
RNNet-MST: A ResNet-50 with Multi-Scale Transformer Blocks for Pulmonary Nodule Classification and Attention-Based Localization on Chest X-Ray Images

[Edrill F. Bilan](#)[†], [Emman T. Manduriaga](#)[†], [Hernando S. Salapare III](#)^{*}, Ymir M. Garcia, Khatalyn E. Mata, [Rose Anna R. Banal](#), [Imelda C. Ang](#), [Wei-Ta Chu](#), [Dan Michael A. Cortez](#)^{*}

Posted Date: 10 April 2026

doi: 10.20944/preprints202604.0738.v1

Keywords: RNNet-MST; pulmonary nodule detection; pulmonary nodule classification; deep learning; ResNet-50; transformers; spatial attention mechanism; chest X-ray analysis; computer-aided detection; false negative reduction



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

RNNNet-MST: A ResNet-50 with Multi-Scale Transformer Blocks for Pulmonary Nodule Classification and Attention-Based Localization on Chest X-Ray Images

Edrill F. Bilan ^{1,†}, Emman T. Manduriaga ^{1,†}, Hernando S. Salapare III ^{2,3,*}, Ymir M. Garcia ^{4,5}, Khatalyn E. Mata ¹, Rose Anna R. Banal ⁶, Imelda C. Ang ⁶, Wei-Ta Chu ⁷ and Dan Michael A. Cortez ^{1,*}

¹ College of Information Systems and Technology, Pamantasan ng Lungsod ng Maynila, Manila, Philippines

² Institut de Science des Matériaux de Mulhouse (IS2M), CNRS, UHA, UMR 7361, Mulhouse 68057, France

³ University Research Center, Pamantasan ng Lungsod ng Maynila, Manila, Philippines

⁴ School of Chemical, Biological, Materials Engineering and Sciences, Mapúa University, Manila, Philippines

⁵ School of Health Sciences and Nursing, Mapúa University, Makati, Philippines

⁶ College of Medicine, Pamantasan ng Lungsod ng Maynila, Manila, Philippines

⁷ Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

* Correspondence: hissalaapare@plm.edu.ph (H.S.S.III); dmacortez@plm.edu.ph (D.M.A.C.)

† These authors contributed equally to this work.

Abstract

Background/Objectives: Lung cancer survival depends on early detection; however, in the Philippines, high radiologist workloads and the anatomical complexity of chest X-rays (CXRs) contribute to missed pulmonary nodules and false-negative diagnoses. This study aims to develop an enhanced deep learning model to improve nodule classification and localization sensitivity. **Methods:** We propose RNNNet-MST, an extension of ResNet-50 that incorporates Multi-Scale Transformer blocks for global context modeling and a custom spatial attention mechanism for attention-based weak localization of disease-relevant regions. The model was trained and evaluated on the NODE21 chest X-ray dataset and compared with a baseline ResNet-50 using classification metrics, with attention maps used for weak localization analysis. **Results:** RNNNet-MST demonstrated improved performance across evaluated metrics relative to the baseline model. Nodule Recall increased from 86.18% to 93.09% (+6.91%), reducing false negatives. Test Accuracy reached 95.16% (+0.51%), and the Nodule F1-Score improved to 91.40% (+1.50%), indicating better detection of small and subtle nodules. **Conclusions:** The integration of multi-scale transformer features improved classification sensitivity, while the attention mechanism provided weak localization cues that aligned more closely with annotated nodule regions than the baseline. RNNNet-MST shows potential as a diagnostic support tool, warranting further validation on larger and more diverse clinical datasets to reduce perceptual errors and facilitate early lung cancer detection in resource-constrained settings.

Keywords: RNNNet-MST; pulmonary nodule detection; pulmonary nodule classification; deep learning; ResNet-50; transformers; spatial attention mechanism; chest X-ray analysis; computer-aided detection; false negative reduction

1. Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, and its prognosis is highly dependent on early detection [1]. In the Philippine context, the burden of life-

threatening diseases continues to pose a major public health concern. Recent data indicates that neoplastic diseases rank among the top five leading causes of death nationwide [2]. This highlights an urgent need for improved diagnostic approaches, particularly for lung cancer, where early identification is crucial for improving survival outcomes.

Chest radiography (CXR) is the most widely used imaging modality in clinical practice due to its accessibility, low cost, and relatively low radiation dose compared to computed tomography (CT) [3]. CXR serves primarily as a first-line screening tool. However, detecting pulmonary nodules—small growths typically less than 3 cm—on CXRs is a highly challenging task.

Consequently, CXR interpretation is highly susceptible to diagnostic limitations and false-negative findings. A prospective study by Miki et al. demonstrated that radiologists significantly tend to miss lung nodules located in anatomically complex areas, such as the bilateral hilar regions, revealing inherent limitations in human visual search behavior [4]. Further evidence from Digumarthy et al. utilized simulation-based approaches to show that even with targeted training and education, substantial diagnostic blind spots persist in complex anatomical locations [5].

While CT offers superior spatial resolution and sensitivity for nodule detection, it is considerably more expensive, delivers a substantially higher radiation dose, and requires specialized infrastructure that is often unavailable in low- and middle-income settings [3]. As a consequence, the majority of high-performing deep learning Computer-Aided Detection (CAD) systems in the literature have been developed and evaluated primarily on CT or low-dose CT (LDCT) datasets [17,19]. CXR-based nodule detection has received comparatively less attention, despite CXR remaining the predominant—and in many rural or resource-constrained facilities, the only—available imaging modality [18]. This creates a significant translational gap: state-of-the-art CT-focused models are not clinically deployable in settings where CT infrastructure is absent, yet the CXR-focused literature reports considerably lower sensitivity benchmarks. Critically, the goal of a CXR-based CAD system in such environments is not to replace CT diagnosis, but to serve as a low-cost, computationally lightweight second opinion—flagging suspicious regions that a fatigued radiologist might overlook and thereby triaging patients who warrant further investigation.

To contextualize these limitations within current real-world diagnostic practices, the researchers conducted expert interviews with radiologists at the Lung Center of the Philippines (LCP). The interviews revealed that radiologists face a substantial daily workload, interpreting approximately 100 to 200 radiographic images alongside other imaging modalities. This intensity causes severe cognitive load and professional burnout, which are known risk factors for perceptual and interpretive errors. Furthermore, the experts noted a lack of locally developed medical imaging technologies tailored to these specific systemic constraints.

These insights reveal a critical diagnostic gap that contributes to delayed diagnoses. Given the widespread reliance on CXR screening, there is a vital need for intelligent, automated Computer-Aided Detection (CAD) systems. Deep convolutional neural networks, specifically ResNet-50, have demonstrated strong performance in extracting fine-grained features for medical imaging using residual learning [6]. While standard ResNet-50 performs well on binary classification, its limited ability to capture multi-scale features and model long-range dependencies reduces its sensitivity to small nodules [7]. Furthermore, standard deep vision models often lack the spatial attention required to isolate subtle, overlapping abnormalities [8].

