BMJ Open Respir Research

Deep learning with test-time augmentation for radial endobronchial ultrasound image differentiation: a multicentre verification study

Kai-Lun Yu,1,2 Yi-Shiu-Tsen Tseng,3 Han-Ching Yang,1 Chia-Jung Liu,1 Po-Chih Kuo,3 Meng-Rui Lee,4 Chun-Ta Huang,4 Lu-Cheng Kuo,4 Jann-Yuan Wang,4 Chao-Chi Ho,4 Yin-Yuan Shih,2,4 Chong-Jen Yu.4

ABSTRACT
Purpose Despite the importance of radial endobronchial ultrasound (rEBUS) in transbronchial biopsy, researchers have yet to apply artificial intelligence to the analysis of rEBUS images.
Materials and methods This study developed a convolutional neural network (CNN) to differentiate between malignant and benign tumours in rEBUS images. This study retrospectively collected rEBUS images from medical centres in Taiwan, including 769 from National Taiwan University Hospital Hsin-Chu Branch, Hsinchu Hospital for model training (615 images) and internal validation (154 images) as well as 300 from National Taiwan University Hospital NTUH-TPE and 92 images were obtained from National Taiwan University Hospital Hsin-Chu Branch, Biomedical Park Hospital (NTUH-BIO) for external validation. Further assessments of the model were performed using image augmentation in the training phase and test-time augmentation (TTA).
Results Using the internal validation dataset, the results were as follows: area under the curve (AUC) (0.88 (95% CI 0.83 to 0.92)), sensitivity (0.80 (95% CI 0.73 to 0.88)), specificity (0.75 (95% CI 0.66 to 0.83)). Using the NTUH-TPE external validation dataset, the results were as follows: AUC (0.76 (95% CI 0.71 to 0.80)), sensitivity (0.58 (95% CI 0.50 to 0.65)), specificity (0.92 (95% CI 0.88 to 0.97)). Using the NTUH-BIO external validation dataset, the results were as follows: AUC (0.72 (95% CI 0.64 to 0.82)), sensitivity (0.71 (95% CI 0.55 to 0.86)), specificity (0.76 (95% CI 0.64 to 0.87)). After fine-tuning, the AUC values for the external validation cohorts were as follows: NTUH-TPE (0.78) and NTUH-BIO (0.82). Our findings also demonstrated the feasibility of the model in differentiating between lung cancer subtypes, as indicated by the following AUC values: adenocarcinoma (0.70; 95% CI 0.64 to 0.76), squamous cell carcinoma (0.64; 95% CI 0.54 to 0.74) and small cell lung cancer (0.52; 95% CI 0.32 to 0.72).
Conclusions Our results demonstrate the feasibility of the proposed CNN-based algorithm in differentiating between malignant and benign lesions in rEBUS images.

WHAT IS ALREADY KNOWN ON THIS TOPIC
⇒ Radial endobronchial ultrasound (rEBUS) is an important technique for localising lung lesions in transbronchial biopsy.

WHAT THIS STUDY ADDS
⇒ We present a convolutional neural network (CNN) to differentiate between malignant and benign lesions in rEBUS images.
⇒ The feasibility of using the CNN model to interpret rEBUS images was demonstrated via validation using both internal and external cohorts.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY
⇒ Our CNN model serves as a valuable adjunct to rEBUS in transbronchial biopsy.
⇒ This non-invasive procedure could potentially be used to facilitate the identification of lung lesions with a high probability of malignancy.

INTRODUCTION
Lung cancer is the leading cause of cancer-related mortality worldwide, accounting for 1.8 million deaths annually.1 Despite advances in imaging techniques, a deeper understanding of underlying biological mechanisms and improved treatment options, the 5-year survival rate for lung cancer remains below 30%.2 3 Early detection and timely intervention are crucial to improving patient outcomes. Radial endobronchial ultrasound (rEBUS)-guided transbronchial biopsy (rEBUS-TBB) has emerged as a valuable tool for lung cancer diagnosis.4 5 rEBUS represents a significant advancement in the visualisation and localisation of peripheral pulmonary lesions during bronchoscopy with a corresponding improvement of outcomes for TBB.6 7 The utilisation of rEBUS in peripheral bronchoscopy has been widely accepted and incorporated in recent guidelines.8 9 Nonetheless, rEBUS is used primarily to localise lesion in clinical practice, due to its limited utility...
for differential diagnosis. Some rEBUS features, such as homogeneous internal echoes and concentric circles, are suggestive of malignant lesions; however, the definition of these features remains somewhat arbitrary and dependent on the subjective interpretation of the operator. Bronchoscopists are hindered in the selection of appropriate criteria for use in diagnosing lesions as benign or malignant, due to its heavy reliance on individual experience with EBUS.

