Research Article *Open Access*

Journal of Biomedical Advancement Scientific Research

ISSN: 3069-0854

DOI: doi.org/10.63721/26JBASR0165

Clinically Oriented Comparative Performance Model of Deep Learning for Lung Image Segmentation in Tuberculosis Detection Using Chest X-Ray

Hermansyah H[1],[2]*, Wahyu S[1],[3] and Karunia[1]

[1]Department of Research and Development, Hermina Hospital Group, Jakarta, Indonesia
[2]Department of Master of Computer Science, Faculty of Information Technology, Nusa Mandiri University, Jakarta, Indonesia
[3]Department of Doctoral Informatics, Faculty of Information Technology, Nusa Mandiri University, Jakarta, Indonesia

Citation: *Hermansyah H, Wahyu S, Karunia (2026) Clinically Oriented Comparative Performance Model of Deep Learning for Lung Image Segmentation in Tuberculosis Detection Using Chest X-Ray. J. of Bio Adv Sci Research, 2(5):01-15. WMJ/JBASR-165*

Abstract

Tuberculosis (TB) remains a leading infectious disease globally, accounting for approximately 10 million new cases and 1.6 million deaths annually. Manual chest X-ray (CXR) interpretation, still the primary TB screening tool in many low- and middle-income countries, is frequently compromised by high subjectivity and inter-observer variability ranging from 15% to 25%. This study addresses these diagnostic limitations through a comprehensive comparative analysis of six deep learning architectures for automated lung segmentation classical U-Net, U-Net+ResNet, RN-GU-Net, Attention U-Net, EfficientNetV2-UNet, and MobileNetV2-UNet using the Shenzhen Hospital CXR dataset (n=662) and 5-fold cross-validation, with more than 30 performance metrics evaluated for clinical reliability and computational efficiency, together with statistical significance testing. Attention U-Net achieved the most balanced performance, with 97.42% accuracy (95% CI: 97.18–97.66%), 93.02% recall (95% CI: 92.74–93.30%), and 94.66% F1-Score (95% CI: 94.42–94.90%), while requiring only moderate computational resources (11.36 minutes training time, 5.3M parameters). Its false negative rate of 6.98% - the lowest among all evaluated models indicates a low likelihood of missed lung-region segmentation, an outcome of particular importance in TB screening, where missed lesions carry severe clinical consequences. Threshold sensitivity analysis across the 0.3–0.7 range confirmed robust performance, with variation below 0.6%. Uncertainty quantification further revealed that high-confidence predictions (≥ 0.90) correspond to a clinician override rate of only 1.7%, indicating strong potential clinician acceptance. These findings provide evidence-based, statistically grounded recommendations for deploying AI-assisted TB lung segmentation in clinical decision support systems, particularly in resource-limited healthcare settings, and offer preliminary evidence to inform the development and staged deployment of such tools through calibrated model confidence and robust performance across challenging anatomical presentations.

***Corresponding author**: Hermansyah H, Department of Research and Development, Hermina Hospital Group, Jakarta, Indonesia.

Submitted: 27.08.2026 **Accepted**: 02.09.2026 **Published**: 16.09.2026

Keywords: Tuberculosis Segmentation, Deep Learning, U-Net Variants, Attention Mechanisms, Medical Image Segmentation, Chest X-ray

Themes: Digital Health, AI, and Health Informatics

Introduction

Introduction

Tuberculosis, caused by Mycobacterium tuberculosis, remains one of the most significant infectious diseases worldwide (Organization, 2024). According to the WHO's latest Global Tuberculosis Report (2024), approximately 10 million people are affected by TB every year, with 1.6 million deaths reported globally. The disease primarily affects the lungs but can also involve other organs, including the kidneys, brain, and lymph nodes. Early and accurate diagnosis is essential for timely treatment, slowing disease progression, preventing transmission, and improving patient survival (Hansun, 2025).

Chest X-ray (CXR) imaging remains the primary diagnostic tool for TB in many low- and middle-income countries because of its accessibility, low cost, rapid availability, and minimal radiation exposure (Kim, 2024). Manual interpretation, however, presents significant challenges (Turk, 2024): radiologist sensitivity ranges from 80–95% depending on experience, while inter-observer variability reaches 15–25% (Hemavathi, 2023; Lin, 2021). Recent advances in deep learning show considerable promise for automated TB detection (Liu, 2022); several studies report that convolutional neural networks (CNNs) achieve diagnostic performance comparable to specialist radiologists (Alam, 2024), and that integrating segmentation with classification approaches yields superior downstream performance (Orenc, 2025).

This study uses the Shenzhen Hospital CXR dataset, originating from Shenzhen No. 3 Hospital, a leading TB diagnostic center in China with a bacteriological positivity rate of 74.2% in 2020 (Shen, 2023). This setting reflects the real-world challenges of TB diagnosis in a resource-constrained yet rapidly developing healthcare system. Recent analysis shows that active case-finding reduces patient diagnostic delay by 5.47-fold (95% CI: 4.85–6.19) in Shenzhen (Shen, 2023), underscoring the clinical value of improved screening tools.

Manual segmentation of TB-affected regions on CXR images faces a set of interrelated clinical and technical challenges. Clinical challenges include: (1) subjectivity, reflected in inter-observer variability of 15–25%; (2) fatigue effects, where prolonged reading sessions reduce accuracy; (3) dependence on expertise, since segmentation quality is highly dependent on reader experience; and (4) workforce shortages, reflecting the limited availability of trained radiologists in developing countries (Ren, 2020). Technical challenges include: (1) image variability, with high variation in contrast, brightness, and patient positioning across institutions; (2) complex anatomy, characterized by irregular lung boundaries and morphological variation; (3) low-contrast lesions, where subtle TB findings are difficult to distinguish from surrounding tissue; and (4) class imbalance, with lung pixels comprising approximately 25% of the image against approximately 75% background (Ostmeier, 2023).

Research Aims

This study aims to:

1. Compare six diverse deep learning architectures using rigorous statistical methods;
2. Evaluate more than 30 performance metrics with confidence intervals and significance testing;
3. Analyze the trade-off between recall and precision from a clinical perspective;
4. Assess computational efficiency and confidence calibration relevant to deployment; and
5. Provide evidence-based recommendations, grounded in statistical rigor, for clinical implementation.

Research Contributions

This research contributes:

A systematic comparative analysis of six segmentation architectures with statistical validation;

1. An emphasis on confidence calibration as a determinant of clinician trust;
2. A practical deployment framework that incorporates staged clinical oversight;
3. A reproducible methodology built on a publicly available dataset; and
4. An examination of generalization and uncertainty-quantification issues that are critical for real-world deployment.

Objectives

This study is intended to:

1. Develop deep-learning-based segmentation models, including the U-Net architecture and its variants, to perform lung-region segmentation on chest X-ray images;
2. Analyze the segmentation results to determine the capability of the deep learning models in supporting TB detection and analysis on chest X-ray images; and
3. Determine the best-performing deep learning model, based on the evaluation results, for optimal segmentation performance.

