Test Report

RadiSen AXIR

โดย บริษัท ไทย จีแอล จำกัด

รายงานผลการทดสอบ
โดยราชวิทยาลัยรังสีแพทย์แห่งประเทศไทย

ทดสอบใช้กับภาพรังสีทรวงอก ในกรณี

- คัดกรอง (screening) วัณโรคปอด
- อ่านผลซ้ำ (double reading) ให้กับรังสีแพทย์ เพื่อเพิ่มคุณภาพการวินิจฉัย
- เพิ่มความแม่นยำในการค้นหาพยาธิสภาพให้กับรังสีแพทย์
- ประมาณความยาก-ง่ายในการแปลผล
- จัดลำดับความเร่งด่วน (triage) ในการแปลผลให้แก่รังสีแพทย์

Report on the Test Performance of Artificial Intelligence for Tuberculosis Screening in Chest X-Ray Images of the Thai Population

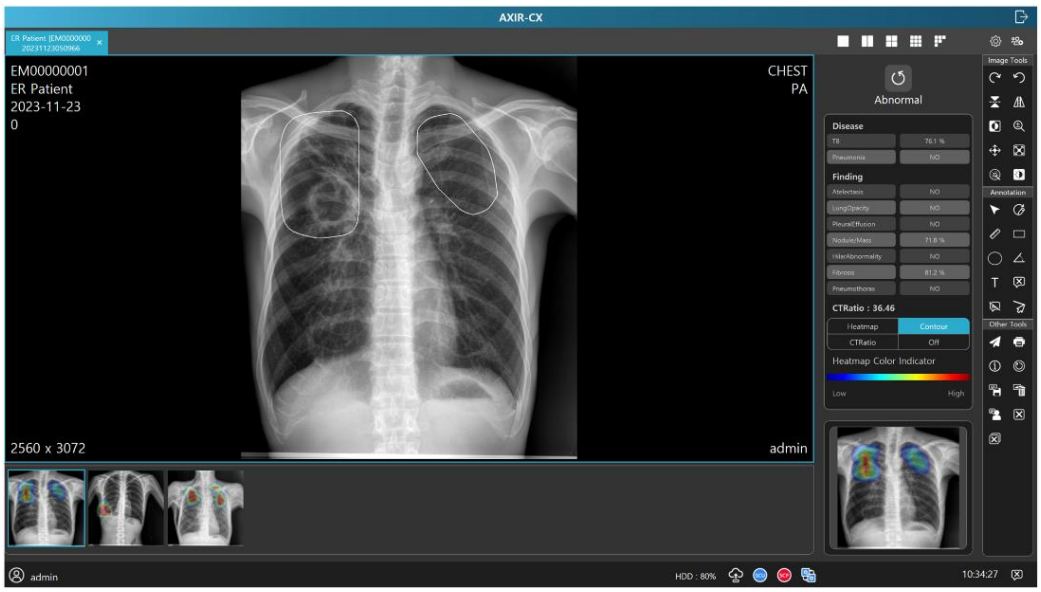
Filer Name

Company	Thai GL Co., Ltd.
Address	388 Muban Town in Town Soi Ladphrao 94 Ladphrao Rd., Phlabphla, Wangthonglang, Bangkok 10310
Contact	Woravut Kumin

Developer Company

Company	Radisen Co., Ltd
Address	4F, 46 Hoam-ro 26ga-gil, Gwanak-gu, Seoul, 08812
Country	South Korea
Website	https://radisentech.com/

Software

Name	RadiSen AXIR
Version	Not specified
Description	Product specifications excerpted from https://radisentech.com/portfolio/axir/  AXIR-CX is an automated AI system to detect pulmonary abnormalities and diseases. The AXIR software is designed for use by radiologists and radiology technicians for annotation in the Chest X-ray images.

Dataset

Reference No.	1A2A
Number of Images	808
Internal Validation	Consistent

Data Characteristics

The dataset consists of 808 randomly selected chest radiographic images from a pool of 1,500 images carefully curated from Songklanagarind Hospital in Songkhla Province, Chiangrai Pracharuk Hospital in Chiang Rai Province, Udon Thani Hospital in Udon Thani Province, Suttawet Hospital in Maha Sarakham Province, and the Tuberculosis Division of the Department of Disease Control, Ministry of Public Health. Each image was read by three B Readers. Our goal is to utilize high-quality datasets that are read by B Readers, who are trained and certified radiologists.