Artificial intelligence (AI) has been widely applied to the interpretation of medical images (e.g., chest X-rays, CT, and magnetic resonance), and has achieved remarkable success in the detection and classification of diseases. In fact, the accuracy of AI models has been shown to match or exceed that of human physicians. Nonetheless, the application of deep learning algorithms to ultrasound images has been hindered by artefacts and noise, thereby limiting its applicability to the rEBUS images used for lung cancer detection or characterisation. One recent study achieved satisfactory discriminant performance when applying a convolutional neural network (CNN) to radial endobronchial ultrasound images; however, those results were based on a relatively small dataset and lacked external (multicentre) validation and inapplicable to clinical use. Researchers have yet to explore the application of AI to the subtyping of lung cancers based on rEBUS images.

The objective in the current study was to assess the feasibility of using deep learning algorithms to differentiate between malignant and benign lesions in rEBUS images, based on a multicentre cohort of patients who underwent rEBUS-TBB.

METHOD

Patient and public involvement
The data used in this study were collected retrospectively. Participants did not receive feedback on the results and were not directly involved in the study.

Study population
Patients who underwent rEBUS-TBB as part of routine care were recruited from three hospitals in northern Taiwan: National Taiwan University Hospital (NTUH-TPE), National Taiwan University Hospital Hsin-Chu Branch, Hsinchu Hospital (NTUH-HC) and National Taiwan University Hospital Hsin-Chu Branch, Biomedical Park Hospital (NTUH-BIO) between January 2016 and August 2021.

Bronchoscopic procedures
The bronchoscopic procedure used in this study was rEBUS-TBB for sampling peripheral pulmonary nodules. Most of the procedures were performed under local anaesthesia (lidocaine); however, some patients underwent mild-to-moderate sedation using intravenous fentanyl and midazolam, based on the judgement of the bronchoscopist. Conventional bronchoscopy (BF-P260F, BF-P290 or BF-1T; Olympus, Tokyo, Japan) was used to examine the trachea and bronchi, after which rEBUS images were obtained using an endoscopic ultrasound device (Olympus EU-ME2 at NTUH-HC and NTUH-BIO; Olympus EU-M30S at NTUH-TPE or Fujifilm SP-900 at NTUH-BIO) with a 20 MHz radial ultrasonic probe (Olympus UM-S20-17S or UM-S20-20R at NTUH-HC, NTUH-TPE and NTUH-BIO or Fujifilm PB-2020M at NTUH-BIO). The position of the rEBUS probe was reported to be within or adjacent to the target lesion. Once the lesion was identified, the radial probe was withdrawn from the working channel, to allow the insertion of biopsy forceps for TBB. In some instances, bronchial brushing and/or washing were performed. Specimens obtained in the TBB were prepared for pathological and cytological examination by a qualified pathologist.

Image datasets
All of the collected rEBUS images were submitted with corresponding pathological reports for review by board-certified pathologists. The rEBUS images were divided into four major categories, based on pathological and cytological reports. The four major categories were as follows: malignant, benign, non-specific findings and other findings. The malignant category included the following subcategories: adenocarcinoma, squamous cell carcinoma, non-small cell lung cancer-not otherwise specified, small cell carcinoma, atypia and other malignancies. The benign category included the following subcategories: benign tumours, acute and chronic inflammation, granulomatous inflammation and organising pneumonia. In our institution, patients who presented atypical cells were routinely scheduled to undergo repeated biopsies using rEBUS-TBB or alternative diagnostic modalities. In the current study, cases that initially presented atypical cells, but were later confirmed as malignant through other biopsy methods, were reclassified as atypia indicative of malignancy. Note that no benign tumours (e.g., hamartomas) were identified via rEBUS-TBB. Cases involving lesions that were pathologically confirmed as benign were subject to clinical follow-up for at least 6 months to confirm the diagnosis. The non-specific findings category included cases without evidence of malignancy as well as those with non-specific findings. The other findings category included fibrosis and haemorrhage. This study selected only images corresponding to pathologically confirmed cases of malignant or benign tumours. The diagnostic yield of EBUS-TBB was approximately 70%. In several notable studies that focused on the diagnostic yield of rEBUS-TBB, cases involving fibrosis in the absence of tumour cells or specific inflammation were commonly classified as non-diagnostic. Categorising fibrosis as benign could potentially impede the development of an effective diagnostic model. Images designated as non-specific or other findings were excluded from further analysis.
Automated image cropping

