

# Physician-Guided Learning with Attention Mechanism for Non-Tuberculosis Mycobacterium Disease Classification Using 3D Chest CT Images

Yueh-Chun Liu<sup>1</sup>  
 Chia-Jung Liu<sup>2</sup>  
 Yu-Hsuan Chen<sup>3,4</sup>  
 Chang-Wei Wu<sup>2</sup>  
 Meng-Rui Lee<sup>2,4</sup>  
 Po-Chih Kuo<sup>1</sup>

EUGENELIU1998@GAPP.NTHU.EDU.TW  
 M10082100@GMAIL.COM  
 C2724201@HOTMAIL.COM  
 SINCEADIDAS@HOTMAIL.COM  
 LEEMR@NTU.EDU.TW  
 KUOPC@CS.NTHU.EDU.TW

<sup>1</sup> Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

<sup>2</sup> Department of Internal Medicine, National Taiwan University, Hsin-Chu branch, Hsin-Chu, Taiwan

<sup>3</sup> Department of Critical Care Medicine, Min-Sheng General Hospital, Taoyuan, Taiwan

<sup>4</sup> Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

**Editors:** Under Review for MIDL 2026

## Abstract

The growing demand for interpreting Computed Tomography (CT) scans has driven the adoption of deep learning techniques, particularly Convolutional Neural Networks (CNNs), to assist physicians in managing their increasing workload. However, CNNs typically require large, well-annotated datasets to learn clinically meaningful imaging patterns, and such datasets remain limited due to the volumetric and resource-intensive nature of CT imaging. Incorporating physician expertise into the training process offers a promising path to overcoming this data limitation. In this study, we introduce a Human-in-the-Loop (HITL) workflow that enables physicians to directly guide CNN training by (1) providing feedback on model-generated attention maps and (2) validating or correcting predicted labels. This interactive process integrates domain knowledge into both representation learning and model interpretability. We evaluate the proposed HITL workflow on a chest CT dataset of patients with Non-tuberculous Mycobacteria (NTM) infection, a clinically challenging pulmonary disease that often requires long-term imaging follow-up to assess disease progression. Experimental results demonstrate that the HITL approach improves disease progression classification and produces clinically meaningful attention patterns, highlighting the value of physician-guided learning in medical imaging AI.

**Keywords:** Computed Tomography, Human-in-the-Loop, Expert-Guided Learning, Attention Mechanism

## 1. Introduction

Computed Tomography (CT) scans provide rich volumetric information for evaluating pulmonary abnormalities, but their interpretation is time-consuming and requires considerable expertise. Deep learning methods, particularly 3D Convolutional Neural Networks (CNNs), have shown promise in assisting this process; however, their performance is often

constrained by the scarcity of large, well-annotated datasets and the variability of physician assessments (Singh et al., 2020; Thanoon et al., 2023). These challenges are amplified in 3D chest CT analysis, where the large spatial complexity of volumetric data and differences in clinical judgment among physicians can lead to inconsistent or ambiguous labels. Such ambiguity is especially pronounced in certain pulmonary diseases—including Non-tuberculous Mycobacteria (NTM) infection—making it difficult for models to learn stable, clinically meaningful representations. Improving both model generalizability and interpretability for these challenging CT datasets therefore remains a critical need.

NTM are environmental organisms found in natural and treated water sources and can cause chronic pulmonary infections in humans. Prior studies report substantial 5-year all-cause mortality rates (13%-45%) among patients with NTM pulmonary disease (Novosad et al., 2017). Early identification of high-risk patients is thus essential for timely treatment and improved outcomes. However, the standard diagnostic workflow relies on microbiological species identification and culture testing, which are inherently slow and may delay clinical decision-making (Griffith et al., 2007). To mitigate this delay, multiple physicians often provide a preliminary consensus-based evaluation of suspected NTM cases. Although faster, this approach is inherently subjective and less standardized than formal diagnostic criteria (Kwak et al., 2016), leading to inconsistent assessments. Clinically, these initial evaluations typically categorize patients into *Colonization* or *Disease* groups, reflecting lower or higher likelihood of true NTM infection (Catanzaro, 2002; Van Ingen et al., 2018). The resulting ambiguity in labels becomes a significant obstacle for training reliable deep learning models.

Integrating physician knowledge directly into the model training process provides a promising strategy to address these challenges. Attention mechanisms can help 3D CNNs focus on clinically relevant lung regions (Liu et al., 2023; Almahasneh et al., 2025), while Human-in-the-Loop (HITL) frameworks allow physicians to interact with the model during training by refining predictions, validating ambiguous cases, and guiding the model’s attention patterns. Such expert-driven feedback has the potential to improve both predictive performance and clinical interpretability, especially in datasets where label ambiguity plays a critical role.

In this study, we first propose a 3D CNN with an attention mechanism that integrates 3D chest CT scans with patient demographic and microbiological data for predicting the likelihood of NTM pulmonary infection. We further introduce a **physician-guided HITL workflow** in which clinicians review model predictions and attention maps, re-examine ambiguous cases, and provide feedback used to refine labels and guide attention during fine-tuning—enhancing interpretability and improving model behavior.

## 2. Related Work

### 2.1. Attention Mechanism for 3D Chest CT

Deep learning techniques, particularly CNNs, have achieved remarkable success in 2D medical imaging (Gu et al., 2019; Li et al., 2019b; Rai and Chatterjee, 2021). However, applying 2D CNNs with slice-based approaches (Grewal et al., 2018; Zhang et al., 2020; Yang et al., 2021) directly to volumetric modalities such as chest CT scans often leads to the loss of

depth-dependent information, motivating the transition to 3D CNN architectures that better preserve spatial continuity.

