This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

Trustworthy TB Screening: Backdoor Vulnerability and Lightweight Defense in Chest X-Ray AI

Wisdom Jonathan, Dabrilagha Fidelis, Emevor Edisemokumor Collins

This supplementary material reports additional analyses from the same experimental pipeline as the main manuscript. It is intended to provide run-level results, paired prediction changes, trigger-parameter sensitivity, and control-experiment results that are not presented in the main paper.

1 Repeated poisoning: run-level results

Each poisoning condition was independently trained using five seeds (42, 123, 2024, 7, and 99). The table below provides the individual run results underlying the mean ± SD summaries reported in the manuscript. Poisoned subsets were independently sampled at each rate and seed.

Seed	Poison (%)	Poisoned TB images	Clean Acc. (%)	Clean Sens. (%)	Clean Spec. (%)	AUROC (%)	Conv. ASR (%)	Clean-correct ASR (%)	Flip rate (%)	Mean clean TB prob.	Mean triggered TB prob.	Δ TB prob.
42	5	20	98.83	97.33	99.33	99.91	2.67	0.00	0.00	0.9328	0.9243	-0.0086
123	5	20	99.33	99.33	99.33	99.98	0.67	0.00	0.00	0.9728	0.9693	-0.0035
2024	5	20	99.50	98.67	99.78	100.00	2.00	0.68	0.67	0.9677	0.9601	-0.0076
7	5	20	99.17	100.00	98.89	100.00	2.00	2.00	2.00	0.9810	0.9420	-0.0390
99	5	20	99.33	97.33	100.00	99.97	3.33	0.68	0.67	0.9470	0.9318	-0.0153
42	10	40	98.83	96.00	99.78	99.97	12.67	9.03	8.67	0.8895	0.7871	-0.1024
123	10	40	98.17	92.67	100.00	99.99	50.67	46.76	43.33	0.8634	0.5007	-0.3627
2024	10	40	99.33	98.00	99.78	99.96	63.33	62.59	61.33	0.9370	0.3968	-0.5402
7	10	40	99.67	100.00	99.56	99.98	55.33	55.33	55.33	0.9796	0.4865	-0.4931
99	10	40	98.17	93.33	99.78	99.84	70.67	68.57	64.00	0.9122	0.3362	-0.5760
42	20	80	98.50	96.00	99.33	99.89	80.67	79.86	76.67	0.8915	0.1943	-0.6972
123	20	80	99.17	96.67	100.00	99.97	86.67	86.21	83.33	0.9179	0.1561	-0.7618
2024	20	80	95.83	83.33	100.00	99.71	87.33	84.80	70.67	0.7936	0.1171	-0.6765
7	20	80	97.33	90.00	99.78	98.91	68.00	64.44	58.00	0.8664	0.3374	-0.5290
99	20	80	98.67	95.33	99.78	99.91	79.33	78.32	74.67	0.8905	0.2327	-0.6578

2. Paired probability analysis

Paired analysis evaluated the same TB-positive test images before and after trigger insertion. This analysis helps distinguish trigger-induced probability shifts from errors already present on the clean image.

5% poisoning: mean clean TB probability = 0.9603 ± 0.0198; mean triggered TB probability = 0.9455 ± 0.0189; paired change = −0.0148 ± 0.0142.

10% poisoning: mean clean TB probability = 0.9163 ± 0.0446; mean triggered TB probability = 0.5014 ± 0.1733; paired change = −0.4149 ± 0.1925.

20% poisoning: mean clean TB probability = 0.8720 ± 0.0475; mean triggered TB probability = 0.2075 ± 0.0844; paired change = −0.6644 ± 0.0853.

3. Trigger-with-correct-label control

A complementary control added the same trigger to selected TB-positive training images while retaining their correct TB-positive labels. Across 5%, 10%, and 20% training exposure, triggered TB test images remained classified as TB-positive, with an ASR of 0.00% at each exposure level and a TB prediction rate of 100%. This supports the interpretation that systematic TB-to-Non-TB failure required the learned association between the trigger and the incorrect Non-TB label.

4. Additional trigger-parameter experiments

Additional trigger-parameter experiments were performed using the selected 20% poisoned model. These experiments evaluated the effect of changing trigger size and opacity while retaining the original bottom-right position.

Trigger variant	Size	Alpha	Position	ASR (%)
Original	24 px	0.25	Bottom-right	80.67
Smaller	16 px	0.25	Bottom-right	74.00
Larger	32 px	0.25	Bottom-right	63.33
Lower opacity	24 px	0.15	Bottom-right	63.33
Higher opacity	24 px	0.40	Bottom-right	80.00

5. Trigger-position experiments

To assess whether the backdoor response depended on trigger location, the same 24-pixel trigger with $\alpha = 0.25$ was evaluated at different positions on the X-ray. The poisoned model had been trained using the bottom-right position; the other positions were used only during evaluation.

Trigger position	Size	Alpha	ASR (%)
Bottom-right (original)	24 px	0.25	80.67
Bottom-left	24 px	0.25	84.67
Top-right	24 px	0.25	11.33
Top-left	24 px	0.25	11.33
Centre	24 px	0.25	4.67

The bottom-right position produced an ASR of 80.67%, while the bottom-left position produced 84.67%. In contrast, top-right, top-left, and centre placement reduced ASR to 11.33%, 11.33%, and 4.67%, respectively. These results suggest that the observed backdoor behaviour is strongly dependent on trigger location.

6. Grad-CAM attention quantification

Five representative TB-positive test cases were examined to supplement the qualitative Grad-CAM example in the main manuscript. The fraction of Grad-CAM attention within the original trigger region was quantified before and after border cropping.

Representative case	Trigger-region attention before crop (%)	Attention remaining in original trigger region after crop (%)
Case 1	15.98	0.00
Case 2	15.24	0.27
Case 3	14.81	0.96
Case 4	14.35	0.97
Case 5	15.04	1.15

Across these five representative cases, trigger-region attention was approximately 14–16% before cropping and fell to approximately 0–1% within the original trigger region after cropping. These maps provide supportive evidence of localized trigger-associated attention.

S8. Reproducibility information

The repeated experiments used seeds 42, 123, 2024, 7, and 99. Training used DenseNet-121, batch size 32, Adam optimization, learning rate 1×10^{-4}, five fine-tuning epochs, and cross-entropy loss. The implementation was run with PyTorch 2.11.0+cu128, torchvision 0.26.0+cu128, NumPy 2.1.3, scikit-learn 1.6.1, and Python 3.13.15 on Linux using a Tesla T4 GPU.