Figure 1 presents a flow chart of the procedures followed in this study. A median filter was first applied to the rEBUS images to remove noise. Histogram equalisation and a Canny filter were then used to extract edges. Finally, the Hough circle transform was used to detect the (circular) probe and its centre and crop the images to a uniform size (300×300 pixels) with the radial probe located in the centre. Standardising the images in this way made it easier to focus on peripheral tissue and lesion region.

Image denoising

Speckle noise was detected in many of the rEBUS images obtained using rEBUS probes that had undergone repeated use. Denoising was performed by successively applying an opening operation to the images using two basic morphological operators, erosion and dilation. During noise removal, the quality of the rEBUS images was preserved by performing connected component analysis to extract the largest connected area, where the content of the original image was retained and the background was replaced by the denoised image. Examples of denoised images are presented in online supplemental figure 1.

Image augmentation

Data imbalance could lead to model overfitting, that is, the model would be more likely to predict results in the majority class. Data augmentation was used to expand the number of images in the two classes (malignant and benign) to 400 images each. We randomly adjusted the rotation of the images from 0 to 360 degrees and the scale of the images from 0.5 to 1.5. We randomly adjusted the brightness and contrast of the images to within set ranges of (0.3–1.5) and (0.8–1.2), respectively. Examples of image augmentation are presented in online supplemental figure 2.

CNN-based model construction

We developed a CNN to differentiate between malignant and benign tumours in rEBUS images. The proposed EfficientNet model is based on a limited parameter set, which makes it ideal for the analysis of small datasets. EfficientNet-B0 pretrained on ImageNet was used as our backbone model. The outputs of the EfficientNet were fed into a global average pooling layer, to which were added three dense layers with RELU activation functions and dropout layers (dropout rate=0.3) to reduce the likelihood of overfitting. Finally, we added a dense layer with softmax activation to perform binary classification. During the training process, Talos was used to tune the hyperparameters using the validation dataset (33% of the training dataset). The hyperparameters with the highest area under the curve (AUC) were selected.

Test-time augmentation

In evaluating the performance of the model, we applied test-time augmentation (TTA) to both internal and external validation datasets. For both internal validation (NTUH-HC) and external validation (NTUH-TPE and NTUH-BIO), the scale of the rEBUS images was adjusted from 0.6 to 1.4 and each image was used to generate 14 different images for data augmentation. The resulting model was subsequently used to classify a set of 15 images, and the final diagnostic result was determined based on the majority vote among the classifications.

Model evaluation

Our primary objective in this study was to assess the performance of the proposed CNN in differentiating between malignant and benign lesions. Performance in the internal and external validation cohorts was assessed in terms of the AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and F-1 score. Pathological and cytological reports were used as the ground truth for comparison. To simulate real-world clinical utilisation, the model was fine-tuned using 10% of the data from the external validation cohorts (NTUH-TPE and NTUH-BIO). We also assessed the diagnostic performance of the CNN in differentiating specific subtypes of lung cancer (adenocarcinoma, squamous cell carcinoma and small cell lung) from other subtypes.

Statistical analysis

Python V.3.8 and TensorFlow V.2.9.0 were used for image processing, model training and model evaluation. All
categorical variables were analysed using Pearson’s $\chi^2$ tests, except where a small size ($<5$) required the use of Fisher’s exact test. We reported the point estimates and 95% CI of the validation metrics: AUC, sensitivity, specificity, PPV, NPV, accuracy and F1-score. 95% of CIs were computed by bootstrapping the scores of the predictions 1000 times.