To address this challenge, Zunair et al. (Zunair et al., 2020) introduced a modified 17-layer 3D CNN, along with two depth-resampling methods: Subset Slice Selection (SSS) (Zunair et al., 2019) and Spline Interpolated Zoom (SIZ). While SSS extracts a fixed subset of slices, SIZ performs spline-based interpolation along the z-axis to preserve more of the volumetric structure. Comparative analyses show that SIZ maintains far richer semantic continuity and yields superior classification performance under fixed computational budgets.

In parallel, attention mechanisms have emerged as powerful tools for guiding CNNs toward clinically meaningful regions in medical images (Li et al., 2019a; Suman et al., 2021; Rahman and Marculescu, 2023). Traditional ROI-based methods—such as blacking out background pixels (Yang et al., 2018) or cropping region-of-interest patches (Huynh et al., 2023)—often discard important contextual cues. Attention-based approaches overcome this limitation by incorporating ROI masks or attention maps as auxiliary inputs to the network (Li et al., 2020). Prior work demonstrates that introducing the attention map at the earliest convolutional layers provides the most significant performance gains, whereas inserting it at deeper layers or generating multiple attention maps offers limited or no improvement. Furthermore, the choice between additive and multiplicative merging operations has minimal effect, though additive fusion occasionally shows slight advantages (Eppel, 2018).

## 2.2. Human-in-the-Loop (HITL)

HITL (Monarch, 2021) has emerged as a powerful paradigm for developing machine learning systems that are not only accurate but also adaptive, interpretable, and aligned with domain expertise. Traditional machine learning pipelines rely heavily on expert involvement during modeling, implementation, and evaluation, yet these conventional workflows often struggle to maintain performance under data drift and frequently lack mechanisms for explaining model decisions. These limitations have motivated a shift toward HITL on machine learning, which integrates human expertise directly and iteratively into the learning loop.

The rise of Explainable AI (XAI) (Adadi and Berrada, 2018) further strengthens the importance of HITL, especially in domains requiring trustworthy and interpretable decisions. XAI methods help clarify the reasoning behind model outputs, enabling developers to identify sources of error, validate model logic, and extract insights from models that outperform human-level performance. When combined with XAI, HITL creates a collaborative workflow in which humans and machines complement each other—leveraging human knowledge alongside computational scalability and pattern-recognition strengths.

## 3. Dataset

We utilize a private NTM dataset (Liu et al., 2025) to train and evaluate our proposed method. The dataset comprises 3D high-resolution chest CT (HRCT) scans, along with patient demographic information (age and gender) and microbiological test results relevant to NTM infection: the acid-fast bacillus smear test (AFS) and the acid-fast bacillus culture test (culture).

The dataset is divided into four subsets: 278 unique patients for training, 67 for validation, 55 for internal testing, and 196 for external testing. Three senior physicians provided initial consensus-based assessments for each patient. These assessments were converted into labels from 0 (*Colonization*) to 1 (*Disease*), weighted according to the level of consensus among the physicians.

## 4. Method

When evaluating patients suspected of NTM infection, physicians typically review 3D chest CT scans together with demographic information and microbiological test results (Daley et al., 2020). They visually inspect the CT volumes for abnormal regions and integrate these findings with their clinical experience to form an initial assessment of infection likelihood. The degree of agreement among physicians in these assessments reflects the certainty of the diagnosis: lower consensus often indicates greater ambiguity in imaging or clinical information, leading physicians to rely more heavily on subjective judgment. For such low-consensus cases, a structured review process allows physicians to re-examine suspicious regions more carefully and consistently, thereby reducing inter-physician variability and improving the reliability of the initial assessments. We draw inspiration from this structured clinical review process to design our proposed method, which incorporates physician feedback to guide model refinement and improve interpretability.

### 4.1. 3D CNN with an attention mechanism

We propose a 3D CNN with an attention mechanism that helps the model learn to attend to clinically meaningful visual features when generating predictions. Figure 1 presents an overview of the proposed 3D CNN-based model architecture used for NTM infection classification. Our proposed model is a mixture model consisting of two sub-models: (1) a 3D CNN with an attention mechanism for extracting visual features from chest CT scans, and (2) a multi-layer perceptron (MLP) for extracting clinical features from patient demographics and microbiological test results.

For the CNN sub-model, each chest CT scan is paired with its corresponding attention map highlighting regions that the model should focus on. Both the CT scan and its attention map are first processed through separate convolutional layers in parallel for generating their feature maps, respectively, and then merged into a single feature map representing high-level attended CT features. This merged feature map is further processed through the remaining CNN layers to generate visual features for NTM infection likelihood prediction. For the MLP sub-model, fully connected linear layers are applied to compress the patient’s non-imaging data demographics and microbiological test results into clinical features. Finally, the visual features from the CNN and the clinical features from the MLP are concatenated and fed into a classifier for binary prediction (*Colonization* vs *Disease*) of NTM infection.

