and Luo with 4 and 8 years of working experience, respectively) on 5 random cases from the data. The mean time of the two radiologists to locate were 11.30s and 9.98s, and the medians were 11.04s and 9.79s. The two radiologists spent an average of

647.19s and 796.01s in the fine-delineation process, with a median of 305.51s and 382.59s, respectively. The ROI of a large tumor can be greater than 80 layers in axial CT images, so fine segmentation is more time-consuming. The relevant contents have been added in the results section.

Changes in the text:

The relevant contents were added in Page 13, Line 276-280.