Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-21-870

- Reviewer A:

This study aimed to build a deep convolutional neural network for the automatic 30 classification of malignant involvement in thoracic LNs using EBUS. The article was well-written. The results were interesting and clear. My comments were the below.

Comment 1:

The shape of LNs showed higher sensitivity, specificity and accuracy than the VGG-16. This data means malignant prediction by sonographic feature classification is more useful than the VGG-16. To show the usefulness of the VGG-16, please describe the feature of LNs that were predicted as non-malignancy by sonographic feature, but as malignancy by the VGG-1

Reply 1: We greatly appreciate the Reviewer's comments and agree completely. We have examined all the false-negative cases in the sonographic feature that were identified as a malignancy in the VGG 16 model. We identified 18 cases that were false-negatives in terms of shape (the oval shape on the sonographer's on-site evaluation was considered negative). We analyzed the predictive performance of other sonographic features in these false-negatives in terms of shape and have outlined the results in Table 1.

Table 1. Sensitivity of features for LNs that were predicted as non-malignancy by EBUS due to shape, but as malignancy by VGG-1.

			Sonographic features				
Pathology	VGG-16	Shape	Shape	Margin	Echogenicity	Central hilar structure	Coagulation necrosis sign
Malignancy	Malignancy	N, false negative	18	1	5	6	18
		N, true positive	0	17	13	12	0
		N, total	18	18	18	18	18
		Sensitivity (%)	0	94.4	72.2	66.7	0

These results showed large differences in the sensitivity of each sonographic feature in the 18 false-negative cases in terms of shape. In addition, we analyzed the false-negative cases according to every sonographic feature, confirming large differences in the predictive value of malignancy between each ultrasound feature. Therefore, it is possible to improve the predictive value of malignancy by simultaneously and comprehensively analyzing various characteristics on the ultrasound. For this reason, Hylton et al[1]. developed the four-point scoring system using four sonographic features; short-axis diameter, margins, central hilar structure, and necrosis, which demonstrated good performance in identifying malignant LNs. If all these sonographic features can be evaluated comprehensively and simultaneously using a deep learning model, the model will be a robust predictor for the classification of malignant LNs.

Additionally, we analyzed the diagnostic value of shape alone versus shape with VGG-16. The diagnostic performance was further improved across all values with the addition of the VGG-16 model (Table 2).

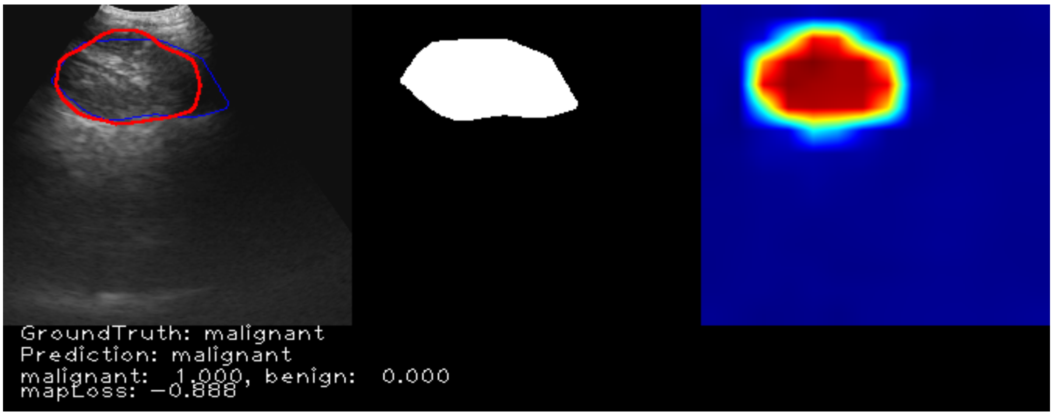
Table 2. Malignancy prediction performance for shape alone versus shape with VGG-16.