### 4.2. HITL Process

To further support decision-making in low-consensus cases, we introduce a **physician-guided HITL workflow**. In this process, physicians interact with the model by reviewing

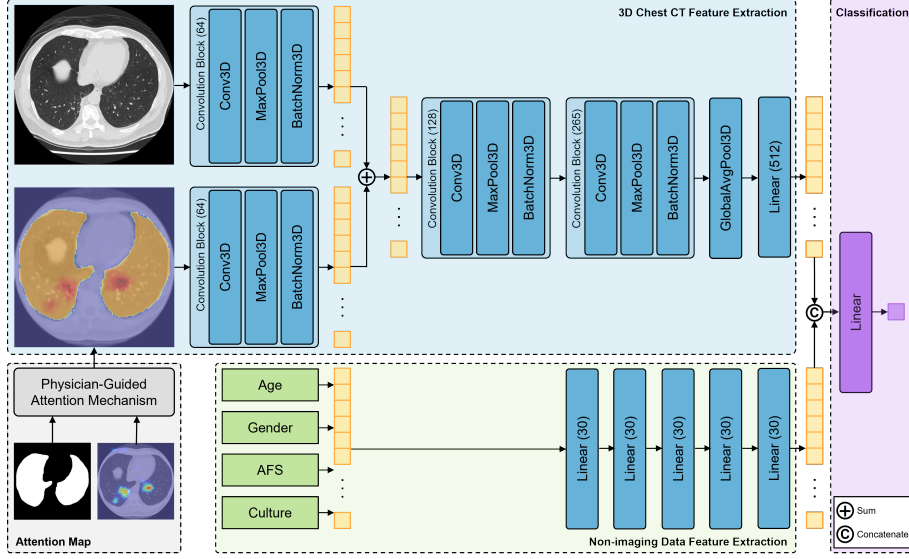


Figure 1: The proposed attention-based model for predicting the NTM infection likelihood. Physician-guided attention is incorporated to refine feature learning and enhance clinical relevance during fine-tuning in HITL process.

predicted NTM infection likelihoods and attention heatmaps highlighting relevant CT regions. They provide feedback on these outputs and record revised assessments. Both the feedback and updated assessments are then used to fine-tune the model, enabling it to adjust its decision-making and reasoning patterns based on expert advisement.

Figure 2 illustrates the proposed workflow of HITL process. In each round, physicians review the model’s NTM infection likelihood predictions and corresponding attention heatmaps on CT volumes, provide re-examined assessments, and offer feedback on the usefulness of the model outputs. This feedback loop allows the model to gradually adjust its decision patterns and align more closely with the expert reasoning.

#### 4.2.1. HITL PROCESS - PREPARATION (ROUND 0)

Before the iterative HITL rounds begin, we prepare the data and train an initial base model.

- Data Preparation** Since only low-consensus cases require physician re-examination, full-consensus and low-consensus cases are partitioned differently. Full-consensus cases are directly divided into training, validation, and testing sets. Low-consensus cases are split into  $R+1$  folds, and these folds are sequentially incorporated into the training set as the HITL process progresses.
- Base Model Training** The base model  $M^0$  is trained using the data combination of the training set of full-consensus cases and the  $0^{th}$  fold of low-consensus cases. The initial physician assessments  $s^0$  are averaged to obtain the initial label  $l^0$ . The lung segmentation mask generated using a pre-trained BCDU-Net (Azad et al., 2019) is used as the initial attention map  $a^0$  guiding the CNN to prioritize lung regions where NTM-related features may appear (Daley et al., 2020).

#### 4.2.2. HITL PROCESS

After Round 0, the iterative HITL process begins. Each round  $r$  involves: (1) training data collection, (2) previous model inference, (3) physician re-examination, and (4) model fine-tuning. A total of  $N$  physicians participate in every re-examination round.

For each round  $r$ , the steps proceed as follows:

- **Step 1: Training Data Collection**

The training data includes the full-consensus training set and the  $r^{th}$  fold of low-consensus cases. Each sample contains a 3D chest CT scans  $x_{CT}$  and its corresponding binary lung segmentation mask  $x_{lungseg}$ .

- **Step 2: Previous Model Inference**

Using the model  $M^{r-1}$  from the previous round we generate: (1) the predicted NTM infection likelihood  $p^r$ , and (2) the Grad-CAM (Selvaraju et al., 2017)–based attention heatmap  $h^r$  on CT volumes using data in  $r^{th}$  fold of low-consensus cases.

- **Step 3: Physician Re-examination**

Each physician  $i$  reviews the predictions  $p^r$  and heatmaps  $h^r$  then provides: (1) a re-examined assessment  $s_i^r$ , and (2) a binary feedback indicator  $f_i^r$ , specifying whether the model outputs were helpful.

The re-examined assessments are considered more reliable than the original ones because they incorporate physicians’ reflections informed by the model. Thus, the new label  $l^r$  for fine-tuning is the averaged only on re-examined assessments  $s_i^r$ :

$$l^r = \frac{1}{N} \sum_{i=1}^N s_i^r \quad (1)$$

- **Step 4: Model Fine-tuning**

- **Step 4-1: Feedback-weighted Attention Map Construction**

Positive feedback indicates that physicians find the model’s output helpful, suggesting that the previous attention heatmap is likely highlighting regions that align with their diagnostic focus. In this case, the model should be encouraged to attend more strongly to those regions. Conversely, negative feedback suggests that the highlighted areas do not match the physicians’ clinical priorities, and attention to those regions should be diminished.

Physician feedback is converted into numeric values:

$$\hat{f}_i^r = \begin{cases} +1 & \text{if feedback } f_i^r \text{ is positive} \\ -1 & \text{if feedback } f_i^r \text{ is negative} \end{cases} \quad (2)$$

Then, the overall feedback weight can be defined as:

$$W_{feedback}^r = \frac{1}{N} \sum_{i=1}^N \hat{f}_i^r \quad (3)$$

The new attention map  $a^r$  combines the lung segmentation mask and the feedback-adjusted Grad-CAM heatmap:

$$\begin{aligned} a_{lungseg}^r &= W_{lungseg} \times x_{lungseg} \\ a_{heatmap}^r &= W_{heatmap} \times W_{feedback}^r \times h^r \\ a^r &= a_{lungseg}^r + a_{heatmap}^r \end{aligned} \quad (4)$$

Positive feedback strengthens the influence of the model’s previous attention; negative feedback attenuates it.