To address these limitations, this study proposes RNNet-MST, an enhanced ResNet-50 architecture for pulmonary nodule classification on chest X-rays, with attention-based weak localization used to highlight disease-relevant regions. The proposed approach integrates Multi-Scale Transformer blocks into the ResNet-50 framework to capture both local and global contextual features [9,10]. Additionally, a custom spatial attention mechanism is implemented to highlight disease-relevant pulmonary regions through attention-based weak localization [11]. Rather than serving as an autonomous diagnostic system, this model is designed to assist radiologists by highlighting potential nodules, thereby optimizing workflow efficiency and reducing the cognitive burden.

The main aim of this work is to develop a more sensitive computer-aided screening model that may help reduce false-negative interpretations in resource-constrained settings. Experimental results showed improved model performance relative to the baseline configuration. Most notably, the proposed RNNet-MST system achieved a 6.91% increase in Nodule Recall (rising to 93.09%), alongside a Test Accuracy of 95.16% and a Nodule F1-Score of 91.40%, successfully outperforming the baseline architecture across all key metrics.

While hybrid CNN--Transformer architectures have been widely explored in medical imaging, their application to pulmonary nodule analysis on chest X-rays remains relatively limited, particularly in resource-constrained screening settings. This study focuses on adapting such hybrid architectures specifically for CXR-based nodule detection, emphasizing sensitivity to small nodules and reduction of false-negative findings.

2. Materials and Methods

2.1. Dataset

This study utilized the NODE21 public dataset, a benchmark repository for pulmonary nodule detection on frontal-view chest radiographs. The dataset aggregates images from multiple sources, including JSRT, PadChest, ChestX-ray14, and Open-I, comprising 4,882 images in total: 1,134 positive cases containing 1,476 annotated pulmonary nodules with radiologist-provided bounding boxes, and 3,748 negative (nodule-free) cases. The dataset exhibits significant class imbalance, with the non-nodule class substantially overrepresented relative to the nodule class. All images were resized to 224×224 pixels to conform to the input layer specifications of the ResNet-50 backbone. The dataset was partitioned into training (70%), validation (15%), and test (15%) sets at the image level, with no overlap between splits.

2.2. Data Preprocessing and Augmentation

To improve generalization while preserving the intrinsic radiographic characteristics of chest X-ray images, controlled augmentation was applied only to the training set. The augmentation pipeline included resizing to 224×224 pixels, random horizontal flipping, small-angle rotation, mild translation, limited brightness/contrast perturbation, and normalization using the standard ImageNet mean and standard deviation. The nodule class was further upsampled to reduce class imbalance during training. PyTorch DataLoader objects were configured with a batch size of 16 and `num_workers = 4`.

2.3. Baseline Model: ResNet-50

The baseline model is a ResNet-50 deep convolutional neural network pretrained on ImageNet [6]. ResNet-50 introduced residual learning through skip connections, enabling training of 50-layer networks without suffering from the vanishing gradient problem. Its final fully connected layer was replaced with a 2-class output layer (Nodule / No Nodule) for binary classification. The baseline was trained for 25 epochs using AdamW ($\text{lr} = 1 \times 10^{-4}$, weight decay = 0.01), a Cosine Annealing learning rate scheduler, and Cross-Entropy Loss.

Despite strong general classification capability, ResNet-50 has two documented limitations addressed in this study: (1) its local receptive fields cannot model long-range spatial dependencies, causing misclassification of smaller-diameter nodules; and (2) its ImageNet-pretrained weights are optimized for natural-image features that differ substantially from the subtle, texture-dependent patterns of pulmonary nodules on grayscale CXRs, resulting in reduced sensitivity on the medical domain.

2.4. Proposed Architecture: RNNet-MST

RNNet-MST extends the baseline ResNet-50 by hierarchically integrating Multi-Scale Transformer (MST) blocks across all four backbone stages (Figure 1). This design was intended to

improve contextual feature modeling across scales and to better adapt pretrained convolutional features to grayscale chest radiographs. The resulting architecture combines convolutional feature extraction with transformer-based global context modeling in a single classification pipeline.

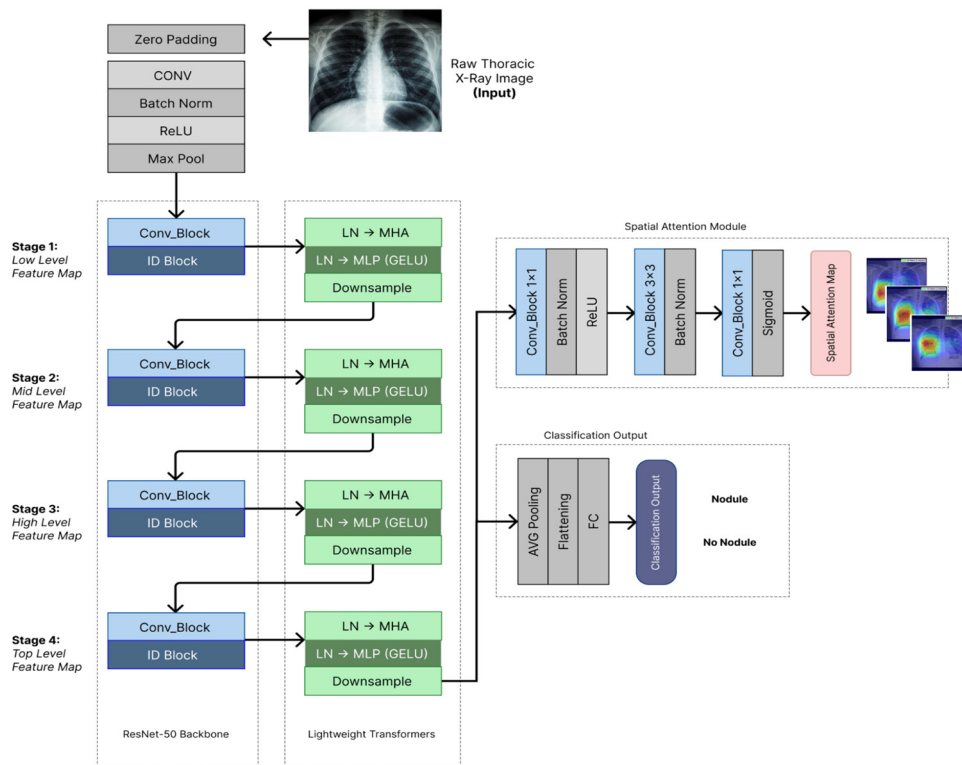


Figure 1. The proposed RNNet-MST architecture: MST blocks are integrated at all four stages of ResNet-50 to jointly capture long-range dependencies (Objective 1) and help mitigate the domain mismatch between ImageNet-pretrained features and CXR-specific patterns (Objective 2) in a single unified pipeline.

