Peer Review File

Article information: https://dx.doi.org/10.21037/tlcr-24-258

Reviewer A

The manuscript describes a deep learning pipeline to classify the high risk and low risk groups of the early-stage invasive lung adenocarcinoma. The deep learning model is trained on the preoperative computed tomography data and electronic medical records. The endpoints of the study were to develop a method to train AI to recognize "high" and "low" risk histologic varients. The general delivery of the manuscript is clear, and the text is easy to follow. However, there are a number of revisions recommended.

Detailed comments:

(1) While the definition of risk categories is eventually provided, it is not clear in the front sections that this pertains only to high and low risk histology, and is not in any way related to eventual outcome. This needs to be clarified in the abstract, introduction, and throughout the manuscript.

Reply: Thanks for this suggestion. We have clarified in the abstract, introduction, and throughout the manuscript that the risk categories pertain only to high and low-risk histology and are not related to the eventual outcome.

Changes in the text: (Abstract, introduction, and methods etc.) Clarifications have been added in the abstract, introduction, and throughout the manuscript where necessary, please check in manuscript.

(2) Line 128, (all pixels + 1250) / 1500 * 255.2. It is strange and unclear why the author would you 1500 and 255.2. The usage of these two numbers cannot distribute the data to [0,255] as an 8-bit image. The authors lost part of the dynamic range of the images. This will also affect the data value distribution in line 140.

Reply: Thanks for this suggestion. We find that this description is prone to misunderstandings. This process involves retaining the pixels in the range of 0-255 in the CT lung window state as a JPG format. We have fixed in manuscript to clarify.

Changes in the text: (Methods) The formula has been refined and explained in detail. 1) Window Width and Level Adjustment: Initially, the CT image underwent window width and bed position adjustment to align with the lung window. All pixel values were transformed using the formula: new_pixel_value = [clip (original_pixel_value, −1250, 250) + 1250] / 1500 × 255; where clip (x, a, b) clips the value x to the range [a, b].

(3) The authors missed a section for training hyperparameters and settings. This will

influence the repeatability of the work. Or the authors can provide GitHub link in the manuscript for the repeatability of the work. The authors can release the full version of the code upon the acceptance of the manuscript.

Reply: Thanks for this suggestion. We have included hyperparameters and settings at the end of the model design section in the paper. For detailed code, we have also sign for the data sharing statement, the researchers would get the code when contact us via email at hengrui_liang@163.com.

Changes in the text: (Methods) Added a section for training hyperparameters and settings. Please check in manuscript.

(4) From Figure 5(B), it seems the learning rate at epoch 70 is still a bit large. How did the authors design the learning rate scheduler? Supposedly, the training and validation curve should have very small variation between adjacent epochs at the convergence.

Reply: Thanks for the question. We adopted the Cosine Annealing learning rate adjustment scheme and meticulously fine-tuned the parameters as suggested. Additionally, we included a section in the paper detailing the training parameters.

Changes in the text: (Methods) Explain in detail for training hyperparameters and settings. Please check in manuscript.

(5) Line 178, at the inference stage, the model weights / parameters are all fixed. The outcome should remain the same no matter how many running times for the same inputs. It is unclear why the authors can obtain variations of the outcome by running several times. If the authors need to verify the model confidence interval, Monte Carlo sampling (perturbations to the model weights) or other Bayesian techniques (modeling the posterior distribution) should be implemented.

Reply: Sorry for the lack of clarity in our text. By "5 times," we meant that during training, we conducted 5-fold cross-validation and selected the best-performing model among them.

Changes in the text: The explanation of the implementation of Monte Carlo sampling has been added. During training, we utilized 5-fold cross-validation, and the best-performing model was selected to obtain 95% confidence interval (CI) of AUC and average sensitivity, specificity and F1 score of the test set.

(6) Line 177, using accuracy for a binary classification requires a balanced testing set. It is unfair to have an imbalanced testing set with accuracy metric to evaluate the classifier.

Reply: We balanced the test set to the extent possible compared to the training set. Additionally, for the partially imbalanced test set, we calculated other metrics including recall, F1 score, etc., to ensure a proper assessment of the model's performance.