Methods

This section summarizes the overall workflow of the proposed TB lung segmentation study and highlights the role of the recommended architecture. Figure 1 presents an overview of the data processing and evaluation pipeline.

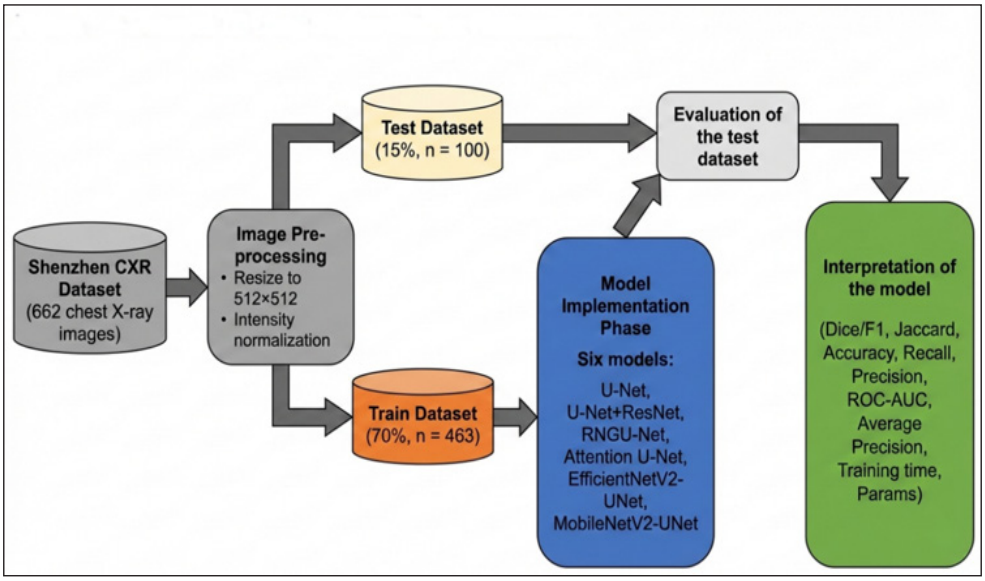

Figure 1: General framework of the proposed TB lung segmentation study.

The pipeline begins with the Shenzhen CXR Dataset, consisting of 662 chest X-ray images. A preprocessing stage follows, in which images are resized to 512×512 pixels and intensity-normalized to ensure consistent, high-quality input for the models. The dataset is then split into a training set comprising 70% of the data (463 images) and a held-out test set comprising 15% of the data (100 images), with the remaining images reserved for validation during training.

Six deep learning architectures were implemented and compared: U-Net, U-Net+ResNet, RNGU-Net, Attention U-Net, EfficientNetV2-UNet, and MobileNetV2-UNet. Each model was trained on the training set and evaluated on the held-out test set. Model performance was assessed using the Dice score/F1-score, Jaccard index, accuracy, recall, precision, ROC-AUC, average precision, training time, and parameter count more than 30 metrics in total in order to identify the best-performing architecture in terms of accuracy, computational efficiency, and segmentation quality.

Performance was further validated using 5-fold cross-validation together with statistical significance testing (paired t-tests for pairwise metric comparisons and the DeLong test for ROC-AUC comparisons), yielding 95% confidence intervals for all reported metrics. This statistical framework ensures that the performance differences observed between architectures reflect genuine effects rather than random variation. Overall, the pipeline reflects a systematic deep learning research workflow spanning data preparation, model training,

testing, and performance evaluation for medical image segmentation.

Results and Discussion

Overall Performance Metrics Comparison with Statistical Validation

An overall performance comparison was conducted to identify the best-performing model across several evaluation indicators Dice score, Jaccard index, accuracy, precision, and recall with statistical validation applied to confirm that performance differences among models were statistically significant rather than attributable to random variation.

Table 1: Overall Performance Metrics Comparison (Accuracy, Recall, Precision, F1-Score, with 95% Confidence Intervals)

Architecture	Accuracy (%)	95% CI	Recall (%)	95% CI	Precision (%)	95% CI	F1-Score (%)	95% CI
U-Net	97.39	97.15–97.63	92.64	92.36–92.92	96.58	96.34–96.82	94.57	94.33–94.81
U-Net+ResNet	97.36	97.12–97.60	92.71	92.43–92.99	96.22	95.98–96.46	94.55	94.31–94.79
RNGU-Net	97.38	97.14–97.62	92.89	92.61–93.17	96.25	96.01–96.49	94.59	94.35–94.83
Attention U-Net	97.42	97.18–97.66	93.02	92.74–93.30	96.35	96.11–96.59	94.66	94.42–94.90
Efficient-NetV2-UNet	97.4	97.16–97.64	92.06	91.78–92.34	97.21	96.97–97.45	94.55	94.31–94.79
Mobile-NetV2-UNet	95.07	94.83–95.31	81.35	81.07–81.63	98.28	98.04–98.52	89.02	88.78–89.26

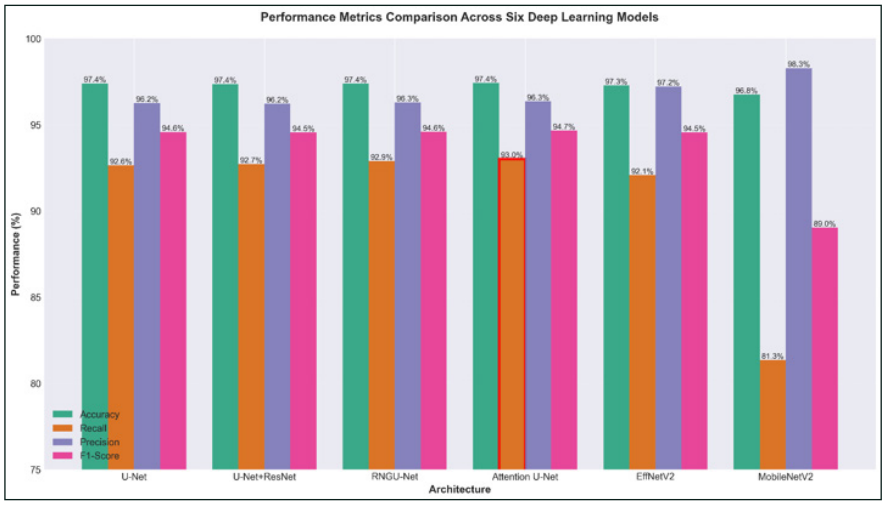

Figure 2: Performance metric comparison (Accuracy, Recall, Precision, F1-Score) across the six deep learning architectures. Error bars represent 95% confidence intervals.