To capture long-range dependencies and global contextual information, MST blocks were hierarchically integrated at each of the four stages of the ResNet-50 backbone. Feature maps are extracted at four distinct stages: Stage 1 – Low-Level Features; Stage 2 – Mid-Level Features; Stage 3 – High-Level Features; and Stage 4 – Top-Level Features.

The output of each ResNet stage is passed into a corresponding Lightweight Transformer Block consisting of Layer Normalization (LN), Multi-Head Self-Attention (MHSA), and a Multi-Layer Perceptron (MLP) with GELU activation. Given an input image $I \in \mathbb{R}^{H \times W \times C}$, the feature map at stage s is produced by the ResNet stage function $f_s(\cdot)$ applied to the output of the previous stage:

$$X_s = f_s(X_{s-1}), \quad X_0 = I \quad (1)$$

where $X_s \in \mathbb{R}^{H_s \times W_s \times C_s}$ is the feature map at stage s , with H_s , W_s , and C_s denoting its spatial height, width, and number of channels respectively. Each X_s is then passed into a corresponding Lightweight Transformer block $T_s(\cdot)$:

$$Y_s = T_s(X_s) \quad (2)$$

where $Y_s \in \mathbb{R}^{H_s \times W_s \times C_s}$ is the transformer-enhanced feature map. To manage computational cost, X_s is downsampled to 14×14 before the attention operation and upsampled back to $H_s \times W_s$ afterward. Let $S_{MLP} \in \mathbb{R}^{H_s \times W_s \times C_s}$ denote the output of the MLP sub-layer within $T_s(\cdot)$, upsampled and reshaped back to the spatial dimensions of X_s . The final output feature map F_{out} is then obtained by combining S_{MLP} with the original feature map X_s via a residual connection to preserve low-level spatial information:

$$F_{\text{out}} = X_s + \text{Upsample}(\text{reshape}(S_{\text{MLP}})) \quad (3)$$

where $F_{\text{out}} \in R^{H_s} \times W_s \times C_s$ is the final enriched feature map passed to the next stage or classification head.

This hierarchical design ensures that global contextual information is captured continuously across all scales — from fine-grained textures in Stage 1 to high-level semantic structures in Stage 4 — overcoming ResNet-50's inherent local-receptive-field constraint.

In addition, the same MST integration simultaneously addresses the domain gap between ImageNet pretraining and medical radiographs. ResNet-50's convolutional filters are optimized for the distinct edges and RGB textures of natural scenes, not the subtle grayscale texture patterns of pulmonary nodules on CXR. By applying self-attention across the entire CXR image at each feature scale, the transformer blocks enable the model to contextualize local convolutional features within the global thoracic structure, learning CXR-specific representations that compensate for ResNet-50's natural-image inductive bias.

The MST blocks were integrated at all four stages to ensure that global contextual information is captured across multiple feature scales, from low-level textures to high-level semantic representations. Downsampling to 14×14 was used to balance computational efficiency with sufficient spatial resolution for attention modeling. Freezing the ResNet-50 backbone stabilizes training and preserves pretrained low-level feature representations, allowing the MST blocks to focus on adapting features to the target domain.

2.5. Training Configuration

During training, the ResNet-50 backbone weights (previously fine-tuned on NODE21) were frozen. Only the added MST blocks and the classification head were made trainable. The model was compiled with the following configuration:

- Optimizer: AdamW ($lr = 1 \times 10^{-4}$, weight decay = 0.01);
- Scheduler: Cosine Annealing Learning Rate;
- Loss Function: Weighted Binary Cross-Entropy (WBCE)

$$\mathcal{L}_{WBCE} = -\frac{1}{N} \sum_{i=1}^N [w_1 y_i \log(\hat{y}_i) + w_0 (1 - y_i) \log(1 - \hat{y}_i)] \quad (4)$$

where y_i is the ground-truth label, \hat{y}_i is the predicted probability, and w_1, w_0 are class weights computed as:

$$w_c = \frac{N}{|C| \times n_c} \quad (5)$$

yielding $w_1 = 1.7684$ for the Nodule class and $w_0 = 0.6971$ for the Non-Nodule class. Training ran for 25 epochs with a batch size of 16, saving model checkpoints whenever the total validation loss improved.

Model selection was based on the lowest validation loss. For reproducibility, key training details include the hardware configuration, software library versions, random seed, and decision threshold used for positive-class prediction. In this study, results are reported from a fixed train-validation-test split.

2.6. Evaluation Metrics

Model performance was assessed using the following metrics:

- Accuracy: $(TP + TN) / (TP + TN + FP + FN)$
- Precision: $TP / (TP + FP)$
- Nodule Recall (Sensitivity): $TP / (TP + FN)$
- Nodule F1-Score: $2 \times (\text{Precision} \times \text{Nodule Recall}) / (\text{Precision} + \text{Nodule Recall})$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

To evaluate performance on small nodules, a dedicated small-nodule subset was isolated from the test set comprising 171 CXR images whose radiologist-annotated bounding boxes measured below 70×70 pixels in either dimension (Figure 2), targeting small and irregular nodules. Performance on this subset was measured by detection rate and false negative count. To evaluate Objective 2, Nodule Recall improvement was measured on the full test set, reflecting the reduction in domain-gap-related false negatives. Both evaluations use the same single trained model — the distinction is analytical, not architectural.

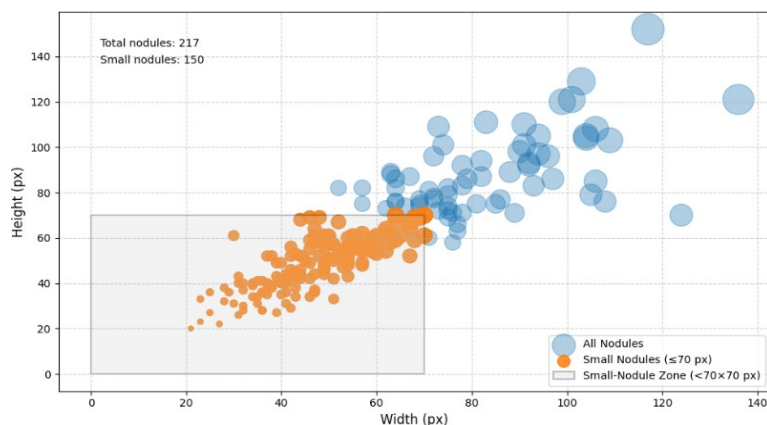


Figure 2. Distribution of BBox (width and height) within the selected small-nodule test subset, consisting of 171 CXR images with BBox width or height below 70 pixels.