A B Reader is a qualified radiologist who is certified by the National Institute for Occupational Safety and Health (NIOSH) in the United States. B Readers are specifically trained to interpret and classify chest radiographs for the presence of pneumoconiosis, a group of lung diseases.

Characteristics of the radiographic images:

- Chest radiographic images of patients aged 15 years and above were included, taken with a computed radiography machine.
- No images from patients with a positive HIV Serology status.
- No images from patients with other opportunistic pulmonary infections or co-infections, such as Mycobacterium tuberculosis, Histoplasmosis, Cryptococcosis, Melioidosis, and Acinetobacter baumannii.

To assess the inter-rater reliability, the following metrics were employed:

- Pairwise Agreement: The average level of agreement among each pair of B readers.
- Intraclass Agreement (ICC): The average Pearson's correlation using ICC(2,3) when three B readers read the randomly selected radiographic images.
- Pairwise Cohen's Kappa and Fless' Kappa statistics for the analysis of agreement between assessors

Number of Findings

Table 1 presents the number of findings annotated by B Readers for chest X-ray images in Dataset 1A2A, which consists of 808 images. Each image in the dataset was independently assessed by three randomly selected B Readers from a pool of six B Readers. $N_{\text{Individual Reader}}$ represents The number of findings that each individual B reader labelled, while $N_{\text{Concensus}}$ represents the number of findings where the majority of the B Readers agreed.

Table 1 Number of findings annotated annotated by B Readers in Dataset 1A2A

Finding	$N_{Individual\ Reader}$	$N_{Concensus}$
Abnormalities	1,575	513
Small opacity	1,252	421
Primary nodular	929	324
Primary reticular	308	58
Secondary nodular	718	242
Secondary reticular	455	110
Large opacity	1,240	422
Mass/nodule	497	136
Cavity	881	298
Fibrosis	742	243
Calcification	299	58
Pleural effusion	327	109
Pleural thickening	556	179
Pneumothorax	14	4
Hilar adenopathy	316	72
Mediastinal adenopathy	96	17
Consistent with tuberculosis	1,270	416
Active Tuberculosis	1,222	408
Patchy infiltration	930	336
Cavity with surrounding consolidation	813	280
Unilateral hilar/paratracheal lymph node enlargement	147	30
Pleural effusion	165	49
Miliary nodules	310	76
Indeterminate tuberculosis	48	6
Reticulonodular infiltration	28	4
Destroyed lung or bronchiectasis	5	0
Inconsistent with tuberculosis	1,154	392

Inter-rater Reliability

Table 2 Inter-rater reliability measures for each finding in Dataset 1A2A (808 images). Each finding was interpreted by three B Readers. The reliability was measured using statistical metrics such as Pairwise Agreement, ICC(2,3), Pairwise Cohen's kappa, and Fleiss' kappa.

Finding	Agreement	ICC	Cohen's	Fleiss'
Abnormalities	0.9208	0.9345	0.826	0.826
Small opacity	0.8589	0.8841	0.7175	0.7175
Primary nodular	0.8276	0.8395	0.6352	0.6352
Primary reticular	0.8069	0.3092	0.1296	0.1297
Secondary nodular	0.7063	0.5576	0.2953	0.2955
Secondary reticular	0.7434	0.3615	0.1587	0.1585
Large opacity	0.9043	0.9269	0.8085	0.8085
Mass/nodule	0.7748	0.5734	0.309	0.309
Cavity	0.8688	0.8837	0.7168	0.7165
Fibrosis	0.7632	0.7051	0.4429	0.4426
Calcification	0.8177	0.3586	0.157	0.1569
Pleural effusion	0.9389	0.8945	0.7381	0.7384
Pleural thickening	0.8457	0.7951	0.5639	0.5636
Pneumothorax	0.9967	0.8816	0.7095	0.7126
Hilar adenopathy	0.8399	0.5564	0.2952	0.294
Mediastinal adenopathy	0.9398	0.4438	0.2062	0.2082
Consistent with tuberculosis	0.9604	0.9721	0.9206	0.9206
Active Tuberculosis	0.9538	0.9672	0.9076	0.9076
Patchy infiltration	0.8284	0.8407	0.6371	0.6371
Cavity with surrounding consolidation	0.8507	0.8565	0.6651	0.665
Unilateral hilar/paratracheal lymph node enlargement	0.9051	0.3763	0.1651	0.1672
Pleural effusion	0.9406	0.7733	0.5314	0.5318
Miliary nodules	0.8234	0.4418	0.2087	0.2084
Indeterminate tuberculosis	0.9686	0.4183	0.197	0.1923
Reticulonodular infiltration	0.9802	0.3178	0.1643	0.1328
Destroyed lung or bronchiectasis	0.9959	-0.006	-0.0019	-0.0021
Inconsistent with tuberculosis	0.9604	0.9721	0.9206	0.9206