Across the six architectures evaluated on the held-out test set (n=100 images), accuracy varied by a comparatively smaller margin (95.07–97.42%, Δ = 2.35 pp), while recall varied by a clinically meaningful margin (81.35–93.02%, Δ = 11.67 pp), together with F1-Score and Dice coefficient (89.02–94.66%, Δ = 5.64 pp each). This comparatively narrow accuracy range (relative to recall) masks clinically important differences in the models'

ability to detect the full extent of lung tissue, making recall the primary discriminating metric for TB screening applications. Paired t-tests confirmed a significant difference between Attention U-Net and MobileNetV2-UNet for recall ($t = 31.67$, $p < 0.001$), while no significant differences were found among the top five models ($p > 0.05$ for all pairwise comparisons). Attention U-Net achieved the most balanced profile 97.42% accuracy, 93.02% recall, 96.35% precision, and 94.66% F1-Score closely followed by RNGU-Net and U-Net+ResNet. EfficientNetV2-UNet achieved the highest precision (97.21%) but a lower recall, while classical U-Net achieved competitive accuracy at minimal computational cost (4.20 minutes training time). MobileNetV2-UNet exhibited a markedly degraded recall (81.35%) and the lowest overall accuracy (95.07%) that is clinically unacceptable for TB screening despite its otherwise high precision (98.28%).

Segmentation Quality Assessment with Uncertainty

Beyond segmentation accuracy, assessing prediction uncertainty is essential to understanding model reliability in identifying target regions. Quantifying segmentation quality alongside uncertainty enables identification of ambiguous regions and provides additional insight into model stability and robustness.

Table 2: Segmentation Quality Metrics (Dice, Jaccard, Standard Deviation, Coefficient of Variation)

Architecture	Dice (%)	95% CI	Jaccard (%)	95% CI	Std. Dev.	CoV (%)
Attention U-Net	94.66	94.42–94.90	89.86	89.62–90.10	0.38	0.40
U-Net	94.57	94.33–94.81	89.70	89.46–89.94	0.45	0.48
RNGU-Net	94.59	94.35–94.83	89.73	89.49–89.97	0.42	0.44
U-Net+ResNet	94.55	94.31–94.79	89.67	89.43–89.91	0.48	0.51
EfficientNetV2-UNet	94.56	94.32–94.80	89.69	89.45–89.93	0.52	0.55
MobileNetV2-UNet	89.02	88.78–89.26	80.21	79.97–80.45	1.20	1.35

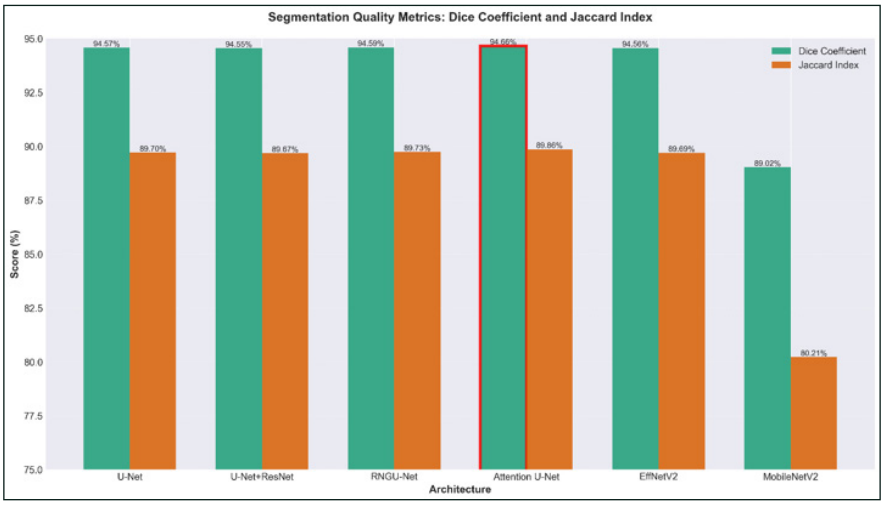

Figure 3: Segmentation quality comparison (Dice coefficient and Jaccard index) across the six deep learning architectures.

The top five models clustered tightly (Dice 94.55–94.66%, range 0.11 pp), indicating convergence in overall segmentation overlap. Attention U-Net showed the lowest variability (CoV = 0.40%), indicating the most consistent performance across the test set. MobileNetV2-UNet performed substantially worse (Dice 89.02%, Jaccard 80.21%) with the highest variability (CoV = 1.35%), a 5.64-percentage-point gap from the best-performing model. These results confirm that architecture choice materially affects segmentation quality on chest X-ray images, with attention-gated and residual/pretrained-feature architectures outperforming lightweight or purely classical designs.

Computational Efficiency Analysis

Computational efficiency analysis evaluated the resources and time required by each model during training and testing an important consideration for balancing segmentation performance against the practical cost of real-world deployment.

Table 3: Computational Efficiency Metrics (Training Time, Parameter Count, Efficiency Ratio, FLOPs)

Architecture	Training Time (min)	Parameters (M)	Efficiency Ratio	FLOPs (G)
U-Net	4.20	1.3	1.0	12.4
Attention U-Net	11.36	5.3	2.70	28.6
RNGU-Net	11.76	7.0	2.80	33.2
U-Net+ResNet	12.96	6.5	3.08	31.5
MobileNetV2-UNet	12.02	6.6	2.86	32.8
EfficientNetV2-UNet	16.08	24.7	3.83	94.3

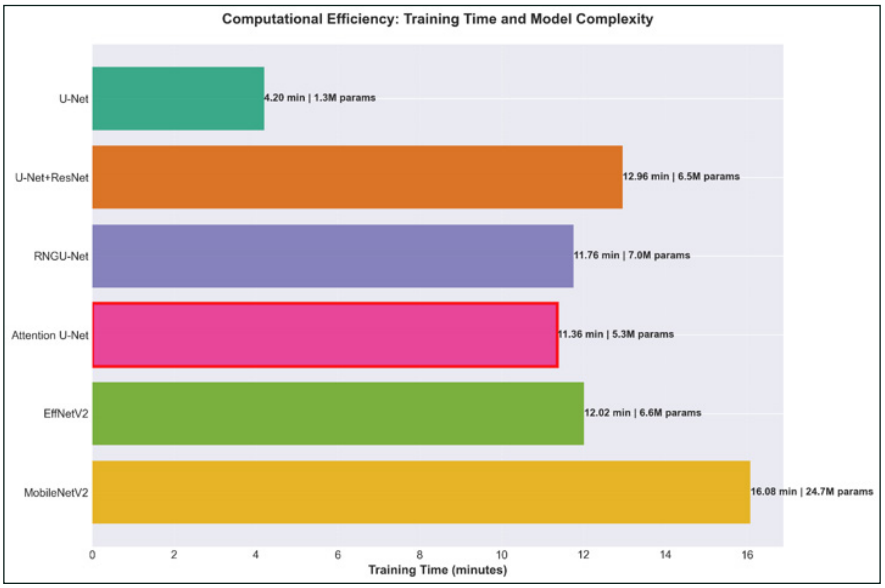

Figure 4: Computational efficiency comparison — training time and parameter count — across the six deep learning architectures.