	Sensitivity (%)	Specificity (%)	NPV (%)	PPV (%)	Accuracy (%)	P value

Shape	83.6	86.1	88.6	80.3	85.1	<0.001
Shape with VGG-16	94.0	86.1	93.5	87.0	90.0	<0.001

§ Abbreviation: NPV, negative predictive value; PPV, positive predictive value

We have outlined an example of a false-negative case that was identified as non-malignant by EBUS due to shape, but as malignancy by the VGG-1 model (Figure 1).



```
GroundTruth: malignant
Prediction: malignant
malignant:  1.000, benign:  0.000
mapLoss: -0.888
```

Figure 1. An example of a proven malignant lymph node, mistaken for benign by EBUS due to shape.

Changes in the text: This has been reflected in discussion part on page 12, lines 213 ~ 220. Table 1, 2, and Figure 1 are the results of additional analyses for the reviewer and are not included in the manuscript.

Comment 2;

Please describe how many times EBUS-TBNA was performed for each LNs. It has an impact on the diagnostic rate.

Reply 2: We appreciate the Reviewer's helpful suggestion. We agree with the Reviewer's comment and have retrospectively reviewed the procedure records of all participants (310 patients). The median number of EBUS-TBNA performed for each lymph node was 2 per LN (with IQR 1.0-2.0). We have changed the script according to the result as suggested.

Changes in the text: We have added this result on page 9, line 154.

Comment 3;

Please describe the reason that the threshold was set to 0.5 for decision of whether the LN is malignant or benign. Does the threshold of 0.5 means an 50% chance of malignancy, like 85% in Figure 4?

Reply 3: We greatly appreciate the Reviewer's comments. We set the threshold to 50% during training to optimize the AI function, extinguishing benign lymph nodes to obtain more information about the lymph nodes that are likely to be malignancy, while sustaining enough sensitivity. A threshold of 0.5 refers to a 50% possibility of malignancy prediction each time the AI evaluates the lymph nodes in real-time. That is, the value of 50% is the value used as the threshold during training, however the actual value of malignancy prediction of VGG-16 is showed in the range of 0-100%.

Changes in the text: We have modified the text as advised; on page 6, we deleted line 93, and added a modified phrase on page 8, lines 125-129.

- Reviewer B

This manuscript describes a retrospective study of the application of convolutional neural networks to EBUS images of thoracic lymph nodes in an attempt to automate the determination of benign and malignant status. As the authors mention, there are limits to human judgment, and the application of AI is sympathetic. However, for the reasons listed in the major comments, it is difficult to think that the subjects of the analysis are appropriate, and the results obtained are not credible. The research content is interesting, and I recommend the authors to resubmit it again with the appropriate subjects for analysis.

Major comments

1. The authors should be aware of the need to base the analysis on pathological results, which are the golden standard, in order to construct the system. In fact, from the description in lines 167–168, 340 malignant lymph nodes and 548 benign lymph nodes were confirmed. Nevertheless, the final analysis includes 935 malignant lymph nodes and 1,459 benign lymph nodes, as described in lines 157–158 and shown in Figure S1. I do not know the details of this discrepancy in the numbers as they are not described, but I do not understand why only one image per lymph node was not analyzed. In other words, if the number of images extracted for each lymph node is different, the weights affecting the analysis of each lymph node will be different, and the correct results cannot be obtained.

Reply 1: We greatly appreciate the Reviewer's comments, and fully agree. The lymph node described in lines 167-168 and 340 refers to the actual number of mediastinal lymph nodes after pathologic confirmation, either by EBUS-TBNA or surgical resection. However, the number of lymph nodes used in the final analysis is the number of "images" taken during EBUS-TBNA. As the reviewer mentioned, the number of images extracted from each lymph node is different in this study. We take multiple images of the same lymph nodes; thus, one lymph node may have more than two images.

However, artificial intelligence learning is not related to the similarity of images; the pathology results and an accurate outline of the lymph nodes are the most crucial factors. In this study, every image was double-checked for image quality and accurately outlined by an experienced bronchoscopist. Therefore, training AI with similar images of the same lymph node has no effect on analysis, regardless of multiple images of the same lymph node being used.