Table 3 Interpretation of ICC and Kappa Values according to Landis and Koch (1977)[1]

ICC/Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

[1] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. In Biometrics (Vol. 33, Issue 1, p. 159). JSTOR. https://doi.org/10.2307/2529310

Results

The inter-rater reliability is measured using Pairwise Agreement, which is the average similarity between each pair of B Readers and RadiSen AXIR, as well as Pairwise Cohen's Kappa, which is the average of Cohen's Kappa statistics between each pair of B Readers and RadiSen AXIR. This is done to compare the agreement between B Readers and RadiSen AXIR ("B" vs AI) and among B Readers themselves ("B" vs "B").

Table 4 Reliability Measures Within B Readers (“B” vs “B”) and Between the System and B Readers (“B” vs AI)

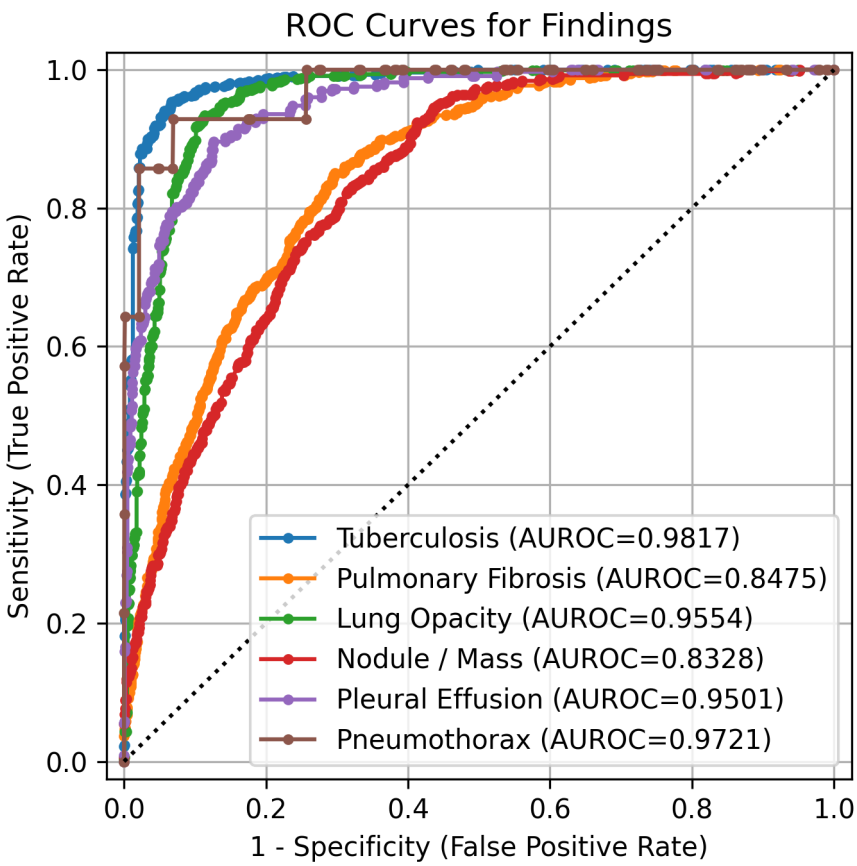
Finding	N	Threshold	Pairwise Agreement		Cohen’s Kappa	
			“B” vs “B”	“B” vs AI	“B” vs “B”	“B” vs AI
Tuberculosis	1,270	0.30	**0.9602**	0.9431	**0.9186**	0.8862
Primary Fibrosis	742	0.30	0.7614	**0.7789**	0.4363	**0.4527**
Lung Opacity	1,240	0.30	**0.9060**	0.8960	**0.8104**	0.7912
Nodule / Mass	497	0.20	**0.7695**	0.6588	**0.3413**	0.3196
Pleural Effusion	327	0.50	**0.9391**	0.9303	**0.7304**	0.6482
Pneumothorax	14	0.30	**0.9966**	0.9769	**0.6659**	0.2361