**RESULTS**

Online supplemental figure 3 presents a flow diagram providing an overview of the cases included in this study as well as the number of cases that were excluded and underlying reasons. After excluding images with poor quality and images designated as non-specific or other findings, a total of 769 rEBUS images depicting 265 lesions in NTUH-HC were collected from 260 patients for model training (615 images, 207 lesions) and internal validation (154 images, 58 lesions). This training cohort included 139 malignant lesions, most of which were adenocarcinoma (n=92, 66.2%), followed by squamous cell carcinoma (n=16, 11.5%). A total of 68 lesions were benign, most of which were inflammation (80.9%). The internal validation cohort included 37 malignant lesions and 21 benign lesions. A total of 300 rEBUS images of 190 lesions (malignancy, n=119, 62.6%) were collected from the NTUH-TPE external validation cohort and 92 images of 35 lesions (malignancy, n=16, 45.7%) were collected from the NTUH-BIO external validation cohort for external testing. Table 1 details the characteristics of the patients, lesions, rEBUS images and pathologies.

**Figure 2** illustrates the process of image collection, processing and analysis. Table 2 lists the diagnostic performance of the proposed CNN in identifying malignancies using various analysis techniques. When using conventional CNN analysis, the model obtained the following internal validation results: AUC (0.88 (95% CI 0.83 to 0.93)), accuracy (0.81 (95% CI 0.71 to 0.88)), sensitivity (0.85 (95% CI 0.78 to 0.92)) and specificity (0.97 (95% CI 0.88 to 0.99)). When using the NTUH-TPE external validation dataset, the results were as follows: AUC (0.75 (95% CI 0.71 to 0.81)), accuracy (0.68 (95% CI 0.63 to 0.74)), sensitivity (0.53 (95% CI 0.49 to 0.58)) and specificity (0.86 (95% CI 0.82 to 0.90)). When using the NTUH-BIO external validation dataset, the results were as follows: AUC (0.65 (95% CI 0.61 to 0.71)), accuracy (0.64 (95% CI 0.53 to 0.75)), sensitivity (0.67 (95% CI 0.57 to 0.85)) and specificity (0.62 (95% CI 0.56 to 0.78)).

When the TTA technique was applied, the AUC for internal validation was 0.88 (95% CI 0.83 to 0.92). When using the NTUH-TPE external validation dataset, the results were as follows: AUC (0.76 (95% CI 0.71 to 0.80)), accuracy (0.73 (95% CI 0.68 to 0.78)), sensitivity (0.58 (95% CI 0.50 to 0.65)), specificity (0.92 (95% CI 0.88 to 0.97)) and F1-score (0.71 (95% CI 0.64 to 0.78)). When using the NTUH-BIO external validation dataset, the results were as follows: AUC (0.72 (95% CI 0.64 to 0.82)), accuracy (0.74 (95% CI 0.65 to 0.83)), sensitivity (0.71 (95% CI 0.55 to 0.86)), specificity (0.76 (95% CI 0.64 to 0.87)) and F1-score (0.69 (95% CI 0.56 to 0.80)).