3. Results

3.1. General Classification Performance

Figure 3 presents the test set classification performance of the baseline ResNet-50 vs RNNNet-MST. The baseline model achieved a Nodule Recall of 86.18%, indicating that 13.82% of positive cases in the test set were missed. Following the integration of MST blocks, RNNNet-MST achieved a Nodule Recall of 93.09%, corresponding to fewer false-negative classifications than the baseline model. A direct comparison of all primary metrics is provided in Table 1.

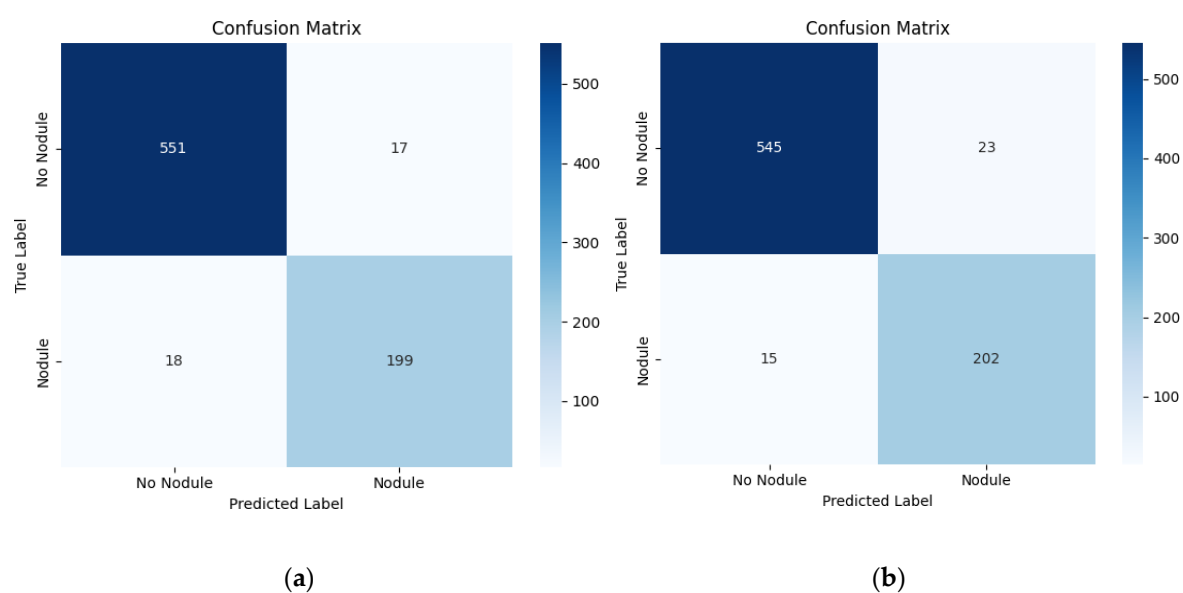


Figure 3. Test set classification performance confusion matrices. (a) Baseline ResNet-50. (b) Proposed RNNNet-MST.

Table 1. Performance metrics comparison of ResNet-50 and RNNNet-MST on the NODE21 test set.

Metric	Baseline ResNet-50	RNNNet-MST	Improvement
Accuracy	94.65%	95.16%	+0.51%
Nodule Recall	86.18%	93.09%	+6.91%
Nodule F1-Score	89.90%	91.40%	+1.50%

3.2. Comparative Evaluation

The proposed RNNNet-MST was evaluated against the baseline ResNet-50 on the NODE21 test set across two complementary analyses: overall classification performance on the full test set, and targeted detection performance on an isolated small-nodule subset. Both analyses compare the same two models — Baseline ResNet-50 and RNNNet-MST — reflecting the single unified architectural enhancement introduced in Section 2.

3.2.1. Overall Classification Performance

As shown in Table 1 and Table 2, the integration of MST blocks improved performance across all primary metrics. Nodule-Recall increased from 86.18% to 93.09% — an absolute gain of 6.91% — directly reducing the rate of false-negative interpretations. Test Accuracy improved by 0.51% and Nodule F1-Score improved by 1.50%.

Nodule Recall is equivalent to sensitivity for the positive class. The proposed model shows a marked improvement in Nodule Recall from 0.86 to 0.93 at the cost of a marginal decrease in Nodule Precision from 0.94 to 0.90, reflecting the expected sensitivity-precision trade-off in screening models.

Nodule Precision showed a marginal decrease in the proposed model. This reflects a well-established trade-off in medical AI: as a model becomes more aggressive in detecting positive cases, it may flag additional ambiguous regions as positive, marginally reducing Precision. In a screening context, this is clinically acceptable. As established by Luo et al. [14], recall (sensitivity) is the most critical metric in pulmonary nodule detection, since false negatives — patients with true nodules who are missed — can delay diagnosis and treatment, whereas false positives can be resolved through radiologist review.

The observed 6.91% recall improvement is consistent with the hypothesis that MST blocks improve contextual modeling and help adapt pretrained features to chest X-ray-specific patterns. Raghu et al. [13] demonstrated that ImageNet pretraining provides limited benefit for medical imaging tasks because visual feature distributions differ substantially between natural images and radiographs. By embedding MST blocks that apply self-attention over the image, the proposed model may better adapt ResNet-50 features to the visual characteristics of chest radiographs. This is further supported by Fu et al. [15], who demonstrated that hybrid CNN–Transformer architectures significantly outperform pure CNNs by overcoming spatial information loss and the style-over-content bias inherent in standard convolutional operations.

Table 2. Per-class classification performance of baseline ResNet-50 and proposed RNNNet-MST on the NODE21 test set.

Metric	Class	ResNet-50	RNNNet-MST
Precision	No Nodule	0.95	0.97
	Nodule	0.94	0.90
	Macro Average	0.94	0.94
	Weighted Average	0.95	0.95
Nodule Recall	No Nodule	0.98	0.96
	Nodule	0.86	0.93
	Macro Average	0.92	0.95
	Weighted Average	0.95	0.95
Nodule F1-Score	No Nodule	0.96	0.97

Nodule	0.90	0.91
Macro Average	0.93	0.94
Weighted Average	0.95	0.95

3.2.2. Small-Nodule Detection Performance

To further assess the model's sensitivity to the most diagnostically challenging cases, both models were additionally evaluated on an isolated small-nodule subset of 171 images whose radiologist-annotated bounding boxes measured below 70×70 pixels. As shown in Table 3, the baseline ResNet-50 correctly identified 138 nodules (80.7%) while failing to detect 33 cases (19.3%) as false negatives. RNNet-MST correctly classified 159 of 171 positive cases in the small-nodule subset, corresponding to a sensitivity of 93.0% and 12 false-negative cases (7.0%) — a 12.3% absolute improvement in sensitivity on the small-nodule subset.

Table 3. Comparative analysis of baseline ResNet-50 vs. RNNet-MST on the small-nodule subset ($\leq 70 \times 70$ px bounding boxes).