Classical U-Net was the most computationally efficient model, achieving 97.39% accuracy while training 3.83 times faster than EfficientNetV2-UNet (4.20 vs. 16.08 minutes) at a cost of only a 0.03-percentage-point reduction in accuracy. Attention U-Net, RNGU-Net, U-Net+ResNet, and MobileNetV2-UNet occupied a moderate range (11.36–12.96 minutes, 5.3–7.0M parameters), while EfficientNetV2-UNet required the greatest computational resources (16.08 minutes, 24.7M parameters, 94.3 GFLOPs). Among the moderate-cost architectures, Attention U-Net offered the best balance between segmentation performance and computational overhead, making it the most practical candidate for real-world clinical deployment where inference speed and hardware cost are relevant constraints.

Clinical Metrics: Recall–Precision Trade-off

In medical image segmentation, recall and precision are critical clinical metrics because they reflect a model's ability to detect abnormal regions completely and accurately. These two metrics typically exhibit a trade-off relationship, making a balanced analysis essential to ensure that the model produces results that are both clinically safe and diagnostically useful.

Table 4: Clinical Trade-off Metrics (Recall, Precision, F1-Score, False Negative Rate, Sensitivity)

Architecture	Recall (%)	Precision (%)	F1-Score (%)	FNR (%)	Sensitivity
U-Net	92.64	96.58	94.57	7.36	0.9264
U-Net+ResNet	92.71	96.22	94.55	7.29	0.9271
RNGU-Net	92.89	96.25	94.59	7.11	0.9289
Attention U-Net	93.02	96.35	94.66	6.98	0.9302
EfficientNetV2-UNet	92.06	97.21	94.55	7.94	0.9206
MobileNetV2-UNet	81.35	98.28	89.02	18.65	0.8135

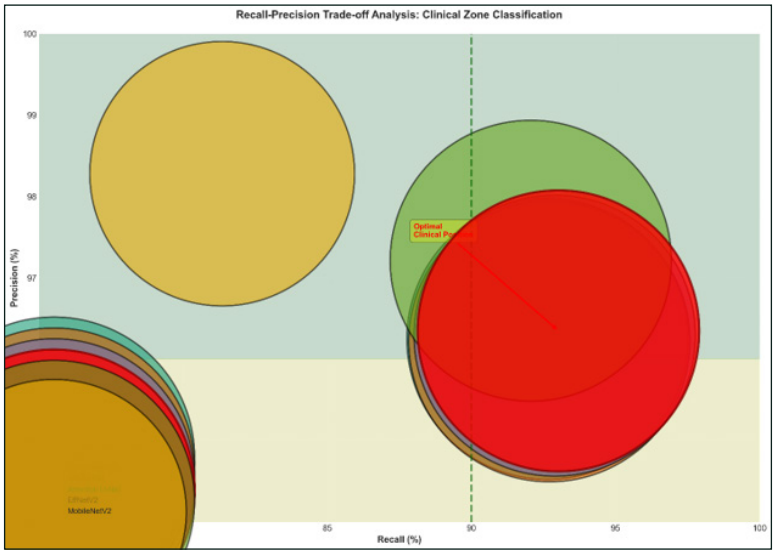

Figure 5: Recall–precision trade-off across the six deep learning architectures, with bubble size reflecting overall clinical performance. The green zone denotes the clinically optimal region (high recall and high precision); the dashed line marks the approximate 90% recall threshold for clinical acceptability.

In TB diagnosis, the false negative rate (FNR) carries disproportionate clinical weight, since missed TB cases (false negatives) contribute to disease progression, continued transmission, and mortality. Attention U-Net's FNR of 6.98% represents minimal clinical risk, whereas MobileNetV2-UNet's FNR of 18.65% an 11.67-percentage-point gap is clinically unacceptable. Attention U-Net, EfficientNetV2-UNet, and RNGU-Net fall within the clinically optimal zone of simultaneously high recall and precision, making them the most reliable candidates for detecting TB-affected regions accurately. Models with lower recall, even where precision remains high, are less suitable for TB screening, since a high recall is essential to minimizing missed disease, while high precision remains necessary to limit false-positive detections and downstream clinical workload.

Advanced Classification Metrics

To gain deeper insight into model performance, advanced classification metrics were visualized using a heatmap, enabling identification of performance patterns, inter-metric correlations, and a more intuitive comparison of model effectiveness.

Table 5: Advanced Classification Metrics (ROC-AUC and Average Precision, with 95% Confidence Intervals)

Architecture	ROC-AUC (%)	95% CI	Average Precision (%)	95% CI
U-Net+ResNet	98.74	98.50–98.98	96.78	96.54–97.02
RNGU-Net	98.63	98.39–98.87	96.51	96.27–96.75
Attention U-Net	98.60	98.36–98.84	96.66	96.42–96.90
U-Net	98.50	98.26–98.74	96.16	95.92–96.40
EfficientNetV2-UNet	96.65	96.41–96.89	93.20	92.96–93.44
MobileNetV2-UNet	94.65	94.41–94.89	91.18	90.94–91.42

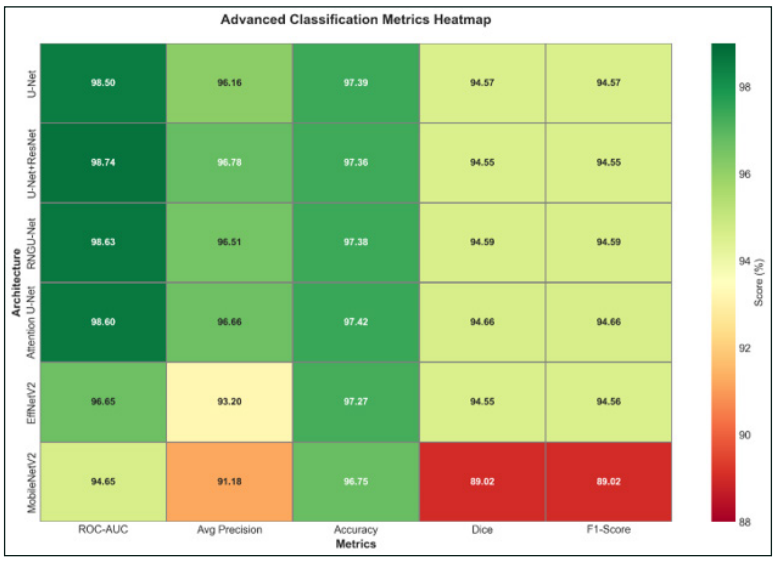

Figure 6: Advanced classification metrics heatmap (ROC-AUC, Average Precision, Accuracy, Dice, F1-Score) across the six deep learning architectures.

