Peer Review File

Article Information: Available at http://dx.doi.org/10.21037/tlcr-20-370.

Major comments:

1) The data in Abstract/Results were inconsistent with the data in the Results and Figure 3,4. Please correct.

Reply: Thank you for your comments. We have already corrected the statement for Abstract/results, to make sure that the results /figures were consistent.

Changes in the text:

The AUC for classification of AAH/AIS with MIA were 0.891, 0.841 and 0.779 for Deep-RadNet, XimaNet and XimaSharp respectively. The AUC for classification of MIA with IAC were 0.889, 0.785 and 0.778 for three networks and AUC for classification of AAH/AIS&MIA with IAC were 0.941, 0.892 and 0.827 respectively. The performance of deep_RadNet was better than the other two models with the Z-test ($p<0.05$).

2) The authors only mentioned Deep-RadNet that could improve the predictive value. If so, the authors should further compare the diagnostic value among these three methods via Z test.

Reply: Thank you for your suggestive comments. We have applied the Z-test accordingly and described the method and result both in statistical analysis and result.

Changes in the text:

Statistical analysis

Then, Z test was applied to evaluate the difference of performance among models.

Result:

Moreover, the Z-test was used to compare the performance among 3 models. P

values of the Z-test was 0.021($p < 0.05$) between deep_RadNet and XimaNet, 0.019 ($p<0.05$) between deep_RadNet and XimaSharp and 0.98 ($p>0.05$) between XimaNet and XimaSharp, which indicated that the deep_RadNet revealed the best performance.

Minor comments:

1) The data (line 313-314) was inconsistent with the data in Table 3.

Reply: Thank you for pointing out the error. We have already corrected the data in result. We have merged table 2 with table 3 together with description for table 2.

Changes in the text:

Table 2 showed deep-RadNet presented with the highest accuracy, "weighted average F1-score" and MCC in comparison with other two models in all the three classification tasks.

2) There were overall five figures and four tables. The author may consider to reduce or merge them.

Reply: Thank you for your suggestions. We have merged table2 and table3 as table 2, figure 3 and figure 4 as figure 3, table 5 changed as table 4. We deleted precious Table4, and the contents of Table 4 are described in the results section.

Changes in the text:

The new merged tables and figures were revised in the manuscript.

3) The legends of the figures are difficult to understand. And the legend of figure 5 might be in error.

Reply: Thank you for suggestions. We have revised the legend to make sure they are understandable.

Changes in the text:

Tables:

Table 1: Number of nodules for training, validation, and testing

Table 2: Classification performance of three network models

Table 3: Performance by the lesion sizes of GGNs merely

Figure Legends:

Figure 1:

a) Structure of XimaNet. Convolutional neural network (CNN) algorithm development for classification, 3D patches with size of 64*64*64 pixel were used as input. They were first fed into a BN-convolution-BN module with 64 kernels. These feature maps then went through 6 building blocks followed by a GAP module.

b) Structure of building block of XimaNet. The first building block used a convolution with stride of 1 while the other building blocks used stride of 2 for down sampling.

Figure 2:

The structure of fully connected layer network in Deep-RadNet. The numbers below each layer is the number of neurons.

Figure3:

Figure 3(a) was receiver operating characteristic curve (ROC) of AAH/AIS versus MIA. Figure 3(b) was ROC of MIA versus IAC. Figure 3(c) was the ROC of AAH/AIS&MIA versus IA.

Figure 4:

The figure illustrated the results for algorithm learning and automatic segmentation. The first to last columns were lung nodule examples selected from AAH, AIS, MIA and IAC respectively. (a) The first row showed the original CT images of

tumor area. (b) The second row showed the heat maps of the corresponding tumor area. Grad-CAM method was used to visualize the region of interest learned by XimaNet. The color bar on the most right illustrated the attention degree the algorithm paid on. (c) The third row was the segmentation result predicted by XimaSharp (red circle areas were the automatic segmentation result and blue circle areas were the ground truth)

4) The Statistical analysis part should be more specified, such as ROC curve. And the authors might consider using professional medical statistics software, such as SPSS or SAS or R.

Reply: According to your valuable suggestion. We repeated the process by Matlab, and statistical analysis have been specified already.

Changes in the text:

All statistical analysis were performed in Matlab (version 2019a；MathWorks, Narick, Mass). Receiver operating characteristic curves（ROCs）as well as areas under receiver operating characteristic curves (AUCs) were used to assess overall classification performance of the three models. Then, the Z-test was applied to evaluate the difference of performance among models. Bootstrapping (1000 boot-strap samples) was used to calculate 95% CIs and the associated P values. P value < 0.05 was considered a statistically significant difference. The performance by the size of GGNs was evaluated by the T-test. The optimal cut-off diameter for GNNs classification was calculated by searching in the dataset to maximize accuracy. Double tail distribution and double sample equal variance hypothesis were selected for parameters for tails and type, and p values were calculated under the optimal cut-off size.

5) Line 93 "At present, it is still controversial whether adjuvant treatment is needed after stage I lung cancer surgery" was invalid.

Reply: Thanks for your correction. In order to make it clear, we need to revise this sentence and quote it again.

Changes in the text:

While for patients with IAC, lobectomy and mediastinal lymph node dissection are performed; moreover, if the postoperative adjuvant treatment is applied for those, the survival rate may be improved, which is of great significance for the individualization of treatment[10,11].

10. Russell PA, Wainer Z, Wright GM, et al. Does lung adenocarcinoma subtype predict patient survival?: A clinicopathologic study based on the new International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary lung adenocarcinoma classification. J Thorac Oncol 2011;6:1496-504.

11. Ettinger DS, Akerley W, Borghaei H, et al. Non-small cell lung cancer, version 2.2013. J Natl Compr Canc Netw 2013;11:645-53; quiz 53.