Model	Correctly Detected	False Negatives	Detection Rate
Baseline ResNet-50	138 / 171	33	80.7%
RNNet-MST	159 / 171	12	93.0%

On the full nodule test subset containing nodules of varying sizes (Table 4), the baseline ResNet-50 correctly identified 176 nodules (81.1%) and missed 41 cases (18.9%). RNNet-MST correctly classified 195 of 217 positive cases in the full nodule subset, corresponding to a sensitivity of 89.9% and 22 false-negative cases (10.1%) — an 8.8% improvement in sensitivity across heterogeneous nodule sizes.

Table 4. Comparative analysis of baseline ResNet-50 vs. RNNet-MST on nodules of varying sizes.

Model	Correctly Detected	False Negatives	Detection Rate
Baseline ResNet-50	176 / 217	41	81.1%
RNNet-MST	195 / 217	22	89.9%

These results suggest that the hierarchical MST integration helps mitigate the limitations of ResNet-50's restricted receptive field by enabling broader contextual modeling across the CXR. This is consistent with Raghu et al. [10], who demonstrated that Vision Transformer layers preserve spatial location information more effectively than ResNet layers, and with Dai and Gao [12], who showed that hybrid CNN-Transformer models consistently outperform pure CNNs in medical image classification.

All experiments were conducted using a fixed train-validation-test split; future work will include repeated runs and statistical analysis to assess performance variability and robustness.

4. Discussion

4.1. Interpretation of Performance Gains

The primary objective of this study was to address the critical diagnostic gap in pulmonary nodule detection, characterized by high rates of false-negative interpretations due to the anatomical complexity of CXRs and radiologist fatigue. The proposed RNNet-MST architecture improved Nodule Recall by 6.91 percentage points relative to the baseline ResNet-50 model.

These findings are consistent with the view that standard convolutional neural networks are limited by local receptive fields when subtle abnormalities must be interpreted within broader anatomical context. By integrating Multi-Scale Transformer (MST) blocks, the proposed model effectively captured the long-range spatial dependencies required to contextualize subtle

radiographic abnormalities. This improvement was most pronounced in the targeted small-nodule subset, where the RNNet-MST model increased the detection rate by 12.3% and reduced false negatives from 33 to just 12 cases. These results align with Raghu et al. [10], who demonstrated that Vision Transformers preserve spatial location information more effectively than pure CNNs, and corroborate the findings of Fu et al. [15] regarding the superiority of hybrid architectures in overcoming CNN style-over-content biases in medical imaging.

4.2. Comparison with Related CXR-Based Nodule Detection Work

To situate the performance of RNNet-MST within the broader literature, Table 5 summarizes representative deep learning methods for pulmonary nodule detection on chest radiographs. A consistent observation across the surveyed works is that CXR-based detection methods report substantially lower sensitivity benchmarks than their CT counterparts, underscoring the inherent difficulty of the task and the importance of continued research on this modality. Behrendt et al. [20] achieved the highest reported sensitivity in the CXR domain through an ensemble of four state-of-the-art object detectors trained on the same NODE21 dataset, winning the Node21 challenge; however, their approach carries a substantially higher computational cost due to the multi-model ensemble strategy. Against this backdrop, RNNet-MST achieved a Nodule Recall of 93.09% on the NODE21 test set, representing one of the strongest recall values among the CXR-based methods summarized here, although direct cross-study comparison remains limited by differences in datasets, task definitions, and evaluation protocols.

Table 5. Comparison of RNNet-MST with related deep learning methods for pulmonary nodule detection on chest radiographs (CXR).

Study	Year	Method/Model	Dataset	Sensitivity/ Recall	Key Notes
Yoo et al. [16]	2020	Deep learning algorithm (commercial CAD)	NLST	74.0%	5,485 participants; specificity 73%; AUC 0.86
Schultheiss et al. [17]	2021	RetinaNet / U-Net CNN	Synthetic (from LIDC-IDRI CT)	wAFROC: 0.81	201 synthetic radiographs; p=0.49 vs. radiologists
Chiu et al. [18]	2022	YOLOv4 + U-Net lung segmentation	TVGH + JSRT	79.0%	3.04 FP/image; 254 CXRs tested
Shimazaki et al. [19]	2022	CNN segmentationbased DL model	In-house (Osaka City Univ.)	73.0%	0.13 mFPI; lower sensitivity in blind spots (50–64%)
Behrendt et al. [20]	2023	Ensemble (Faster-RCNN, RetinaNet, EfficientDet-D2, YOLOv5)	NODE21	FROC25%: ~0.84	Node21 competition winner; AUROC + FROC metric
Present study	2026	RNNet-MST	NODE21	93.09%	Reports strong recall relative to the CXR-based methods summarized here

*mFPI = mean false positive indications per image; FROC_{25%} = sensitivity at 25% false positive rate; wAFROC = weighted alternative free-response ROC figure of merit; AUC = area under the ROC curve. Note: Direct comparison across studies should be interpreted cautiously due to differences in datasets, evaluation protocols, and metric definitions.

4.3. Clinical Significance of the Precision-Recall Trade-off

A critical point of interpretation in this study is the observed precision-recall trade-off. While Nodule Recall improved to 93.09%, Nodule Precision experienced a marginal decrease from 0.94 to 0.90. In a first-line screening context, this trade-off may be acceptable because higher sensitivity can reduce the risk of missed nodules. As emphasized by Luo et al. [14], maximizing sensitivity is paramount in cancer screening; a false positive prompts a secondary CT scan, whereas a false negative—a missed malignant nodule—can fatally delay treatment. The improved sensitivity of RNNet-MST is relevant to the diagnostic blind spots described in prior studies of missed nodules on chest radiographs.

4.4. Workflow Optimization in Resource-Constrained Settings

Furthermore, these performance gains carry significant implications for the Philippine healthcare system. Expert interviews conducted at the Lung Center of the Philippines highlighted that radiologists routinely interpret up to 200 images daily, leading to intense cognitive load. By compensating for the domain gap between natural-image pretraining and grayscale radiographs, RNNet-MST provides attention maps that may support sensitive visual assessment of suspicious regions. This functions not as an autonomous diagnostic replacement, but as an intelligent visual aid designed to reduce perceptual errors, streamline workflow, and alleviate professional burnout in resource-constrained environments.

It is important to note that the proposed system operates as a classification-based computer-aided detection (CAD) model with weak localization via attention maps, rather than a fully supervised object detection framework. While object detection models such as Faster R-CNN or DETR provide explicit bounding box predictions, they typically require dense annotations and higher computational resources. In contrast, the proposed approach prioritizes computational efficiency and reduced annotation requirements, making it more suitable for screening applications in resource-constrained clinical environments.

4.5. Study Limitations and Future Research Directions

This study has several limitations. First, evaluation was conducted on a fixed train-validation-test split without repeated runs, confidence intervals, or formal statistical significance testing. Second, the model was assessed on a single public dataset, which limits conclusions about generalizability across institutions and acquisition settings. Third, the localization analysis was attention-based and therefore should be interpreted as weak localization rather than detector-level lesion localization.