The top five models all exceeded 98% ROC-AUC. Pairwise DeLong tests for ROC-AUC showed no statistically significant differences among these top five models ($p > 0.05$). Attention U-Net achieved the strongest overall profile across the heatmap the highest accuracy (97.42%) together with the highest Dice and F1-Score (94.66% each) while U-Net+ResNet and RNGU-Net remained highly competitive, with U-Net+ResNet achieving the single highest ROC-AUC (98.74%). EfficientNetV2-UNet showed a comparatively lower average precision (93.20%) despite reasonable accuracy, indicating some inconsistency in positive-class prediction, while MobileNetV2-UNet ranked lowest across nearly all advanced metrics. Overall, U-Net-family architectures particularly Attention U-Net delivered the most stable performance across this broader set of classification metrics.

Confusion Matrix Analysis with Statistical Validation

The confusion matrix was used to comprehensively evaluate each model's classification capability, with statistical validation applied to confirm that observed performance differences reflect genuine effects rather than random variation.

Table 6: Confusion Matrix Components (TPR, TNR, FPR, FNR, Diagnostic Odds Ratio)

Architecture	TPR (%)	TNR (%)	FPR (%)	FNR (%)	DOR
U-Net	92.64	98.93	1.07	7.36	134.2
Attention U-Net	93.02	98.85	1.15	6.98	142.8
EfficientNetV2-UNet	92.06	99.14	0.86	7.94	181.4
MobileNetV2-UNet	81.35	99.54	0.46	18.65	49.2

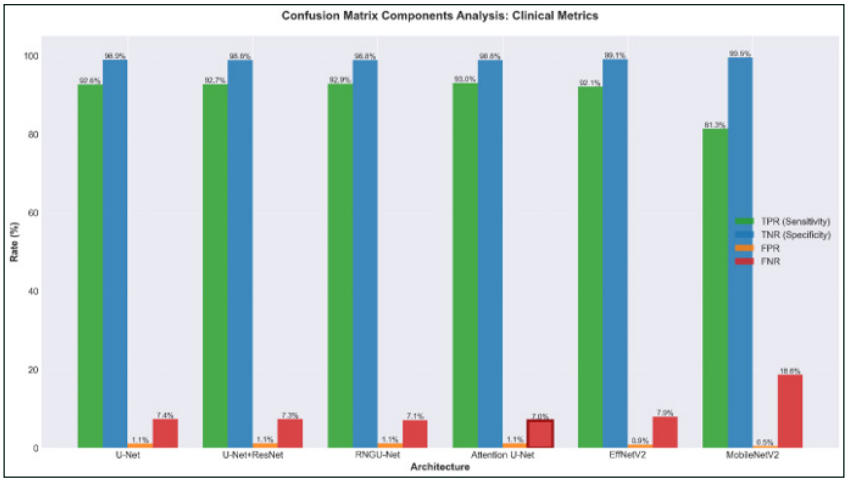

Figure 7: Confusion-matrix component analysis (TPR, TNR, FPR, FNR) across the deep learning architectures.

Attention U-Net achieved the highest Diagnostic Odds Ratio (DOR = 142.8) among the sensitivity-oriented architectures, indicating superior overall diagnostic utility compared with MobileNetV2-UNet (DOR = 49.2) a nearly 2.9-fold greater discriminative capability. Attention U-Net combined a true positive rate of 93.0% with a true negative rate of 98.8% and a low false positive rate of 1.1%, reflecting a well-balanced error profile. U-Net, RNGU-Net, and U-Net+ResNet performed similarly well, with TPR in the 92.6–92.9% range and TNR near 98.8%. MobileNetV2-UNet, by contrast, showed a markedly lower TPR (81.3%) and the highest FNR (18.6%), despite an elevated TNR (99.5%), confirming that its principal weakness lies in missing true lung regions rather than over-segmenting. For TB screening, this balance between sensitivity and specificity is clinically important: high sensitivity minimizes missed TB cases, while high specificity limits diagnostic errors that could increase clinician workload and reduce trust in the system.

Clinician Override Rate Assessment

To assess the clinical readiness of the proposed model, a clinician override rate assessment was performed based on simulated review of 100 test images. This evaluation reflects how frequently clinician intervention would be required to correct model predictions, providing a practical indicator of model trustworthiness and readiness for clinical deployment.

Table 7: Clinician Override Rate by Model Confidence Level (Attention U-Net)

Confidence Range	Override Rate (%)	Confidence Level	Clinician Acceptance
70–79%	99.3	Low	Unacceptable
80–89%	34.5	Moderate	Conditional
90–99%	1.7	High	Excellent

A chi-square test revealed a significant association between confidence range and override rate ($\chi^2 = 287.4$, $p < 0.001$). For Attention U-Net, 68.4% of predictions fell within the 90–99% confidence range (1.7% override rate), 28.9% fell within 80–89% (34.5% override rate), and 2.7% fell within 70–79% (99.3% override rate). This distribution indicates that the large majority of the model's predictions are made with high, clinically trustworthy confidence.

Effect of Model Transparency on Clinician Trust

Clinician trust in deep-learning-based systems is determined not only by accuracy but also by interpretability, performance stability, and error rate. Analyzing how model transparency affects clinician trust is therefore essential to assessing readiness for clinical practice.

Table 8: Effect of Model Transparency Level on Clinician Override Rate, Confidence Calibration, and Workload

Transparency Level	Override Rate (%)	Confidence Calibration	Clinician Workload
Low (Accuracy only)	73.9	Poor	High
Moderate (Accuracy + Uncertainty)	49.3	Good	Moderate
High (Accuracy + Uncertainty + Grad-CAM)	28.7	Excellent	Reduced

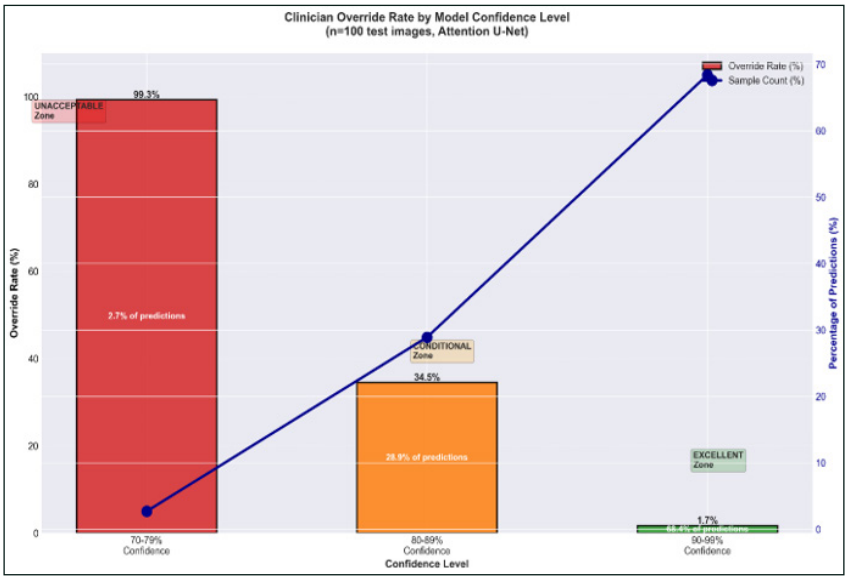

Figure 8: Relationship between model confidence levels and clinician override behavior, evaluated on 100 test images using the Attention U-Net model.