– **Step 4-2: Case-level Attention in the Loss Function**

To further emphasize uncertain cases, we apply higher case-specific weights on cases which are low-consensus at previous round in the binary cross-entropy loss based on how inconsistent the previous model’s predictions  $p_{r-1}$  were with the new re-examined label  $l_r$ :

$$W_{BCELoss} = \begin{cases} 1 & \text{if } l_{r-1} \text{ is full-consensus} \\ 1 + (p_{r-1} - l_r) & \text{if } l_{r-1} \text{ is low-consensus} \end{cases} \quad (5)$$

The HITL process concludes once all  $R + 1$  folds of low-consensus data have been processed.

After completing the HITL procedure, each participating physician is also asked to fill out a quantitative questionnaire evaluating the overall usefulness and clinical reasonableness of the model’s predicted NTM infection likelihood and its corresponding attention heatmaps on the 3D chest CT volumes. These responses are used to assess how effectively the HITL workflow supports physicians during the re-examination of uncertain NTM cases. The questionnaire design, the physicians’ responses collected during the HITL process, and the corresponding analysis incorporating their re-examination results are provided in Appendix B.

## 5. Implementation Details

The base model  $M^0$  was trained for 100 epochs, and in each subsequent HITL round, the model  $M^r$  was fine-tuned for 50 epochs. The base model was initialized with random weights, whereas each fine-tuned model inherited the weights from the model obtained in the previous round. All models—including both the base model and the fine-tuned models—were trained in a fully unfrozen setting, with no layers held fixed.

A batch size of 8 was used for all experiments. Optimization was performed using the Adam optimizer (Kingma, 2014) with a learning rate of  $10^{-2}$ . All training procedures were conducted on a GPU equipped with 24 GB of memory.

Figure A.1 summarizes the number of cases in each data split for both the full-consensus and low-consensus groups. In each HITL round, 3 senior physicians are invited to participate in the re-examination process. Both  $W_{lungseg}$  and  $W_{heatmap}$  are set to 1, assigning equal importance to the lung segmentation mask and the model-generated attention heatmap during model fine-tuning. The attention map  $a^r$  is then normalized to the range  $[0, 1]$ .