Table 1  Characteristics of patients, lesions and rEBUS images

<table>
<thead>
<tr>
<th>Patient numbers</th>
<th>NTUH-HC training</th>
<th>NTUH-HC internal validation</th>
<th>NTUH-TPE</th>
<th>NTUH-BIO</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (mean±SD)</td>
<td>66.2±12.9</td>
<td>69.0±12.6</td>
<td>64.7±12.2</td>
<td>66.6±11.9</td>
<td>0.081</td>
</tr>
<tr>
<td>Gender, male (%)</td>
<td>126 (62.4)</td>
<td>35 (60.3)</td>
<td>115 (61.2)</td>
<td>20 (57.1)</td>
<td>0.820</td>
</tr>
<tr>
<td>Lesion numbers/Image number</td>
<td>207/615</td>
<td>58/154</td>
<td>190/300</td>
<td>35/92</td>
<td></td>
</tr>
<tr>
<td>Malignant (%)</td>
<td>139 (69.1)/382 (62.1)</td>
<td>37 (63.8)/96 (62.3)</td>
<td>119 (62.6)/168 (56.0)</td>
<td>16 (45.7)/38 (41.3)</td>
<td>0.110/0.001</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>92 (66.2)/256 (67.0)</td>
<td>20 (54.1)/64 (66.7)</td>
<td>70 (58.8)/99 (58.9)</td>
<td>9 (56.3)/19 (50.0)</td>
<td>0.121/0.075</td>
</tr>
<tr>
<td>SqCC</td>
<td>16 (11.5)/35 (9.2)</td>
<td>4 (10.8)/9 (9.4)</td>
<td>13 (10.9)/19 (11.3)</td>
<td>3 (18.8)/5 (13.2)</td>
<td>0.113/0.781</td>
</tr>
<tr>
<td>NSCLC-NOS</td>
<td>7 (5.0)/16 (4.2)</td>
<td>3 (8.1)/4 (4.2)</td>
<td>7 (5.9)/13 (7.7)</td>
<td>2 (12.5)/7 (18.4)</td>
<td>0.820/0.002</td>
</tr>
<tr>
<td>SCLC</td>
<td>4 (2.9)/12 (3.1)</td>
<td>2 (5.4)/4 (4.2)</td>
<td>5 (4.2)/7 (4.2)</td>
<td>1 (6.3)/3 (7.9)</td>
<td>0.816/0.559</td>
</tr>
<tr>
<td>Metastatic malignancy</td>
<td>4 (2.9)/24 (6.3)</td>
<td>4 (10.8)/6 (6.3)</td>
<td>3 (2.5)/4 (2.4)</td>
<td>1 (6.3)/4 (10.5)</td>
<td>0.105/0.173</td>
</tr>
<tr>
<td>Atypia</td>
<td>16 (11.5)/39 (10.2)</td>
<td>4 (10.8)/9 (9.4)</td>
<td>21 (17.6)/26 (15.5)</td>
<td>0 (0)/0 (0)</td>
<td>0.113/0.033</td>
</tr>
<tr>
<td>Benign (%)</td>
<td>68 (32.9)/233 (37.9)</td>
<td>21 (36.2)/58 (37.7)</td>
<td>71 (37.4)/132 (44.0)</td>
<td>19 (54.3)/54 (58.7)</td>
<td>0.110/0.001</td>
</tr>
<tr>
<td>Inflammation</td>
<td>55 (80.9)/179 (76.8)</td>
<td>14 (66.7)/45 (77.6)</td>
<td>51 (71.8)/94 (71.2)</td>
<td>18 (94.7)/52 (96.3)</td>
<td>0.098/0.003</td>
</tr>
<tr>
<td>Granulomatous inflammation</td>
<td>8 (11.8)/29 (12.4)</td>
<td>3 (14.3)/7 (12.1)</td>
<td>9 (12.7)/18 (13.6)</td>
<td>1 (5.3)/2 (3.7)</td>
<td>0.811/0.268</td>
</tr>
<tr>
<td>Organising pneumonia</td>
<td>3 (4.4)/14 (6.0)</td>
<td>2 (9.5)/4 (6.1)</td>
<td>8 (11.3)/16 (12.1)</td>
<td>0 (0)/0 (0)</td>
<td>0.245/0.023</td>
</tr>
<tr>
<td>Fungal infection</td>
<td>2 (2.9)/11 (4.7)</td>
<td>2 (9.5)/2 (3.4)</td>
<td>3 (4.2)/4 (3.0)</td>
<td>0 (0)/0 (0)</td>
<td>0.436/0.319</td>
</tr>
</tbody>
</table>

NSCLC-NOS, non-small cell lung cancer not otherwise specified; NTUH-BIO, National Taiwan University Hospital Hsin-Chu Branch, Biomedical Park Hospital; NTUH-HC, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu Hospital; NTUH-TPE, National Taiwan University Hospital; rEBUS, radial endobronchial ultrasound; SCLC, small cell lung cancer; SqCC, squamous cell carcinoma.
Fine-tuning further enhanced the performance of the CNN model with TTA. Using the NTUH-TPE external validation dataset, the results were as follows: AUC (0.78 (95% CI 0.73 to 0.82)), accuracy (0.80 (95% CI 0.74 to 0.83)), sensitivity (0.76 (95% CI 0.70 to 0.82)), specificity (0.80 (95% CI 0.72 to 0.86)) and F1-score (0.79 (95% CI 0.74 to 0.84)). Using the NTUH-BIO external validation dataset, the results were as follows: AUC (0.82 (95% CI 0.73 to 0.90)), accuracy (0.79 (95% CI 0.71 to 0.87)), sensitivity (0.79 (95% CI 0.65 to 0.91)), specificity (0.87 (95% CI 0.77 to 0.69)) and F1-score (0.80 (95% CI 0.68 to 0.89)).