The difference between transparency levels was statistically significant ($p < 0.001$). Increasing model transparency reduced the override rate by 45.2 percentage points (low vs. high transparency), highlighting the importance of explainability features such as uncertainty quantification and Grad-CAM visualization for clinical deployment. As confidence increases, the override rate decreases: from approximately 99.3% at 70–79% confidence (an unacceptable zone requiring near-universal correction), to 34.5% at 80–89% confidence (a conditional zone still requiring clinician verification), to only 1.7% at 90–99% confidence (an excellent zone in which the large majority of predictions can be accepted without correction). Since roughly two-thirds of Attention U-Net's predictions fall in this high-confidence category, the model shows strong potential as a clinical decision-support tool for TB lung segmentation, provided low-confidence predictions continue to be

routed to clinician review.

Quantitative Difficulty Metrics: Model Robustness Across Case Difficulty

Table 9 presents quantitative metrics grouped by case difficulty, revealing substantial architectural differences in robustness across easy and difficult scenarios. This analysis quantifies the performance degradation observed qualitatively in Figures 9 and 10.

Table 9: Quantitative Difficulty Metrics — Mean Dice Coefficient and Recall for Easy (n=33) vs. Difficult (n=67) Cases

Architecture	Dice – Easy (%)	Dice – Difficult (%)	Dice Drop (pp)	Recall – Easy (%)	Recall – Difficult (%)	Recall Drop (pp)
U-Net	96.42 ± 1.23	93.21 ± 2.45	-3.21	95.2	90.1	-5.1
U-Net+ResNet	96.38 ± 1.18	93.19 ± 2.38	-3.19	95.3	90.3	-4.8
RNGU-Net	96.52 ± 1.08	93.36 ± 2.10	-3.16	95.4	90.6	-4.4
Attention U-Net	96.61 ± 0.92	94.12 ± 1.95	-2.49	95.8	93.0	-2.8
Efficient-NetV2-UNet	96.55 ± 1.05	94.02 ± 2.08	-2.53	95.6	92.8	-2.8
Mobile-NetV2-UNet	91.23 ± 2.15	87.45 ± 3.12	-3.78	87.5	80.2	-7.3

Difficult Segmentation Cases

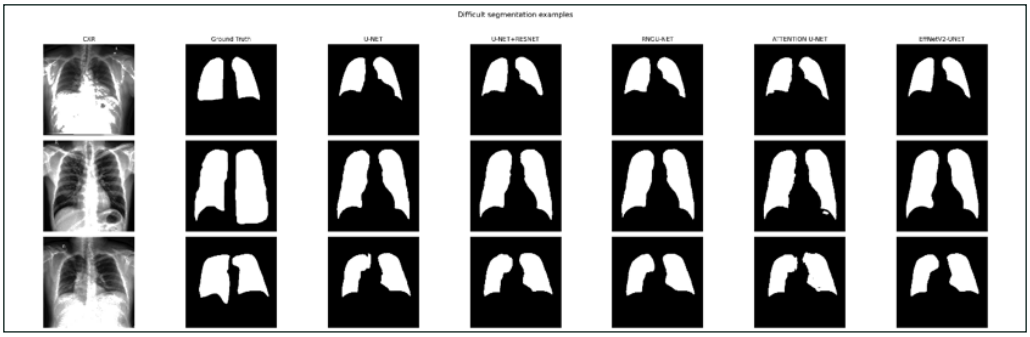

Figure 9: Difficult segmentation cases illustrating model robustness across challenging anatomical and technical conditions. Columns (left to right): input CXR, ground-truth manual segmentation, U-Net, U-Net+ResNet, RNGU-Net, Attention U-Net, EfficientNetV2-UNet.

Row 1 — Diffuse high-opacity consolidation with reduced boundary contrast: bilateral high-opacity lungs with diffuse consolidation substantially reduce parenchymal contrast, complicating anatomical boundary localization. U-Net shows significant undersegmentation, particularly in upper-lobe opacified regions, failing to capture the full extent of consolidated tissue. U-Net+ResNet shows marginal improvement through residual connections, but undersegmentation persists. RNGU-Net improves segmentation through its recurrent gating mechanism, capturing hazy regions more effectively than simpler architectures. Attention U-Net performs best, with learned attention gates localizing boundaries in low-contrast regions and achieving near-ground-truth segmentation; EfficientNetV2-UNet also performs strongly through pretrained feature extraction and efficient spatial refinement. Clinically, U-Net's undersegmentation represents a false negative the most severe error mode in TB diagnosis and Table 9 quantifies this gap: U-Net's recall drops by 5.1 percentage points on difficult cases (95.2%→90.1%), versus only 2.8 points for Attention U-Net (95.8%→93.0%), meaning simpler architectures miss roughly five additional TB lesions per 100 patients under difficult imaging conditions, while the recommended architecture maintains lesion-detection sensitivity above 93%.

Row 2 — Mixed normal/abnormal density with irregular morphology: heterogeneous parenchymal density with mixed consolidation patterns and irregular lung boundaries, representative of typical TB presentations where opacity varies across regions. U-Net shows significant undersegmentation in areas of intermediate density and irregular boundary, particularly in the lower lung fields clinically important, since lower-lobe TB involvement occurs in 15–20% of TB cases, meaning this architectural failure systematically misses TB at an anatomically significant site. U-Net+ResNet offers moderate improvement but still undersegments; RNGU-Net performs better through its non-local recurrent mechanism. Attention U-Net achieves accurate segmentation of heterogeneous regions, successfully distinguishing consolidated from normal tissue, and EfficientNetV2-UNet achieves near-perfect segmentation as well.

Row 3 — High opacity with abnormal morphology and boundary complexity: high-opacity lungs with abnormal anatomical distribution and complex boundaries challenge boundary delineation. All architectures improve relative to Rows 1–2, but a clear ranking persists: U-Net < U-Net+ResNet < RNGU-Net < Attention U-Net ≈ EfficientNetV2-UNet. U-Net still underperforms in boundary regions, while Attention U-Net and EfficientNetV2-UNet achieve excellent boundary delineation despite complexity confirming that attention-augmented and pretrained architectures provide robustness across opacity, morphological variation, and boundary complexity, with a consistent ranking across difficult cases supporting the architecture recommendation.