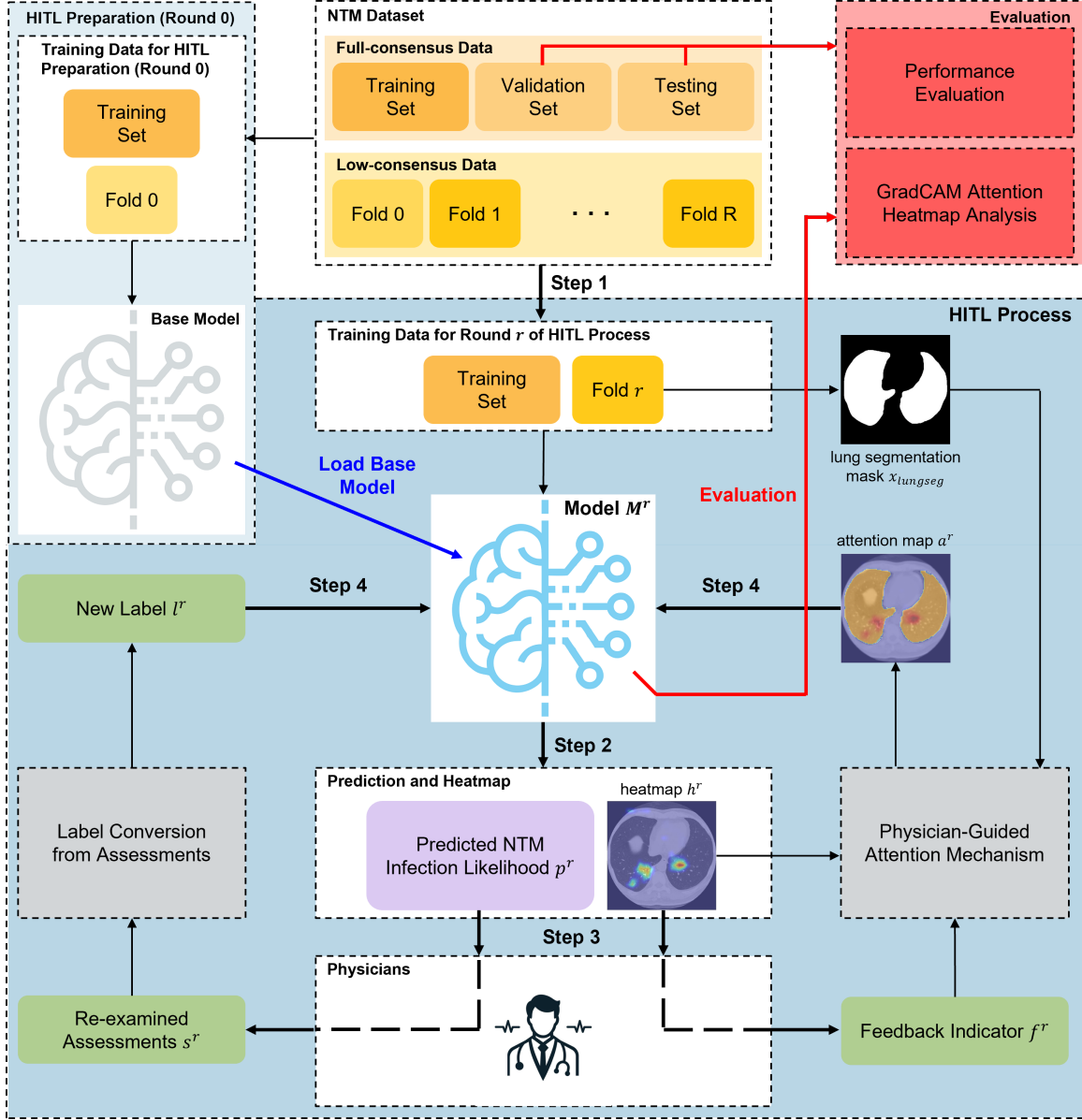


Figure 2: Workflow of the proposed physician-guided attention HITL process. In each HITL round, senior physicians review the model’s predictions and Grad-CAM heatmaps, re-examine for new assessments, and provide feedback indicating model usefulness. These physician-guided annotations are then used to update both the class labels and the attention map, which are incorporated into fine-tuning for the next HITL round.

## 6. Experiment Results

### 6.1. Quantitative Results

As shown in Table 1, on the internal testing set, our model attains an AUROC of 0.9163 ( $+\Delta 0.0380$ ) relative to the baseline. On the external testing set, it achieves an AUROC of 0.8003 ( $+\Delta 0.0784$ ), accuracy of 0.7245 ( $+\Delta 0.0204$ ), sensitivity of 0.7558 ( $+\Delta 0.0930$ ), specificity of 0.7000 ( $-\Delta 0.0364$ ), F1-score of 0.7065 ( $+\Delta 0.0437$ ), and Youden’s index of 0.4558 ( $+\Delta 0.0566$ ). These results demonstrate that the proposed HITL process improves the model’s generalizability to external data while preserving performance on the original data distribution.

Table 1: Evaluation on the internal and external testing sets for models trained without (w/o) and with (w/) the HITL process. The 95% confidence intervals are reported. **Best** results are in bold.

	Metrics	Internal Testing	External Testing
w/o HITL training (baseline)	AUROC	0.8783 (0.8142–0.9337)	0.7219 (0.6439–0.7950)
	Accuracy	<b>0.8167 (0.7500–0.8833)</b>	0.7041 (0.6327–0.7653)
	Sensitivity	<b>0.7143 (0.6429–0.8276)</b>	0.6628 (0.5732–0.7191)
	Specificity	<b>0.9062 (0.8065–1.0000)</b>	<b>0.7364 (0.6415–0.8214)</b>
	F1-score	<b>0.7843 (0.7273–0.8627)</b>	0.6628 (0.5867–0.7356)
	Youden’s index	<b>0.6205 (0.4961–0.7545)</b>	0.3992 (0.2637–0.5224)
w/ HITL training	AUROC	<b>0.9163 (0.8646–0.9554)</b>	<b>0.8003 (0.7568–0.8774)</b>
	Accuracy	0.7833 (0.7167–0.8500)	<b>0.7245 (0.6837–0.7653)</b>
	Sensitivity	0.6429 (0.5172–0.7857)	<b>0.7558 (0.6757–0.8193)</b>
	Specificity	<b>0.9062 (0.7812–0.9355)</b>	0.7000 (0.5981–0.7642)
	F1-score	0.7347 (0.6522–0.8302)	<b>0.7065 (0.6329–0.7500)</b>
	Youden’s index	0.5491 (0.4333–0.6920)	<b>0.4558 (0.3846–0.5449)</b>

### 6.2. Ablation Studies

To assess the contribution of the proposed physician-guided attention mechanism within the HITL framework, we conduct a series of ablation experiments by systematically enabling or disabling 4 key components of the attention design. Specifically, we examine the model’s performance under the following conditions: (1) **Normalization (Norm.)**: whether the attention map is normalized to the range  $[0, 1]$ ; (2) **Convert as Binary (Binary.)**: whether the attention map is converted into a binary mask (analogous to the representation of the lung segmentation mask); (3) **Freeze Convolutional Layers (Freeze.)**: whether the convolutional layers are frozen during fine-tuning; and (4) **Weights of Attentions (Weights.)**: how varying the values of  $W_{\text{lungseg}}$  and  $W_{\text{heatmap}}$ , thereby modulating the relative importance of the lung segmentation mask and the Grad-CAM-derived heatmap, influences overall model performance.

### 6.3. Qualitative Results

Figure 3 and Figure C.1 illustrates the model’s attention heatmaps on 3D chest CT scans generated using Grad-CAM. The results show that the heatmaps shift not only in response to physicians’ revised assessments during re-examination but also according to their feedback regarding the usefulness and reasonableness of the model’s outputs. This demonstrates

Table 2: Ablation study evaluating the impact of different attention mechanism settings. For each group, the best performance is highlighted with a gray while the Best among best of each group is highlighted with a deep gray .

Experiment Settings				Performance Evaluation			
Norm.	Binary.	Freeze.	Weight. ( $W_{\text{lungseg}}, W_{\text{heatmap}}$ )	Internal Testing		External Testing	
				AUROC	Accuracy	AUROC	Accuracy
✓	x	✓	(1, 1)	0.8739	0.8333	0.8074	0.7398
			(2, 1)	0.8717	0.7500	0.7929	0.6684
			(1, 2)	0.8571	0.8000	0.7551	0.6990
✓	x	x	(1, 1)	0.9163	0.7833	0.8003	0.7245
			(2, 1)	0.8772	0.7833	0.7200	0.6531
			(1, 2)	0.8962	0.8000	0.8016	0.7245
x	x	✓	(1, 1)	0.8471	0.7667	0.7842	0.7194
			(2, 1)	0.8895	0.8333	0.7737	0.6684
			(1, 2)	0.8962	0.8500	0.7818	0.6786
x	x	x	(1, 1)	0.8438	0.7667	0.7087	0.6429
			(2, 1)	0.8549	0.7500	0.6790	0.6786
			(1, 2)	0.8996	0.8000	0.8098	0.7194
✓	✓	✓	(1, 1)	0.8850	0.7667	0.7870	0.7245
			(2, 1)	0.8650	0.7833	0.7289	0.6837
			(1, 2)	0.8839	0.7500	0.7197	0.7143
✓	✓	x	(1, 1)	0.8438	0.7833	0.7207	0.6939
			(2, 1)	0.8538	0.7667	0.6572	0.7041
			(1, 2)	0.8147	0.7167	0.5823	0.5714