Figure 3 presents the receiver operating characteristic curves indicating the diagnostic performance of the CNN in identifying malignant lesions using pathological results as ground truth. The diagnostic performance as a function of analysis technique is detailed in the online supplemental tables 1–4. In the NTUH-BIO cohort, we

<table>
<thead>
<tr>
<th>Analysis Technique</th>
<th>NTUH-TPE</th>
<th>NTUH-BIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional method</td>
<td>0.75 (0.71 to 0.81)</td>
<td>0.65 (0.61 to 0.71)</td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>0.77 (0.64 to 0.74)</td>
<td>0.69 (0.52 to 0.77)</td>
</tr>
<tr>
<td>Test-time augmentation</td>
<td>0.76 (0.71 to 0.80)</td>
<td>0.72 (0.64 to 0.82)</td>
</tr>
<tr>
<td>Fine-tuning plus test-time augmentation</td>
<td>0.78 (0.73 to 0.82)</td>
<td>0.82 (0.73 to 0.90)</td>
</tr>
</tbody>
</table>

AUC, area under the curve; NTUH-BIO, National Taiwan University Hospital Hsin-Chu Branch, Biomedical Park Hospital; NTUH-HC, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu Hospital; NTUH-TPE, National Taiwan University Hospital.

Table 2 Diagnostic performance of proposed model in differentiating between malignant and benign lesions using various analysis techniques

<table>
<thead>
<tr>
<th>Analysis Technique</th>
<th>AUC (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Accuracy (95% CI)</th>
<th>F1-score (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional method</td>
<td>0.75 (0.71 to 0.81)</td>
<td>0.53 (0.49 to 0.58)</td>
<td>0.86 (0.82 to 0.90)</td>
<td>0.68 (0.63 to 0.74)</td>
<td>0.64 (0.58 to 0.70)</td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>0.77 (0.64 to 0.74)</td>
<td>0.79 (0.73 to 0.85)</td>
<td>0.60 (0.51 to 0.68)</td>
<td>0.71 (0.65 to 0.76)</td>
<td>0.75 (0.70 to 0.80)</td>
</tr>
<tr>
<td>Test-time augmentation</td>
<td>0.76 (0.71 to 0.80)</td>
<td>0.58 (0.50 to 0.65)</td>
<td>0.92 (0.88 to 0.97)</td>
<td>0.73 (0.68 to 0.78)</td>
<td>0.71 (0.64 to 0.78)</td>
</tr>
<tr>
<td>Fine-tuning plus test-time augmentation</td>
<td>0.78 (0.73 to 0.82)</td>
<td>0.76 (0.70 to 0.82)</td>
<td>0.80 (0.72 to 0.86)</td>
<td>0.80 (0.74 to 0.83)</td>
<td>0.79 (0.74 to 0.84)</td>
</tr>
</tbody>
</table>

AUC, area under the curve; NTUH-BIO, National Taiwan University Hospital Hsin-Chu Branch, Biomedical Park Hospital; NTUH-HC, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu Hospital; NTUH-TPE, National Taiwan University Hospital.
conducted a comparison of the diagnostic performance of the CNN model when applied to rEBUS images acquired from two different manufacturers. Overall, the accuracy obtained using the conventional method (Olympus vs Fujifilm, 24/33 (0.73) vs 43/59 (0.73), p=0.0946) was similar to that obtained using fine-tuning plus the TTA method (Olympus vs Fujifilm, 27/33 (0.82) vs 50/59 (0.85), p=0.983).

Gradient-weighted class activation mapping (Grad-CAM) uses class-specific gradient information to produce a localisation map highlighting important regions. In this study, Grad-CAM was used to gain a detailed understanding of the classification model. Online supplemental figure 4 shows some examples of model visualisation.

Lung cancer subtyping analysis

We also assessed the diagnostic performance of the proposed CNN in identifying subtypes of lung cancer. The overall number of patients with definite pathological reports of lung cancer subtypes was limited, particularly in the NTUH-BIO cohort. Thus, lung cancer subtyping analysis was performed using cross-validation based on the NTUH-HC and NTUH-TPE cohorts. The performance of this model in identifying lung cancer subtypes was indicated by AUC values, as follows: adenocarcinoma (0.70 (95% CI 0.64 to 0.76)) and squamous cell carcinoma (0.64 (95% CI 0.54 to 0.74)). The performance of the model in identifying small cell carcinoma was less satisfactory, as indicated by an AUC of 0.52 (95% CI 0.52 to 0.72). Online supplemental figure 5 presents the receiver operating characteristic curves indicating the performance of the model in identifying subtypes of lung cancer versus pathological results as a reference standard. The diagnostic performance of the model in subtyping lung cancer is detailed in online supplemental table 5.