In addition, the dataset primarily consists of frontal-view radiographs and may not fully capture the variability present in real-world clinical settings, particularly in local Philippine hospitals. The current architecture is also limited to binary classification (nodule vs. no nodule) and does not differentiate between benign and malignant nodules, which restricts its direct clinical interpretability. Furthermore, an ablation study isolating the individual contributions of the Multi-Scale Transformer (MST) blocks and spatial attention module was not conducted. As such, the relative impact of each component cannot be independently quantified.

Future work will include ablation experiments isolating the contributions of the MST blocks and spatial attention module. Future research directions include extending the proposed hybrid backbone into object detection frameworks, such as Faster R-CNN or DETR, to enable precise bounding box prediction rather than relying on classification and attention-based localization. Additionally, incorporating histopathologically confirmed datasets would allow the model to perform multi-class malignancy classification. Finally, validation on larger, independent clinical cohorts — particularly using localized data from Philippine medical institutions — will be essential to ensure clinical robustness and deployment readiness.

5. Conclusions

This study proposed RNNet-MST, a hybrid deep learning architecture for pulmonary nodule classification on chest X-ray (CXR) images, developed by enhancing ResNet-50 with Multi-Scale Transformer (MST) blocks. The proposed model was designed to address two documented limitations of the baseline ResNet-50: its limited ability to capture long-range dependencies relevant to small-nodule classification, and its reduced sensitivity associated with the domain gap between ImageNet pretraining and CXR-specific features.

Experimental results on the NODE21 dataset showed improvements across the primary evaluation metrics. Most critically, the model achieved a Nodule Recall of 93.09%, representing a 6.91% improvement over the baseline and corresponding to fewer false-negative classifications. Test Accuracy improved by 0.51% to 95.16%, and Nodule F1-Score improved by 1.50% to 91.40%. On the isolated small-nodule subset, the proposed model achieved a 12.3% improvement in sensitivity over the baseline.

These findings suggest that combining convolutional feature extraction with multi-scale transformer-based contextual modeling can improve sensitivity in CXR-based pulmonary nodule classification. This may be particularly valuable in resource-constrained clinical settings, where high radiologist workloads and limited access to CT imaging increase the need for assistive screening tools. Future work should extend this architecture toward full bounding box detection using frameworks such as Faster R-CNN or DETR, incorporate benign-to-malignant nodule classification using histopathologically confirmed datasets, and validate the model across additional large-scale and independent datasets to strengthen generalizability.

Author Contributions: Conceptualization, E.F.B., E.T.M.; methodology, E.F.B., E.T.M., H.S.S.III, W.T.C.; software, E.F.B., E.T.M.; validation, E.F.B., E.T.M., H.S.S.III, Y.M.G., K.E.M, R.A.R.B., I.C.A., W.T.C., D.M.A.C.; formal analysis, E.F.B., E.T.M., H.S.S.III, Y.M.G., W.T.C.; investigation, E.F.B., E.T.M.; resources, E.F.B., E.T.M., W.T.C., D.M.A.C.; data curation, E.F.B., E.T.M.; writing—original draft preparation, E.F.B., E.T.M.; writing—review and editing, H.S.S.III, W.T.C.; visualization, E.F.B., E.T.M.; supervision, H.S.S.III, W.T.C., D.M.A.C.; project administration, E.F.B., E.T.M., H.S.S.III, Y.M.G., R.A.R.B., I.C.A., W.T.C., D.M.A.C.; funding acquisition, H.S.S.III. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Department of Science and Technology – Philippine Council for Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD) – Balik Scientist Program (BSP).

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of secondary, fully anonymized data from existing databases, with no direct involvement of human participants and no access to identifiable personal information.

Informed Consent Statement: Not applicable. The study used anonymized secondary data and did not involve direct human participation. No identifiable individual data are included in this manuscript.

Data Availability Statement: The dataset used in this study is the publicly available NODE21 chest X-ray dataset. Additional implementation details and derived materials are available from the corresponding authors.

Acknowledgments: During the preparation of this manuscript/study, the authors used Gemini, ChatGPT, Perplexity AI, and Claude for language editing and improvement of readability. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAD Computer-Aided Detection

CBAM	Convolutional Block Attention Module
CNN	Convolutional Neural Network
CT	Computed Tomography
CXR	Chest X-Ray
FC	Fully Connected
FN	False Negative
FP	False Positive
GELU	Gaussian Error Linear Unit
LCP	Lung Center of the Philippines
LN	Layer Normalization
MHSA	Multi-Head Self-Attention
MLP	Multi-Layer Perceptron
MST	Multi-Scale Transformers
NODE21	Nodule Detection 2021
ResNet	Residual Neural Network
RNNNet-MST	Residual Neural Network with Multi-Scale Transformers
TN	True Negative
TP	True Positive
ViT	Vision Transformer
WBCE	Weighted Binary Cross-Entropy

Appendix A. Quantitative and Qualitative Analysis of Spatial Attention

Appendix A.1. Quantitative Analysis of Baseline vs Enhanced Model

The attention localization performance was evaluated using four complementary metrics designed to assess whether nodule bounding boxes fall within high-attention regions, which is a critical requirement for assistive/computer-aided diagnostic systems.

Table A1. Attention-bounding box alignment performance metrics for baseline ResNet-50 with Grad-CAM and enhanced RNNNet-MST with the spatial attention module.

Metric	Baseline	Enhanced	Improvement
BBox Coverage	28.71%	39.93%	+11.22%
Detection Rate	29.95%	39.63%	+9.68%
Peak Proximity	80.77%	83.80%	+3.03%
Attention Focus	2.1856	3.1337	+0.94

BBox Coverage measures the average attention intensity within ground truth nodule regions (0-1 scale), with higher values indicating stronger highlighting of pathological areas. Detection Rate quantifies the percentage of nodules that fall within high-attention regions (threshold >0.5), providing a binary measure of localization success. Peak Proximity evaluates the normalized distance from the maximum attention point to the nodule center (0-1 scale), where scores approaching 1.0 indicate precise centering of attention on nodules. Attention Focus Ratio compares the mean attention inside versus outside nodule regions, with values >1.0 indicating preferential focus on pathological tissue; this metric is particularly important for reducing false highlights in clinical settings.

As shown in Table A1, the enhanced model with spatial attention demonstrates consistent and substantial improvements across all evaluated metrics compared to the baseline. The Attention Focus Ratio increased from 2.19 to 3.13, indicating a markedly stronger contrast between nodule regions and surrounding background tissue. BBox Coverage improved from 28.71% to 39.93%, reflecting more salient and spatially concentrated highlighting of pathological regions.