that the proposed HITL process effectively enables the model to incorporate physician-guided supervision and adapt its attention accordingly.

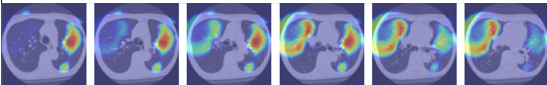
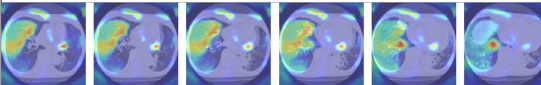
Before Applying HITL Process				After Applying HITL Process			
(a) Model's prediction remains unchanged, while GradCAM heatmap is updated in response to overall negative usefulness feedback from physicians.							
Label	Model's Prediction	Re-examination	Usefulness Feedback	→	Label	Model's Prediction	
Col (0.33)	Col (0.01)	Col / Col / Col	No / No / Yes		Col (0.00)	Col(0.02)	
GradCAM Heatmap					GradCAM Heatmap		
							

Figure 3: An example comparison of model predictions and Grad-CAM attention heatmaps before and after applying the HITL process.

## 7. Conclusion

In this work, we introduce a physician-guided HITL framework that integrates expert feedback into the attention mechanism of a 3D CNN for predicting NTM infection likelihood from chest CT scans and clinical data. Qualitative analysis further demonstrates that the model's Grad-CAM heatmaps adapt in response to physician re-examination and usefulness feedback to model's output, indicating that the HITL process effectively steers the model toward more clinically meaningful attention patterns.

## Acknowledgments

This work is supported by National Taiwan University Hospital Hsin Chu Branch - National Tsing Hua University Joint Research Program (NTUH HCH-NTHU Joint Research Program, No: 114QF017E1).

## References

- Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 2018.
- Majedaldein Almahasneh, Xianghua Xie, and Adeline Paiement. Attentnet: fully convolutional 3d attention for lung nodule detection. *SN Computer Science*, 2025.
- Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Bi-directional convlstm u-net with densely connected convolutions. *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.
- Antonino Catanzaro. Diagnosis, differentiating colonization, infection, and disease. *Clinics in chest medicine*, 2002.
- Charles L Daley, Jonathan M Iaccarino, Christoph Lange, Emmanuelle Cambau, Richard J Wallace Jr, Claire Andrejak, Erik C Böttger, Jan Brozek, David E Griffith, Lorenzo Guglielmetti, et al. Treatment of nontuberculous mycobacterial pulmonary disease: an official ats/ers/escmid/idsa clinical practice guideline. *Clinical Infectious Diseases*, 2020.
- Sagi Eppel. Classifying a specific image region using convolutional nets with an roi mask as input. *arXiv preprint*, 2018.
- Monika Grewal, Muktabh Mayank Srivastava, Pulkit Kumar, and Srikrishna Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. *IEEE International Symposium on Biomedical Imaging*, 2018.
- David E Griffith, Timothy Aksamit, Barbara A Brown-Elliott, Antonino Catanzaro, Charles Daley, Fred Gordin, Steven M Holland, Robert Horsburgh, Gwen Huitt, Michael F Iademarco, et al. An official ats/idsa statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. *American journal of respiratory and critical care medicine*, 2007.
- Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 2019.
- Hoang Nhut Huynh, Anh Tu Tran, and Trung Nghia Tran. Region-of-interest optimization for deep-learning-based breast cancer detection in mammograms. *Applied Sciences*, 2023.
- Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint*, 2014.

- Nakwon Kwak, Chang Hyun Lee, Hyun-Ju Lee, Young Ae Kang, Jae Ho Lee, Sung Koo Han, and Jae-Joon Yim. Non-tuberculous mycobacterial lung disease: diagnosis based on computed tomography of the chest. *European radiology*, 2016.
- Jun Li, Daoyu Lin, Yang Wang, Guangluan Xu, Yunyan Zhang, Chibiao Ding, and Yanhai Zhou. Deep discriminative representation learning with attention map for scene classification. *Remote Sensing*, 2020.
- Liu Li, Mai Xu, Hanruo Liu, Yang Li, Xiaofei Wang, Lai Jiang, Zulin Wang, Xiang Fan, and Ningli Wang. A large-scale database and a cnn model for attention-based glaucoma detection. *IEEE transactions on medical imaging*, 2019a.
- Shaohua Li, Yong Liu, Xiuchao Sui, Cheng Chen, Gabriel Tjio, Daniel Shu Wei Ting, and Rick Siow Mong Goh. Multi-instance multi-scale cnn for medical image classification. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019b.
- Chia-Jung Liu, Yueh-Chun Liu, Yu-Hsuan Chen, Yu-Sen Huang, Po-Chih Kuo, Meng-Rui Lee, Lu-Cheng Kuo, Jann-Yuan Wang, Chao-Chi Ho, Jin-Yuan Shih, et al. Ai-assisted differentiation of nontuberculous mycobacterial pulmonary disease from colonization: a multi-center study. *Insights into Imaging*, 2025.
- Gang Liu, Fei Liu, Jun Gu, Xu Mao, XiaoTing Xie, and Jingyao Sang. An attention-based deep learning network for lung nodule malignancy discrimination. *Frontiers in Neuroscience*, 2023.
- Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.
- Shannon A Novosad, Emily Henkle, Sean Schafer, Katrina Hedberg, Jennifer Ku, Sarah AR Siegel, Dongseok Choi, Christopher G Slatore, and Kevin L Winthrop. Mortality after respiratory isolation of nontuberculous mycobacteria. a comparison of patients who did and did not meet disease criteria. *Annals of the American Thoracic Society*, 2017.
- Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023.
- Hari Mohan Rai and Kalyan Chatterjee. 2d mri image analysis and brain tumor detection using deep learning cnn model leu-net. *Multimedia Tools and Applications*, 2021.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 2017.
- Satya P Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3d deep learning on medical images: a review. *Sensors*, 2020.