DISCUSSION

Our objective in this study was to develop a CNN model for the classification of benign and malignant lesions in rEBUS images. The proposed model demonstrated favourable performance when applied to an internal validation cohort, and satisfactory performance when applied to two independent test cohorts. We observed a decline in performance when applied to the external test sets; however, the negative effects were mitigated by TTA and fine-tuning. Note that this is the first study of its kind to include independent cohorts for external validation and to assess the feasibility of a CNN in identifying lung cancer subtypes.

rEBUS is a valuable tool for localising lung nodules in patients undergoing TBB; however, relatively few studies have investigated its clinical applicability in practice. Chao et al conducted a study using various image features to distinguish between neoplastic and non-neoplastic lesions, such as the margin outside the lesion, homogeneity among internal echoes, hyperechoic dots and concentric circles along the echo probe. Some of these image features were identified as diagnostic markers; however, the interpretation of images remains highly subjective.25 There is a pressing need for an objective method to interpret rEBUS images. This study demonstrated the efficacy of the proposed CNN prediction model in differentiating between malignant and benign lesions during bronchoscopic biopsy.

Numerous researchers have investigated the use of deep learning for the interpretation of medical ultrasound images;26-28 however, there has been very little work on the application of this technology to rEBUS images. Chen et al applied CNN with transfer learning to 164 rEBUS images from 164 patients. Their results (AUC=0.8705, accuracy=85.4% and specificity=82.1%) were similar to the results obtained using the internal validation cohort in the current study.18 Note however that they selected only one rEBUS image for each patient (ie, rather than including all recorded images), which raises serious concerns pertaining to selection bias. Note also that they enrolled patients from only one hospital (ie, no external validation), which may limit the generalisability of their results. Hotta et al used EBUS data from 213 participants to train a CNN algorithm, which achieved accuracy of 83.4%, sensitivity of 95.3% and specificity of 53.6% in differentiating benign from malignant lung lesions.29 Their results provide further support for our assertion that CNN models could be used to differentiate between benign and malignant lesions based on rEBUS images.

One of the major strengths of our study was the enrolment of patients from three different hospitals, using...
different rEBUS probes, and different operating physicians. We also used all recorded images to minimise selection bias. We assessed our CNN model using images generated by EBUS devices from two different manufacturers (Olympus and Fujifilm). We also included in our analysis two different external validation cohorts. Taken together, our results can be considered highly robust and generalisable to real-world clinical settings.

This study assessed a variety of machine learning techniques, including supervised learning, unsupervised learning and reinforcement learning. CNN is an unsupervised learning technique well suited to image classification. CNNs have been applied with considerable success in various medical imaging applications, such as mammography for breast cancer and spine X-rays for scoliosis. They have also been used for outcome prediction in radiation dose planning as well as in the interpretation of serial CT images to assess the response to treatments for lung cancer. CNNs have been shown to achieve accuracy comparable to or even surpassing that of human experts in many studies. However, the repeated use of rEBUS probes can introduce speckle noise, which can interfere with the machine learning process. To address this issue, we developed a denoising technique to reduce the impact of noise. We also employed image augmentation to compensate for imbalances in the training cohort.

It was observed during external validation that the discrimination performance of conventional CNN analysis at NTUH-TPE and NTUH-BIO was lower than that at NTUH-HC, where internal validation was also performed. This discrepancy can perhaps be explained by variations in rEBUS probes and image processors across institutions. In the current study, we addressed this challenge by implementing fine-tuning and TTA. Fine-tuning was performed using 10% of the data from the external validation cohort, which involved selecting images classified at NTUH-TPE (malignant (n=15) and benign (n=15)) and NTUH-BIO (malignant (n=5) and benign (n=5)). Note that the improvement in discrimination performance obtained using TTA was similar to that of fine-tuning. This suggests that in situations where it is not feasible to include rEBUS images from different image processors, TTA could serve as an alternative approach to bridging the performance gap between internal and external validation. Note also that implementing fine-tuning in conjunction with TTA could further enhance discrimination performance.

The TTA method used in this study involves the use of a classifier to make predictions based on multiple augmented test images and determining the final diagnostic result through voting. This approach closely resembles the decision-making process of clinicians, in which a decision is made only after inspecting an image carefully by zooming or rotating it back and forth. Previous studies have reported that TTA can significantly improve prediction performance by helping the classifier to detect objects that might otherwise be missed in the original image. TTA proved to be a valuable technique, leading to superior diagnostic performance compared with conventional methods when applied to external validation cohorts. Scaling augmentation, in particular, was found to enhance diagnostic performance by mitigating the impact of image variations arising from the use of different rEBUS equipment across different institutions.