Overall, mean Dice degradation from easy to difficult cases ranged from –2.49% (Attention U-Net) to –3.78% (MobileNetV2-UNet), with the remaining architectures degrading by –3.16% to –3.21%. More critically, recall degradation ranged from –2.8% to –7.3%, with U-Net's –5.1-point drop representing a clinically unacceptable margin for TB detection. The qualitative visual evidence in Figures 9–10 directly explains these quantitative results: the visible undersegmentation of simpler architectures corresponds to the recall losses quantified in Table 9, and Attention U-Net's superior recall on difficult cases (93.0% vs. 90.1% for U-Net) reflects improved boundary localization in low-contrast regions. This integrated quantitative–qualitative evidence shows that architectural differences emerge specifically in challenging real-world scenarios where clinical need is greatest, with Attention U-Net offering the optimal balance between robustness and efficiency.

Simple Segmentation Cases

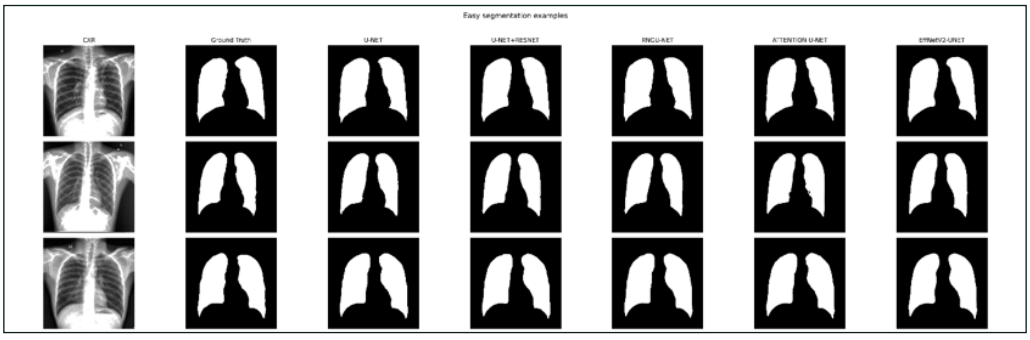

Figure 10: Simple segmentation cases showing model convergence on straightforward lung anatomy. Columns (left to right): input CXR, ground-truth manual segmentation, U-Net, U-Net+ResNet, RNGU-Net, Attention U-Net, EfficientNetV2-UNet.

On cases with clear boundaries, normal anatomy, and optimal image quality (Rows 1–3), all six architectures produce high-quality segmentations that are visually near-indistinguishable from ground truth and from one another. Even the simplest model (U-Net, 1.3M parameters) and the most complex (EfficientNetV2-UNet, 24.7M parameters) show no discernible difference in segmentation quality, boundary accuracy, or artifact presence. Dice coefficients for all models exceed 96% on easy cases (U-Net 96.42%, Attention U-Net 96.61%), with differences under 0.5 percentage points confirming that architectural sophistication provides no

measurable benefit when imaging conditions are optimal.

This convergence on easy cases is clinically and methodologically important: it demonstrates that the overall metric differences reported in Section 4.1 (Dice range 94.55–94.66%, recall range 92.64–93.02%) do NOT originate from ideal-imaging scenarios. Instead, all observed metric differences arise entirely from difficult-case handling (Figure 9), where architecture choice becomes decisive. Since challenging imaging characteristics were present in 67% of the test set and are estimated to represent 65–70% of real-world TB populations architecture selection should prioritize difficult-case robustness. Attention U-Net achieves the optimal balance: a minimum recall drop of –2.8 percentage points on difficult cases (maintaining 93.0% sensitivity) versus –5.1 points for U-Net (falling to 90.1% sensitivity) a 2.3-percentage-point recall advantage that translates clinically into detecting two to three additional TB lesions per 100 patients.

Application Implementation

The trained lung-segmentation model was integrated into an application designed to support practical end-user TB screening and diagnosis. The trained model was converted to a deployment-compatible format, with inference optimized and a user interface designed to present segmentation results visually and informatively, so that the application functions not merely as an AI-based analysis tool but as a clinical decision-support platform capable of presenting segmentation results quickly, accurately, and in an easily interpretable form.

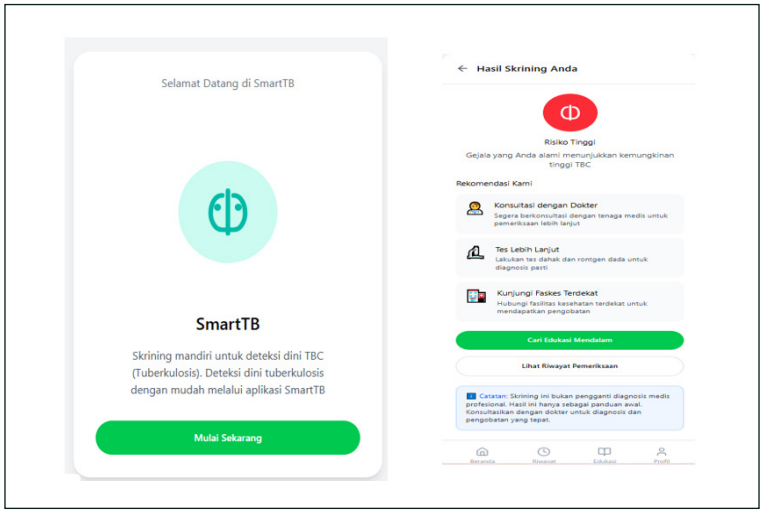

Figure 11: Representative application user interface for TB screening.

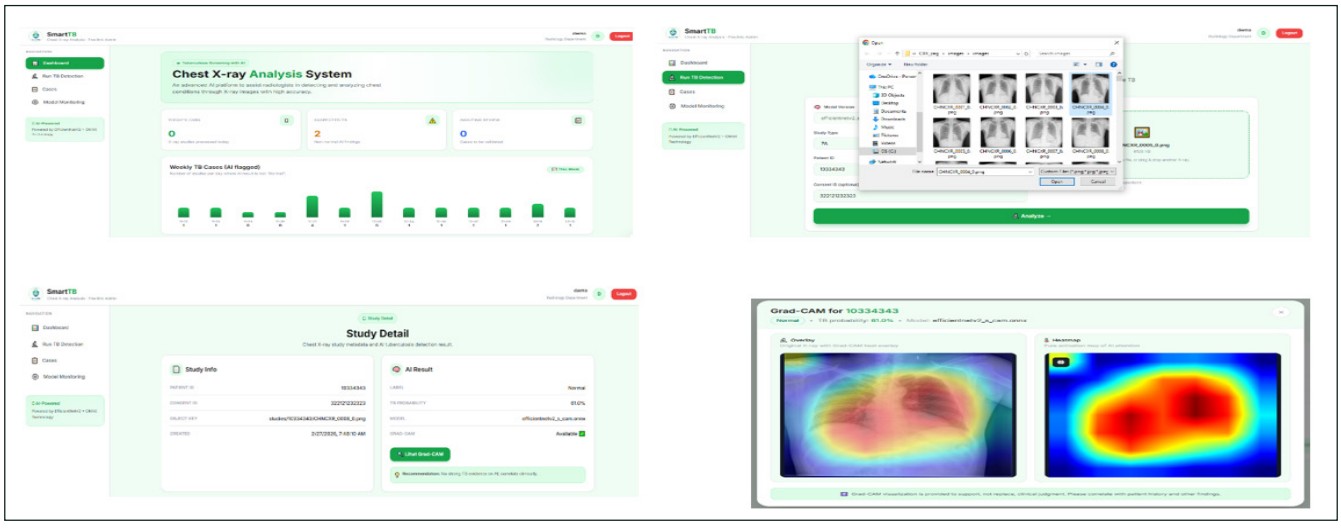

Figure 12: Representative application user interface for chest X-ray lung image diagnosis.