- Sudhir Suman, Gagandeep Singh, Nicole Sakla, Rishabh Gattu, Jeremy Green, Tej Phatak, Dimitris Samaras, and Prateek Prasanna. Attention based cnn-lstm network for pulmonary embolism prediction on chest computed tomography pulmonary angiograms. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.
- Mohammad A Thanoon, Mohd Asyraf Zulkifley, Muhammad Ammirul Atiqi Mohd Zainuri, and Siti Raihanah Abdani. A review of deep learning techniques for lung cancer screening and diagnosis based on ct images. *Diagnostics*, 2023.
- Jakko Van Ingen, Timothy Aksamit, Claire Andrejak, Erik C Böttger, Emmanuelle Cambau, Charles L Daley, David E Griffith, Lorenzo Guglielmetti, Steven M Holland, Gwen A Huitt, et al. Treatment outcome definitions in nontuberculous mycobacterial pulmonary disease: an ntm-net consensus statement. *European Respiratory Journal*, 2018.
- Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, and Bingbing Ni. Reinventing 2d convolutions for 3d images. *IEEE Journal of Biomedical and Health Informatics*, 2021.
- Yang Yang, Weiming Zhang, Dong Liang, and Nenghai Yu. A roi-based high capacity reversible data hiding scheme with contrast enhancement for medical images. *Multimedia Tools and Applications*, 2018.
- Jianpeng Zhang, Yutong Xie, Yan Wang, and Yong Xia. Inter-slice context residual learning for 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2020.
- Hasib Zunair, Aimon Rahman, and Nabeel Mohammed. Estimating severity from ct scans of tuberculosis patients using 3d convolutional nets and slice selection. *CLEF (Working Notes)*, 2019.
- Hasib Zunair, Aimon Rahman, Nabeel Mohammed, and Joseph Paul Cohen. Uniformizing techniques to process ct scans with 3d cnns for tuberculosis prediction. *International workshop on predictive intelligence in medicine*, 2020.

## Appendix A. Data Split of the Full-consensus and Low-consensus Subsets

Figure A.1 summarizes the distribution of cases across the training, validation, and testing splits for both the full-consensus and low-consensus subsets.

## Appendix B. Questionnaire Design, Results and Analysis for Model’s Usefulness and Reasonableness to Physicians in HITL Process

Table B.1 presents the questionnaire design, including both the case-level question and the overall questions used to assess the usefulness and reasonableness of the model outputs. The corresponding responses collected from participating physicians during the HITL process are summarized in Table B.2. Furthermore, Table B.3 provides an integrated analysis of the questionnaire results together with the physicians’ re-examination outcomes.

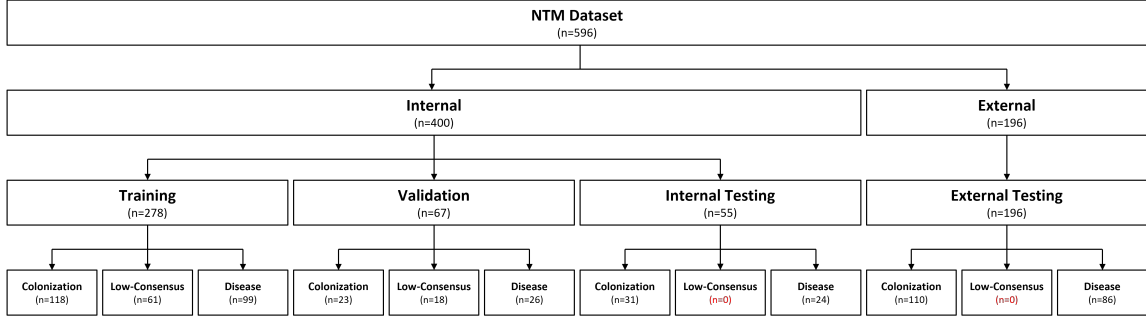


Figure A.1: Data split of the full-consensus and low-consensus subsets. Low-consensus samples in the training set are further partitioned into multiple folds for the HITL process. No low-consensus samples are included in the testing set to ensure an objective and unbiased performance evaluation.

Table B.1: Questionnaire evaluating for each case and the overall usefulness and clinical reasonableness of the model’s predictions and Grad-CAM heatmaps for physicians in a HITL round.

Case Question (C.Q.)	
C.Q.	Do you think the prediction and Grad-CAM heatmap generated by the model for this patient case is <b>useful</b> for re-examination?
Overall Question (O.Q.)	
O.Q.1	In general, do you think the <b>prediction</b> generated by the model in this HITL round is <b>reasonable</b> ?
O.Q.2	In general, do you think the <b>Grad-CAM heatmap</b> generated by the model in this HITL round is <b>useful</b> for re-examination?
O.Q.3	In general, do you think the provided model in this HITL round can <b>help doctors for re-examination</b> ?