The Detection Rate increased from 29.95% to 39.63%, corresponding to a 9.68 percentage point improvement, which indicates that nearly 40% of nodules are captured within visually prominent attention regions despite the absence of bounding box supervision during training. Peak Proximity remained consistently high at 83.80%, showing that attention peaks are reliably centered near true nodule locations, with a modest improvement introduced by the spatial attention module. Overall,

these results confirm that the spatial attention mechanism effectively concentrates computational focus on clinically relevant regions while preserving interpretability for radiologist review.

Appendix A.2. Visual Comparison of Baseline vs Enhanced Model

The resulting heatmaps showed that the model consistently focused on anatomically relevant regions, particularly the lungs, indicating more reliable and clinically meaningful localization. Qualitative analysis of representative cases revealed distinct differences in attention behavior between the baseline and enhanced models.

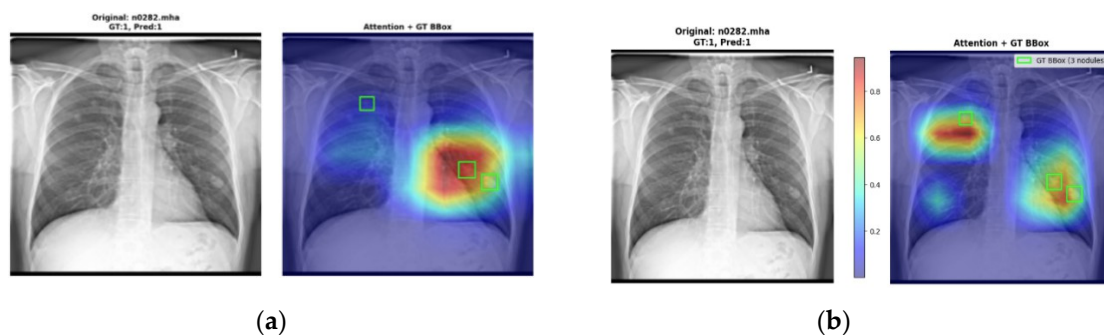


Figure A1. Heatmap Activation comparison of (a) Baseline vs (b) Enhanced Model attention for patient n0282 in the test subset derived from the NODE21 dataset.

Figure A2 presents a particularly notable case involving three nodules (Patient n0282), two in the right lung and one in the left lung. The baseline model captured only the two nodules in the right lung, completely missing the solitary left lung nodule. In contrast, the enhanced model demonstrated attention coverage across all three nodules, suggesting that the multi-scale light transformers embedded in the enhanced ResNet-50 model contribute to improved spatial awareness through long-range dependency modeling and global context aggregation. This ability to simultaneously attend to spatially distant pathological regions indicates that the transformer components enable the model to integrate information across the entire lung field rather than focusing exclusively on locally salient features.

Furthermore, Figure A2 (Patient n0368) demonstrates distributed attention across different lung regions while maintaining peak activation precisely within the bounding box of a very small nodule in the right lung. This behavior suggests that the multi-scale light transformer architecture influences attention allocation decisions by balancing global context awareness with fine-grained local feature discrimination, enabling detection of subtle pathological findings that might otherwise be overlooked in traditional convolutional approaches.

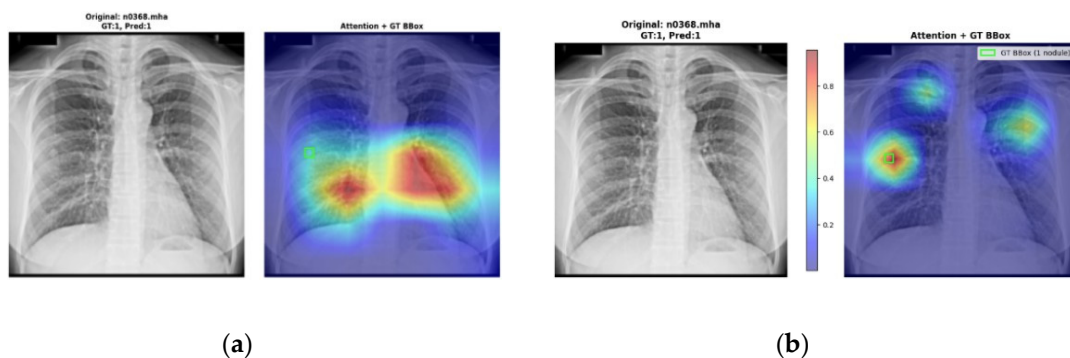


Figure A2. Heatmap Activation comparison of (a) Baseline vs (b) Enhanced Model attention for patient n0368 in the test subset derived from the NODE21 dataset.

Appendix B. Techniques to Address Class Imbalance

To ensure the model remains highly sensitive to pathological findings, specific data augmentation and loss weighting strategies were implemented to address the severe class imbalance and eliminate the inherent bias toward the majority “No Nodule” class.

Since the dataset was heavily imbalanced, the researchers applied data augmentation techniques to the training set. Only minimal augmentation was performed to preserve the intrinsic characteristics of the chest X-rays:

- **Horizontal Translation:** Random shifts of $\pm 2\%$ of image width.
- **Vertical Translation:** Random shifts of $\pm 2\%$ of image height.
- **Rotation:** Small random rotations within ± 3 degrees.
- **Brightness Adjustment:** Multiplicative intensity scaling by random factor between 0.95–1.05 ($\pm 5\%$), clipped to original intensity range.
- **Horizontal Flip:** Left-right mirroring of the image.
- **Combined Transform:** Horizontal shift ($\pm 1\%$) combined with rotation ($\pm 2^\circ$).

However, during the training of the spatial attention layer, augmented images from the upsampled set could not be paired with accurate bounding box coordinates due to random shifts, flips, and rotations. Due to this, the researchers could only utilize the original unbalanced dataset for the training of spatial attention layers.

To address this limitation, the researchers used Weighted Binary Cross-Entropy (WBCE) to compensate for the imbalance when training the spatial attention layer without reliable bounding box information. The class weights were computed using the formula: $\text{class weights} = \frac{\text{total samples}}{(\text{len}(\text{class_counts}) \times \text{class counts})}$.

This weighting strategy helped the model remain sensitive to true nodules even in the absence of spatial labels. The final calculated weights were 1.7684 for the Nodule class and 0.6971 for the Non-Nodule class.

Appendix C. Preliminary Clinical Interface Evaluation and Directions for Future Validation

To obtain initial qualitative insights into the clinical utility and interface usability of the proposed system, a formative pilot evaluation was conducted with a single practicing radiologist from the College of Medicine at the Pamantasan ng Lungsod ng Maynila (PLM) with 6 years of clinical experience. The radiologist interacted with the system interface integrating the proposed AI tool (Figure A3) and completed a structured survey across three domains: Diagnostic Alignment and Accuracy, Clinical Impact and Decision Making, and System Usability, rated on a 5-point scale.