For measuring the diagnostic performance of each disease annotation, criteria such as Sensitivity, Specificity, Positive Prediction Rate (PPR), and Negative Prediction Rate (NPR) are utilized. These metrics are evaluated using the diagnostic threshold specified by the manufacturer, along with the area under the ROC curve.

Table 5 Diagnostic Performance of Each Finding by the System Compared to B Readers

Finding	N	Threshold	Sensitivity	Specificity	PPV	NPV	AUROC
Tuberculosis	1,270	0.30	0.9157	0.9731	0.9740	0.9130	0.9885
Primary Fibrosis	742	0.30	0.5512	0.8793	0.6683	0.8162	0.8475
Lung Opacity	1,240	0.30	0.9718	0.8167	0.8474	0.9651	0.9554
Nodule / Mass	497	0.20	0.8934	0.5983	0.3645	0.9561	0.8328
Pleural Effusion	327	0.50	0.5627	0.9876	0.8762	0.9354	0.9501
Pneumothorax	14	0.30	0.6429	0.9788	0.1500	0.9979	0.9721

Figure 1 ROC Curves Illustrating Diagnostic Performance for Each Finding



Analysis of Results

According to Table 6, when comparing Pairwise Agreement and Cohen's Kappa between B Readers and RadiSen AXIR ("B" vs AI) and among B Readers themselves ("B" vs "B"), RadiSen AXIR demonstrates performance close to that of B Readers (with a difference of less than 5%). For tuberculosis, the Pairwise Agreement of among B readers scored higher than the Pairwise Agreement of each B reader and RadiSen AXIR by 1.71% (N=1,270) and the Cohen's Kappa of among B readers scored higher than the Cohen's Kappa of each B reader and RadiSen AXIR by 3.24% (N=1,270).

Table 6 Differences between Pairwise Agreement and Cohen's Kappa

Finding	Pairwise Agreement			Cohen's Kappa		
	B vs "B"	"B" vs AI	Diff	"B" vs "B"	"B" vs AI	Diff
Tuberculosis	**0.9602**	0.9431	-1.71%	**0.9186**	0.8862	-3.24%
Primary Fibrosis	0.7614	**0.7789**	1.75%	0.4363	**0.4527**	1.64%
Lung Opacity	**0.906**	0.896	-1.00%	**0.8104**	0.7912	-1.92%
Nodule / Mass	**0.7695**	0.6588	-11.07%	**0.3413**	0.3196	-2.17%
Pleural Effusion	**0.9391**	0.9303	-0.88%	**0.7304**	0.6482	-8.22%
Pneumothorax	**0.9966**	0.9769	-1.97%	**0.6659**	0.2361	-42.98%

Regarding the lung tuberculosis screening, RadiSen AXIR, when analyzed on Dataset 1A2A, showed diagnostic performance closely comparable to that of B Readers. It achieved an area under the receiver operating characteristic curve (AUROC) of 0.9885, sensitivity of 0.9157, and specificity of 0.9731 at a threshold of 0.30.

Referring to The Target Product Profiles (TPPs) for a rapid non-sputum-based biomarker test for tuberculosis detection by the World Health Organization (WHO), as shown in Table 7, it can be observed that each test scenario has different criterions for sensitivity and specificity.

Table 7 TPP for a rapid non-sputum-based biomarker test for tuberculosis detection

	Minimal Requirements		Optimal Requirements	
	Sensitivity	Specificity	Sensitivity	Specificity
Smear-replacement test	Overall >80% Positive >99% Negative >60%	98%	Overall >95% Positive >99% Negative >68%	98%
Non-sputum based biomarker test	Overall >65% Positive >98%	98%	Positive >98% Negative >68%	98%
Triage test	90%	70%	95%	80%

Reference: https://academic.oup.com/jid/article/211/suppl_2/S29/2490781

The Minimal Requirements and Optimal Requirements in the WHO TPPs (Target Product Profiles) outline the minimum and ideal thresholds for sensitivity and specificity that such a test should meet.