2. Figure 4 shows a representative case that was analyzed, but the area of the extracted lymph nodes is inappropriate and apparently includes the surrounding soft tissue. Furthermore, one of the reasons for exclusion in Figure S1 is "images of LNs during TBNA", but the image in Figure 4 were taken during TBNA. Line 98 states that a bronchoscopist marked it, but it should also state who was in charge.

Reply 2: Once again, we greatly appreciate the Reviewer’s comments. The AI is trained to distinguish a lymph node from other mediastinal architecture, but in evaluating malignant potency, an adjacent area near the lymph node is also used for training. The white line indicates the area that the AI has evaluated as a possible malignancy, rather than the lymph node itself. Furthermore, Figure 4 is the representative image, however the AI does not depend on only a single image from a brief moment. The AI adjusts its evaluation depending on real-time image flow, as outlined in the submitted movie clip. The movie clip shows that the area that the AI evaluates as a lymph node keeps changing in time, indicating this real-time evaluation.

The bronchoscopist in charge of marking the outlines of the lymph nodes is Dr. Yong, Seung Hyun, the first author of this research. We have changed the manuscript according to the Reviewer's suggestion.

Changes in the text: We added detailed information about LN boundaries extraction in the methods section on page 7, lines 99-102, and 114-116. Statement on bronchoscopist is added on page 5, line 70.

3. At all, the detail of subjects of analysis should be written clearly in the methods.

Reply 3: Thank you for your comment. The Methods section has been clarified in the revised manuscript as you suggested.

Changes in the text: We have reinforced details of the cross-validation, network architecture, training, and statistical analyses in the Methods section on page 6, lines 73-77, 86-88, 91-92, and page 7, lines 99-102, 114-116.

Minor comment

1. Minor errors are noticeable, such as improper articles, missing commas and periods, and spaces between numbers and %.

Reply 1: We appreciate the Reviewer's comments. There have been amended as you suggested.

2. PET should be correctly described as FDG-PET.

Reply 2: We apologize for this omission. We have corrected PET to full acronym as you suggested.

Changes in the text: We have modified our text as advised on page 11, lines 202-203, 204 and 207.

- Reviewer C

Nice results and interesting topic, I congratulate the authors for the work; however, many grammatical changes are needed in the abstract, introduction, and results before resubmission for a second review. Few little suggestions for page 1, as an example, are listed below.

Abstract, page 1, line 30, background part, consider changing to "classification of metastatic malignancies involving thoracic LN diagnosed by EBUS-TBNA.

Reply 1: We greatly appreciate the Reviewer's comments. We fully agree and have made revisions in lines with your suggestion.

Changes in the text: we have modified our text as advised on page 2, lines 6.

Abstract, page 1, line 31.. methods part, consider changing to "Patients who underwent EBUS-TBNAs to assess presence of malignancy in mediastinal lymph nodes during a ten month period at Severance Hospital, Seoul, Korea were included in the study. Corresponding LN ultrasound images, pathology reports, demographic data and clinical history were collected and analyzed.

Reply 2: Once again, thank you for the reviewer's comments. We have revised the manuscript as suggested.

Changes in the text: we have modified our text as advised on page 2, lines 7-10.

Abstract, page 1, line 34.. Results part, consider changing to "A total of 2,394 EBUS images of 1,459 benign LN from 193"

Reply 3: We appreciate the Reviewer's comments. We completely agree and have revised the manuscript as suggested.

Changes in the text: we have modified our text as advised on page 2, line 112.

Consideration to review the grammar for clarity purposes and style in the remainder of the document, particularly in the abstract, introduction, methods and results are required.

Reply 4: We apologize for the errors in grammar and style. We have made appropriate revisions throughout the manuscript, as suggested.

1. Hylton DA, Turner S, Kidane B, Spicer J, Xie F, Farrokhyar F, et al. The Canada Lymph Node Score for prediction of malignancy in mediastinal lymph nodes during endobronchial ultrasound. J Thorac Cardiovasc Surg 2020;159:2499-507.e3.