Histological subtyping in lung cancer plays a crucial role in various aspects of patient management, including molecular testing, treatment planning and prognosis assessment. The histological classification of malignancies using CT or MRI has previously been investigated; however, few researchers have applied machine learning to ultrasound images for the subtyping of malignancies. In the current study, we extended the applicability of the proposed model to the differentiation of lung cancer subtypes. Note however that the diagnostic performance of the model (as indicated by AUC) was not satisfactory. Several factors may have contributed to this less-than-optimal performance. First, detecting subtle differences at the cellular or histological level solely from ultrasound images can be challenging. It should also be noted that some subtypes of lung cancer often share similar echo-textural characteristics, making them even more difficult to differentiate. It is possible that this overlap in imaging features contributed to the relatively low diagnostic performance of the CNN in this study. Further studies using larger datasets of higher quality will be required to fully explore the potential of AI in this type of application.

This study has several limitations that could impact the generalisability of our findings. First, the static rEBUS images were recorded by multiple bronchoscopists, resulting in inconsistent image quality. Second, the rEBUS images were linked to corresponding pathology reports of TBB, despite the fact that in clinical practice, definite results related to lung lesions are not always available (ie, biopsy yield is not 100%). This could have introduced discrepancies between the rEBUS images and the pathology results. However, we made efforts to minimise these effects by ensuring that the images were obtained by experienced bronchoscopists and that an average of four to six biopsy specimens were obtained. We also conducted a clinical follow-up of tumours identified as benign for at least 6 months after rEBUS. In this study, rEBUS-TBB analysis did not detect any benign tumours (eg, hamartomas), due perhaps to the fact that clinicians adopted diagnostic modalities other than rEBUS-TBB when dealing with patients suspected of having benign lesions based on chest CT scans. Lastly, this was a retrospective study in which the brightness and contrast of the rEBUS images were not standardised. However, we attempted to mitigate this variance by adjusting the brightness and contrast through augmentation during the training process. Furthermore, the variance could be considered a strength of the study, as the model demonstrated good diagnostic performance despite variations in parameter settings.
CONCLUSION
Our objective in this study was to assess the feasibility of using the EfficientNet CNN model to differentiate between malignant and benign lesions in rEBUS images. The dataset comprised rEBUS images obtained from three different sites, with one site dedicated to training and internal validation and the remaining two sites used for external validation. We also employed TTA and fine-tuning techniques. The results demonstrated the potential of the CNN model as adjunct method for the interpretation of rEBUS images.

Author affiliations
1Department of Internal Medicine, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu, Taiwan
2Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, Taipei, Taiwan
3Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
4Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

Correction notice
The article has been corrected since it was published online. The co-author Chun-Ta Huang’s name was misspelled as Chun-Da Huang; this has been amended.

Contributors
K-LY, P-CK and M-RL contributed to the study concept and design. K-LY, H-CY, C-JL, M-RL and C-TH conducted data collection. Y-ST performed data preprocessing, model construction, model evaluation and data visualisation. K-LY, Y-ST, P-CCK and M-RL contributed to data analysis and interpretation, and the writing of the manuscript. K-LY, Y-ST, H-CY, C-JL, P-CK, M-RL, C-TH, L-CH, K-JY, C-CH, J-YS, C-JY reviewed and edited the manuscript. All authors approved the final manuscript. M-RL is the author acting as guarantor for this work.

Funding
This study was funded by National Taiwan University Hospital Hsin-Chu Branch, Taiwan (111-HCH095) and National Tsing Hua University, Taiwan (111F7MDKE1).

Competing interests
None declared.

Patient and public involvement
Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication
Consent obtained directly from patient(s).

Ethics approval
This study was approved by the Institutional Review Board of the National Taiwan University Hospital (202109041RNB) and National Taiwan University Hospital Hsin-Chu Branch (110-131-E). This study was retrospective in nature and did not require participants to additional risk, such that the need for informed consent was waived.

Provenance and peer review
Not commissioned; externally peer reviewed.

Data availability statement
Data are available on reasonable request.

Supplemental material
This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access
This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

REFERENCES


37 Karaçiğlů O, Bilge HŞ, Uluer I. Removal of speckle noises from ultrasound images using five different deep learning networks.