Changes in the text: (Methods) We enriched the calculated metrics to ensure the proper evaluation of the model. Explain in detail for training hyperparameters and settings. Please check in manuscript.

(7) Line 188, missing units for the diameter of pulmonary nodules.

Reply: We have added the missing units for the diameter of pulmonary nodules.

Changes in the text: Units for the diameter of pulmonary nodules have been added.

(8) Figure 5(a), too few examples are shown. Four or five cases at least need to be demonstrated with evaluation metrics in the Figure.

Reply: We have added more examples with evaluation metrics to Figure 5(a).

Changes in the text: Additional examples with evaluation metrics have been included in Figure 5(a).

(9) Figure 6, specificity and sensitivity are missing in the Figure.

Reply: Specificity and sensitivity have been added to Figure 6.

Changes in the text: Specificity and sensitivity have been included in Figure 6.

(10) Figures 6 and 7, missing confidence interval in the Figures.

Reply: Confidence intervals have been added to Figures 6 and 7.

Changes in the text: Confidence intervals have been included in Figures 6 and 7.

(11) For both Figures 6 and 7 outcomes, the authors did not specify how do they choose the thresholds or the cut-off points to calculate the specificity and sensitivity. The threshold or the cut-off points should be selected by the optimal point of the validation set or the 0.5 in [0,1] considering using the sigmoid function. A reference paper [1] with a similar task using histology images can be helpful.

[1] Zhou, H., Watson, M., Bernadt, C.T., Lin, S., Lin, C.-y., Ritter, J.H., Wein, A., Mahler, S., Rawal, S., Govindan, R., Yang, C. and Cote, R.J. (2024), AI-guided histopathology predicts brain metastasis in lung cancer patients. J. Pathol., 263: 89-98. https://doi.org/10.1002/path.6263

Reply: We used the sigmoid function as the output layer, with a cut-off point set at 0.5, and we provided additional explanation on this in the text.

Changes in the text: Threshold selection method has been specified and the reference has been added. we applied the Sigmoid function to map the results to the 0-1 interval, with the cut-off point set at 0.5 [21].

(12) Figure 7, it is unclear whether the contribution of electronics medical data is

statistically significant (p-value missing). Line 219-221, it is questionable to draw the conclusion “A comparison between Experiment 1 and Experiment 3 revealed that clinical information remains beneficial for enhancing deep learning classification results.”
Reply: We added the p-values of two sets of ablation experiments in the results section.
Changes in the text: P-value has been included in results. A comparison between Experiment 1 and Experiment 3 revealed that clinical information remains beneficial for enhancing deep learning classification results with significant enhanced AUC performance of 3% (0.87 vs. 0.90, p=0.007). Similarly, comparing Experiment 2 and Experiment 3 demonstrated that by incorporating the merging input approach, the network could learn contextual information surrounding the nodules. However, it should be noted that the AUC improvement under this condition remained relatively modest with enhanced AUC performance of 1% (0.89 vs. 0.90, p=0.241).

(13) Figure 7 is more like a part of ablation study. If the authors want to perform such analysis, I recommend including ablation study for empirical values (probably in supplementary materials) in the manuscripts, such as number of layers for the sequence, thickness dimension of the input, etc.
Reply: We have included model details and hyperparameters in the training details, including parameters such as layer thickness mentioned in the ablation experiments.
Changes in the text: (Methods) Added a section for training hyperparameters and settings. We also re-organized the content in the result part. The Figure 7 was described as the Ablation study. Please check in manuscript.

(14) In the conclusion section, it is better to add comments about the limitation of the current method and the prospective improving directions.
Reply: Thanks for your suggestion. We have discussed the limitation in discussion part already. We are pleased to add more comments about the limitations of the current method and prospective directions for improvement.
Changes in the text: (Conclusion) Comments on limitations and future directions have been added to the conclusion. Using recent AI technologies such as federated learning and large models can potentially improve the data sharing security and increase the accuracy of our models. In the future, we still need to incorporate more centralized data to enhance the generalization of the model.