Table B.2: The results to the questionnaire shown in Table B.1 for case question and overall questions.

	Physician ID	Case Question (C.Q.)	Overall Question (O.Q.)		
		(# of "Yes" over all)	O.Q.1	O.Q.2	O.Q.3
<b>HITL Round 1</b>	Physician 1	2 over 28 (7.14%)	Yes	No	No
	Physician 2	8 over 28 (28.57%)	Yes	No	No
	Physician 3	16 over 28 (57.14%)	Yes	No	Yes

Table B.3: Analysis of the results to the questionnaire with physician’s re-examination.

	Rate of the Same as Previous on Re-examination (for each physician)			Rate of Consensus on Re-examination	Rate of Consensus the Same as Model’s Predictions
	1	2	3		
<b>HITL Round 1</b>	42.86%	39.29%	71.43%	35.71%	28.57%

### Appendix C. Comparison of Model Predictions and Grad-CAM Attention Heatmaps Before and After Applying the HITL Process

Figure C.1 presents a detailed comparison of the model’s predictions and Grad-CAM attention heatmaps before and after applying the HITL process across several representative scenarios. These examples highlight how physician-guided feedback influences both the model’s decision outcomes and its attention behavior, demonstrating the effectiveness of our proposed HITL workflow in refining the model’s focus toward clinically meaningful regions.

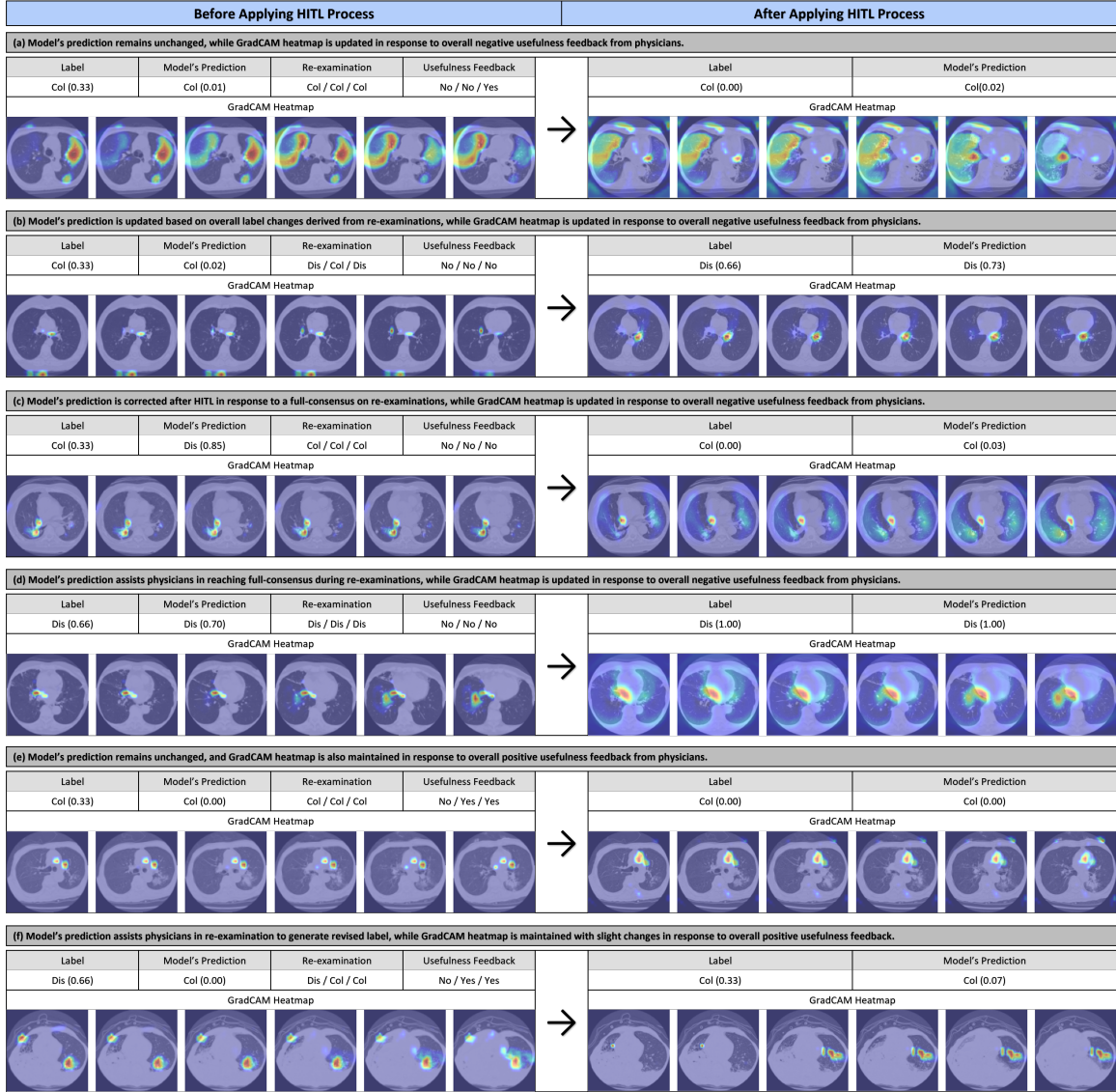


Figure C.1: Examples comparing model predictions and Grad-CAM attention heatmaps before and after applying the HITL process across different representative scenarios.