The Minimal Requirements indicate the minimum acceptable level of sensitivity and specificity that the test should achieve to be considered effective for tuberculosis detection. These criteria serve as a baseline standard for performance.

The Optimal Requirements represent the desired ideal performance levels for sensitivity and specificity. Meeting or exceeding these requirements would indicate a highly accurate and reliable test for tuberculosis detection.

The results of tuberculosis screening using RadiSen AXIR at different thresholds compared to the WHO TPP criteria, with the highest threshold that yields the closest specificity to the WHO TPP, are presented in Table 8.

Table 8 Sensitivity and Specificity Values at Different Thresholds according to WHO TPP Criteria

Threshold	Sensitivity	Specificity
0.8542	0.7409	0.9991
0.7946	0.8061	0.9809
0.3219	0.9110	0.9731
0.0075	0.9811	0.8397
0.0059	0.9866	0.8015

Furthermore, when comparing the results obtained with the WHO TPP criteria, it was found that RadiSen AXIR met the requirements for the Triage test (for both the Minimal Requirements and Optimal Requirements) and the Non-sputum based biomarker test (for the Minimal Requirements criteria). The test outcomes are summarized in Table 9.

Table 9 Results of Tuberculosis Screening by RadiSen AXIR
according to WHO TPP Criteria.

	Minimal Requirements	Optimal Requirements
Smear-replacement test	Pass	Not pass
Non-sputum based biomarker test	Pass	Not pass
Triage test	Pass	Pass

Supplementary Table

Figure 1 illustrates the Receiver Operating Characteristic (ROC) curve which can be used to visualize the performance of a classifier at various thresholds. By adjusting the threshold, one change the trade-off between sensitivity and specificity. Table S1 details different sensitivity and specificity values across varying classification thresholds for abnormalities, respectively.

Table S1 Sensitivity and Specificity Across Varying Classification Thresholds for Tuberculosis. (Manufacturer's recommended threshold value is 0.30)

Threshold	Sensitivity	Specificity
0.9900	0.1809	0.9992
0.9700	0.4354	0.9958
0.9501	0.5507	0.9908
0.8703	0.7414	0.9875
0.8495	0.7586	0.9850
0.8196	0.7848	0.9817
0.7946	0.8061	0.9809
0.7611	0.8265	0.9792
0.6866	0.8584	0.9792
0.5907	0.8789	0.9750
0.5393	0.8863	0.9676
0.5107	0.8936	0.9676
0.5054	0.8953	0.9667
0.4742	0.9002	0.9667
0.4551	0.9034	0.9626
0.4152	0.9034	0.9601
0.4034	0.9067	0.9584
0.3865	0.9083	0.9576
0.3528	0.9157	0.9576
0.3308	0.9190	0.9509

Threshold	Sensitivity	Specificity
0.3219	0.9214	0.9484
0.2748	0.9313	0.9484
0.2579	0.9337	0.9459
0.2303	0.9411	0.9409
0.2020	0.9452	0.9351
0.1929	0.9468	0.9343
0.1716	0.9493	0.9343
0.1600	0.9509	0.9334
0.1444	0.9534	0.9334
0.1268	0.9542	0.9243
0.1086	0.9583	0.9185
0.0961	0.9599	0.9176
0.0595	0.9624	0.9151
0.0421	0.9632	0.9010
0.0395	0.9673	0.9002
0.0322	0.9714	0.8943
0.0290	0.9714	0.8869
0.0175	0.9746	0.8727
0.0107	0.9804	0.8386
0.0071	0.9853	0.8012

Typically, at a lower threshold, the model has high sensitivity but lower specificity. This means it correctly identifies most of the positives but also produces more false positives. At a medium threshold, there's a balance between sensitivity and specificity, which might be a good choice depending on the context. At a higher threshold, the model has high specificity but lower sensitivity. This is suitable when you want to be very certain about the positives but risk missing some.

Choosing the optimal threshold depends on the specific requirements of the task at hand. In some applications, high sensitivity might be more important, while in others, high specificity may be preferred.

4th April 2024