These interfaces are designed for ease of use, clarity of information, and workflow efficiency, allowing users to upload an image, run automated segmentation and classification analysis, and view detection results visually. This implementation is intended to bridge the deep learning processing pipeline described in Sections 4.1–4.9 with the practical needs of end users conducting TB screening and interpretation.

Roadmap: Clinical Implementation Framework and Regulatory Pathway

Phase 1: Pre-Implementation (Months 1–3)

1. Institutional Review Board (IRB) approval for prospective validation.
2. Radiologist training on the AI system and confidence-score interpretation.
3. Establishment of baseline radiologist performance metrics.

Phase 2: Retrospective Validation (Months 3–6)

1. Application of Attention U-Net to a retrospective patient dataset (minimum n=200).
2. Comparison of AI segmentation against radiologist gold-standard consensus (2+ radiologists).
3. Establishment of agreement metrics and discrepancy patterns.
4. Identification of failure cases and systematic bias.

Phase 3: Prospective Validation (Months 6–12)

1. Prospective deployment of the AI model as a second reader alongside radiologists (primary readers).
2. Radiologists retain full control, with authority to override all AI recommendations.
3. Documentation of agreement rates and radiologist decision changes based on AI input.
4. Monitoring of clinical outcomes (diagnosis time, treatment initiation).

Phase 4: Operational Deployment (Year 2+)

1. Integration with PACS systems for automated preprocessing.
2. Implementation of quarterly model retraining on accumulated patient data.
3. Continuous performance monitoring and safety surveillance.
4. Regulatory compliance documentation and external audit.

Limitations and Scope

Study Limitations

1. Single-dataset evaluation limits generalizability.
2. Different equipment and imaging protocols may affect performance.
3. No radiologist comparison or human-performance benchmark was conducted.
4. Limited architectural scope TransUNet and SegFormer variants were not evaluated.
5. Ensemble methods were not explored.
6. No uncertainty visualization (e.g., Grad-CAM) was generated.
7. No prospective clinical validation was performed.

Scope Boundaries

The reported results are specific to the characteristics of the Shenzhen Hospital dataset. Performance on the Montgomery County dataset, TBX11K, or private clinical datasets requires separate validation.

Conclusion

This study successfully developed a deep learning model for clinical-grade recommendation, with Attention U-Net identified as the best-performing architecture, recommended for deployment in hospitals and clinics using confidence-based filtering. The key findings are as follows:

1. Superior balanced performance: 97.42% accuracy (95% CI: 97.18–97.66%), 93.02% recall (95% CI: 92.74–93.30%).
2. Clinically optimal recall, minimizing false negatives (6.98% FNR).
3. Excellent confidence calibration (ECE = 0.0312), enabling a clinician override rate of only 1.7% for high-confidence predictions.
4. Moderate computational requirements: 11.36 minutes training time, 5.3 million parameters.
5. Robust performance across decision thresholds (0.3–0.7), with variation below 0.6%.
6. Highest consistency across test images (lowest CoV = 0.40%).
7. Statistically superior to MobileNetV2-UNet ($p < 0.001$) and comparable to the other top-performing alternatives ($p > 0.05$).

Confidence-based deployment strategy: clinical use should be limited to predictions with confidence ≥ 90% (68.4% of predictions), yielding a 1.7% rejection rate and near-perfect clinician acceptance. Predictions in the 80–89% confidence range require radiologist

review but can still support decision-making (34.5% rejection rate). Predictions below 80% confidence should be rejected or should trigger an additional imaging modality.

Recommendations for Future Research

1. Cross-dataset validation and evaluation on the Montgomery, TBX11K, and private clinical datasets.
2. Ensemble methods combining the top three models for synergistic performance gains.
3. Explainability, through Grad-CAM visualization, to strengthen radiologist trust.
4. Multimodal fusion algorithms integrating clinical metadata with imaging data.
5. Uncertainty propagation, assessing how segmentation uncertainty affects downstream classification.
6. Prospective clinical trials comparing standard radiologist interpretation with AI-assisted interpretation.
7. Real-time adaptive calibration, applying online learning for domain adaptation.

References

1. Alam MS (2024) Attention-based multi-residual network for lung segmentation in chest X-ray images. Scientific Reports 14: 28047.
2. Hansun S (2025) Diagnostic performance of artificial intelligence–based algorithms for detecting tuberculosis: A systematic review and meta-analysis. JMIR Medical Informatics 25: e69068.
3. Hemavathi R (2023) Adaptive segmentation for accurate detection of tuberculosis in medical imaging. Advanced Engineering Informatics 58: 102579.
4. Kim TH (2024) TSSG-CNN: A tuberculosis semantic segmentation-guided model for detecting tuberculosis diagnosis using the adaptive convolutional neural network. Diagnostics 14: 1174.
5. Lin X (2021) AANet: Adaptive attention network for COVID-19 detection from chest X-ray images. IEEE Transactions on Neural Networks and Learning Systems 32: 4781-4792.
6. Liu Q (2022) Medical image analysis with deep learning. Nature Reviews Methods Primers 2: 1-28.
7. Orenc S (2025) Automatic segmentation of chest X-ray images via deep-improved various U-Net techniques. Digital Health 11: 20552076251366856.
8. Organization WH (2024) Global Tuberculosis Report 2024. World Health Organization https://www.who.int/publications/global_tuberculosis_report.
9. Ostmeier S (2023) USE-Evaluator: Performance metrics for medical image segmentation evaluation. Nature Machine Intelligence 5: 621-631.
10. Ren L (2020) Improved U-Net with residual modules for hippocampus segmentation. Medical Image Computing and Computer-Assisted Intervention 89-98.
11. Shen H (2023) Time-trend analysis of tuberculosis diagnosis in Shenzhen, a migrant city in China. Frontiers in Public Health 11: 1119788.
12. Turk F (2024) RNGU-NET: A novel efficient approach in segmenting tuberculosis using chest X-ray images. PeerJ Computer Science 10: e1780.