Reviewer B

This article reported developing and validating a CT scan-based deep-learning model to predict the histologic subtypes and identify high-risk pathology of stage T1 lung adenocarcinoma. This is a timely study that includes relevant methodology and information regarding the importance of predicting subtypes of adenocarcinomas for establishing surgical treatment strategies. Overall, the manuscript is well-written and easy to follow.

Remarks

Methods:
- What were the criteria of the maximal time distance between the evaluated CT scan and surgical resection? Clarification would be needed.
Reply: Thanks for your question. We have clarified the criteria for the maximal time distance between the evaluated CT scan and surgical resection (within 1 month).
Changes in the text: (Methods) The maximal time interval between CT scan and surgical resection are less than one month for all patients.

The authors defined a high-risk group as those with a micropapillary component of 5% or more or a predominant solid pattern. A reference would be needed for this definition.
Reply: Thanks for your suggestion. We have added the appropriate reference for the definition of the high-risk group, please check in manuscript.
Changes in the text: (Methods) Patients were stratified into two risk groups according to IASLC/ATS/ERS classification of lung adenocarcinoma and previous studies [15-17].

- Page 3, Paragraph 2: The distinction between Junior and Senior physicians can be arbitrary. If all physicians involved in the analyses were board-certified, I would recommend omitting the phrases 'junior' and 'senior' and unifying the phrases into board-certified physicians. For example: "The initial diagnosis was conducted by a board-certified pathologist and subsequently validated by another pathologist."
Reply: Thanks for your suggestion. We have unified the phrases into "board-certified physicians."
Changes in the text: (Methods) The distinction between junior and senior physicians has been removed, replaced with "board-certified physicians".

Results:
Page 5, Lines 208-217: These paragraphs describe the methods of analysis using additional clinical features other than the CT scan images. They would better fit in the Methods section.
Reply: We have moved the paragraphs describing methods of analysis using additional

clinical features to the Methods section.

Changes in the text: (Methods) The relevant paragraphs have been moved to the Methods section: Model development for pathological pattern classification based on CT and EMR data.

Discussion:

- Page 6, Paragraph 2: As the authors noted, there are previous reports on AI models developed to predict the histologic pattern of lung cancers. Comparisons of the performance outcomes between the current model and previous models would be of interest to the readers.

Reply: We have added a comparison of performance outcomes between the current model and previous models.

Changes in the text: (Discussion) Comparison with previous models has been added. Several studies have explored the application of AI in the preoperative discrimination of lung invasive adenocarcinoma. For instance, Fan et al.[28] identified a radiomics signature that allows for the preoperative discrimination of lung invasive adenocarcinomas from non-invasive lesions manifesting as ground-glass nodules. This radiomics signature demonstrated a strong ability to differentiate between invasive adenocarcinomas and non-invasive lesions, achieving accuracy rates of 86.3%. In another study, Wang et al.[12] investigated imaging phenotyping by combining radiomics with deep learning to predict high-grade patterns within lung adenocarcinomas. This study included 111 patients who were identified as having ground-glass opacities The proposed method achieved an overall accuracy of 0.913. Compared to other studies, our multi-center research with bigger sample size has validated the generalization of the model to a certain extent. Additionally, the application of deep learning has improved the accuracy of previous radiomics-based studies. Finally, the incorporation of EHR information has allowed our multimodal model to be the first to verify whether EHR data has additional significance in this context.

Minor comments:

- Page 4 Results: Units should be provided (perhaps cm) with the size description of pulmonary nodules.

Reply: We have added the units for the size description of pulmonary nodules.

Changes in the text: Units (cm) have been added (Page 4, line 150).

- Figure 6 and Figure 7: It needs to be clarified which graph matches with which cohort. Clarification and annotations would be needed.

Reply: We have added annotations to clarify which graph matches with which cohort.

Changes in the text: Annotations have been added to Figures 6 and 7 (Page 8, Figures 6 and 7). Figure 6: Orange curve represent the AUC of cohort 2 while the Blue curve represent the AUC of cohort 1. Figure 7 was conducted in test set of cohort 1.