System usability received the highest scores (5/5 across all usability metrics), suggesting that the interface design — including heatmap visualization and zoom functionality — is intuitive and ready for broader clinical testing. Diagnostic alignment and automated classification agreement received neutral scores (3/5), while the utility of the AI in identifying subtle nodules in cases of discrepancy was rated lower (2/5). Rather than undermining the proposed model, these scores honestly reflect the current performance ceiling of a single-dataset, classification-based system and directly motivate the directions for future work outlined in Section 4.5. Importantly, despite the neutral alignment scores, the radiologist reported that the AI output prompted reconsideration or refinement of the final diagnosis (4/5) and increased confidence in the final diagnosis after using the tool (4/5), suggesting potential clinical value as a second-opinion aid even at this early stage.

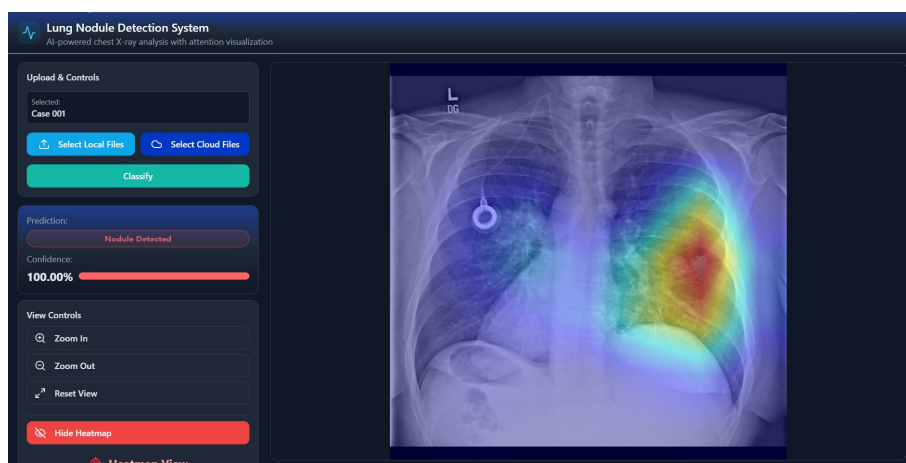


Figure A3. The system interface integrated with the enhanced spatial attention model for radiologist evaluation.

The primary strength noted qualitatively was system usability, described as user friendly and easy to navigate. The key area for improvement identified was the need for training on more diverse samples to improve nodule detection precision — consistent with the generalizability limitations discussed in Section 4.5.

These preliminary findings should be interpreted as formative and exploratory, intended to identify early interface and performance trends rather than provide statistically generalizable clinical conclusions. They directly inform the design of a planned prospective validation study involving a larger cohort of radiologists across multiple Philippine clinical institutions, incorporating standardized usability instruments such as the System Usability Scale (SUS) and blinded diagnostic comparison against radiologist ground truth. Full clinical validation, including multi-reader agreement analysis and evaluation on locally acquired Philippine hospital data, is reserved as a dedicated subsequent study.

References

1. Malhotra, J.; Malvezzi, M.; Negri, E.; La Vecchia, C.; Boffetta, P. Risk factors for lung cancer worldwide. *Eur. Respir. J.* **2016**, *48*, 889–902.
2. Rivera Medical Center Inc. Top 5 leading causes of death in the Philippines, 2025. Available online: <https://www.rmci.com.ph/top-5-leading-causes-of-death-in-the-philippines-2025/> (accessed on 14 March 2026).
3. Çalli, E.; Sogancioglu, E.; van Ginneken, B.; van Leeuwen, K.G.; Murphy, K. Deep learning for chest X-ray analysis: A survey. *Med. Image Anal.* **2021**, *72*, 102125.
4. Miki, S.; Nomura, Y.; Hayashi, N.; Hanaoka, S.; Maeda, E.; Yoshikawa, T.; Masutani, Y.; Abe, O. Prospective Study of Spatial Distribution of Missed Lung Nodules by Readers in CT Lung Screening Using Computer-assisted Detection. *Acad. Radiol.* **2021**, *28*, 647–654.
5. Digumarthy, S.R.; Gullo, R.L.; Levesque, M.H.; Sayegh, K.; Rao, S.; Raymond, S.B.; Otrakji, A.; Kalra, M.K. Cause determination of missed lung nodules and impact of reader training and education: Simulation study with nodule insertion software. *J. Cancer Res. Ther.* **2020**, *16*, 780–787.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
7. Bush, I. Lung nodule detection and classification. Technical Report, Stanford University, Stanford, CA, USA, **2016**.
8. Borji, A. Addressing a fundamental limitation in deep vision models: Lack of spatial attention. *arXiv* **2024**, arXiv:2407.01782.

9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
10. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do Vision Transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.
11. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 3–19.
12. Dai, Y.; Gao, Y.; Liu, F. TransMed: Transformers Advance Multi-Modal Medical Image Classification. *Diagnostics* **2021**, *11*, 1384.
13. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding transfer learning for medical imaging. *arXiv* **2019**, arXiv:1902.07208.
14. Luo, D.; Yang, I.; Bae, J.; Woo, Y. Research on Performance Metrics and Augmentation Methods in Lung Nodule Classification. *Appl. Sci.* **2024**, *14*, 5726.
15. Fu, X.; Lin, R.; Du, W.; Tavares, A.; Liang, Y. Explainable hybrid transformer for multi-classification of lung disease using chest X-rays. *Sci. Rep.* **2025**, *15*, 6650.
16. Yoo, H.; Kim, K.H.; Singh, R.; Digumarthy, S.R.; Kalra, M.K. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. *JAMA Netw. Open* **2020**, *3*, e2017135.
17. Schultheiss, M.; Schmette, P.; Bodden, J.; Aichele, J.; Müller-Leisse, C.; Gassert, F.G.; Gassert, F.T.; Gawlitza, J.F.; Hofmann, F.C.; Sasse, D.; et al. Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance. *Sci. Rep.* **2021**, *11*, 15857.
18. Chiu, H.-Y.; Peng, R.H.-T.; Lin, Y.-C.; Wang, T.-W.; Yang, Y.-X.; Chen, Y.-Y.; Wu, M.-H.; Shiao, T.-H.; Chao, H.-S.; Chen, Y.-M.; Wu, Y.-T. Artificial intelligence for early detection of chest nodules in X-ray images. *Biomedicine* **2022**, *10*, 2839.
19. Shimazaki, A.; Ueda, D.; Choppin, A.; Yamamoto, A.; Honjo, T.; Shimahara, Y.; Miki, Y. Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. *Sci. Rep.* **2022**, *12*, 727.
20. Behrendt, F.; Bengs, M.; Bhattacharya, D.; Krüger, J.; Opfer, R.; Schläefer, A. A systematic approach to deep learning-based nodule detection in chest radiographs. *Sci. Rep.* **2023**, *13*, 10120.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.