Article

Not peer-reviewed version

External Validation of an Artificial Intelligence Triaging System for Chest X-Rays: A Retrospective Independent Clinical Study

André Coutinho Castilla *, Iago de Paiva D'Amorim, Maria Fernanda Barbosa Wanderley, Mateus Aragão Esmeraldo, André Ricca Yoshida, Anthony Moreno Eigier, Márcio Valente Yamada Sawamura

Posted Date: 17 October 2025

doi: 10.20944/preprints202510.1351.v1

Keywords: artificial intelligence; chest radiography; triage; external validation; deep learning;radiology; diagnostic accuracy; medical imaging

Article

External Validation of an Artificial Intelligence Triaging System for Chest X-Rays: A Retrospective Independent Clinical Study

André Coutinho Castilla [1,*], Iago de Paiva D'Amorim [2], Maria Fernanda Barbosa Wanderley [1], Mateus Aragão Esmeraldo [3], André Ricca Yoshida [1], Anthony Moreno Eigier [1] and Márcio Valente Yamada Sawamura [2]

[1] NeuralMed A2 Tecnologia Ltda., Rua Pe João Manoel 1212, Cj 41, São Paulo, SP 01411-010, Brazil
[2] Department of Radiology and Oncology, Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo (HCFMUSP), Av. Dr. Enéas de Carvalho Aguiar 255, São Paulo, SP 05403-000, Brazil
[3] Department of Radiology, Stanford University School of Medicine, Palo Alto, CA, USA
* Correspondence: andre@neuralmed.com

Abstract

Background: Chest radiography (CXR) is the most frequently performed radiological exam worldwide, yet reporting backlogs due to radiologist shortages remain critical in emergency care. Artificial intelligence triage systems may alleviate this challenge by differentiating normal from abnormal studies and prioritizing urgent cases. This study aimed to externally validate TRIA, a commercial AI-powered CXR triage algorithm (NeuralMed, São Paulo, Brazil). **Methods**: TRIA uses a two-stage deep learning approach: an image segmentation module to isolate the thoracic region, followed by a classification model trained to recognize common cardiopulmonary pathologies. The system was trained on 275,399 CXRs from multiple public and private datasets. External validation was performed retrospectively on 1,045 CXRs (568 normal, 477 abnormal) from a teaching university hospital. Ground truth was derived from radiologist reports using a large language model–assisted extraction pipeline; a subset of 300 reports was independently reviewed by a radiologist (accuracy 0.98; 95% confidence intervals (CI) 0.978–0.988). Four ensemble decision strategies for abnormality detection were compared. Performance metrics included sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUROC) with 95% CI). **Results**: The general abnormality classifier achieved strong performance (AUROC 0.911). Individual pathology models for cardiomegaly, pneumothorax, and effusion showed excellent results (AUROC 0.968, 0.955, and 0.935, respectively). The weighted ensemble demonstrated the best balance, with accuracy 0.854 (95% CI 0.831–0.874), sensitivity 0.845 (0.810–0.875), specificity 0.861 (0.830–0.887), and AUROC 0.927 (0.911–0.940). Sensitivity-prioritized methods (>0.92) produced lower specificity (<0.69). False negatives were mainly subtle or equivocal cases, although many were still flagged abnormal by the general classifier. **Conclusions**: TRIA achieved robust and balanced accuracy in distinguishing normal from abnormal CXRs. Integration into clinical workflows could reduce reporting delays, prioritize urgent cases, and improve patient safety. These findings support its clinical utility and warrant prospective multicenter validation.

Keywords: artificial intelligence; chest radiography; triage; external validation; deep learning; radiology; diagnostic accuracy; medical imaging

1. Introduction

The chest radiography (CXR) remains the most common radiological examination globally, serving as a critical first-line diagnostic tool[1]. However, the high volume of CXRs, coupled with a worldwide shortage of radiologists, often leads to significant reporting backlogs. This delay is particularly acute in emergency department (ED) settings, where timely interpretation is crucial for patient management.

 doi:10.20944/preprints202510.1351.v1

In many such settings, initial CXR readings are performed by general practitioners or ED physicians who may lack the specialized expertise of a radiologist, potentially leading to diagnostic errors or delays.

As in many other industries, artificial intelligence (AI) has emerged in healthcare as a promising approach to address persistent challenges. The advent of deep learning has enabled novel diagnostic applications, with algorithms achieving levels of accuracy comparable to those of expert clinicians in disease recognition[2]. One of the earliest and most impactful applications is triage, defined as the automated prioritization of medical cases[3]. A system capable of reliably differentiating between normal and abnormal chest radiographs (CXRs) can improve physicians ability to recognize abnormalities in CXR [4] and has potential to optimize clinical workflows by directing abnormal or critical studies for immediate evaluation by a radiologist. Furthermore, radiology represents the medical specialty with the highest number of AI software approvals, highlighting the imperative for rigorous post-implementation validation of these systems[5].

This study aims to perform an external validation of TRIA, a commercial AI-powered chest radiograph triaging system manufactured by NeuralMed (São Paulo, Brazil), registered with Brazilian Health Regulatory Agency (ANVISA) as a Class II medical device (Register 8261971001)[6]. We retrospectively evaluated its post-implementation performance on Hospital das Clínicas, Faculty of Medicine, University of São Paulo (HCFMUSP) to determine its accuracy in classifying CXRs as normal versus abnormal, a critical step for validating its clinical utility. This study supports ANVISA requirements for manufacturers to maintain clinical evaluation and post-market surveillance[7] and it is funded by Neuralmed. This study was also approved by the institution's Ethics Committee board.

2. Methods

2.1. AI Algorithm (TRIA)

The algorithm is based on a dual-pronged deep learning approach, consisting of a lung field segmentation pipeline followed by a pathology classification model. A key design choice of the methodology is the prioritization of intrathoracic findings, especially those related to the cardiopulmonary system, at the expense of musculoskeletal and spinal conditions.

2.1.1. Training Dataset

The models were developed using a large-scale, multi-source dataset of 275,399 anonymized, unique chest radiographs. The dataset was curated from public and private sources and features a multi-label annotation schema (Table 1). Part of the dataset originated from private clinical sources collected under research and commercial agreements in compliance with Brazilian General Data Protection Law [8].

The public datasets included: the Shenzhen Chest X-ray Set[9]; the Indiana University Chest X-ray Collection[10]; the JSRT database from the Japanese Society of Radiological Technology[11]; the NIH ChestX-ray14 dataset[12]; and the PadChest dataset[13]. All public datasets were used in compliance with their respective open-access agreements.

Before training, the images were proofed by medical radiology specialists using a proprietary annotation application (Figure 1). The dataset exhibits a significant class imbalance reflective of real-world pathology prevalence. For training binary classifiers, studies with definitive labels (1.0 for present, 0.0 for absent) were used, while studies with missing labels for a specific finding were excluded from the loss calculation for that task. Table 2 shows the distribution for a selection of key labels. The list of pathologies was selected based on the availability of labeled data, balanced with clinical importance for ED. The 'normal' and 'abnormal' pools included cases with pathologies beyond the specifically targeted list; these were categorized under the general "yesfinding" class.

 doi:10.20944/preprints202510.1351.v1

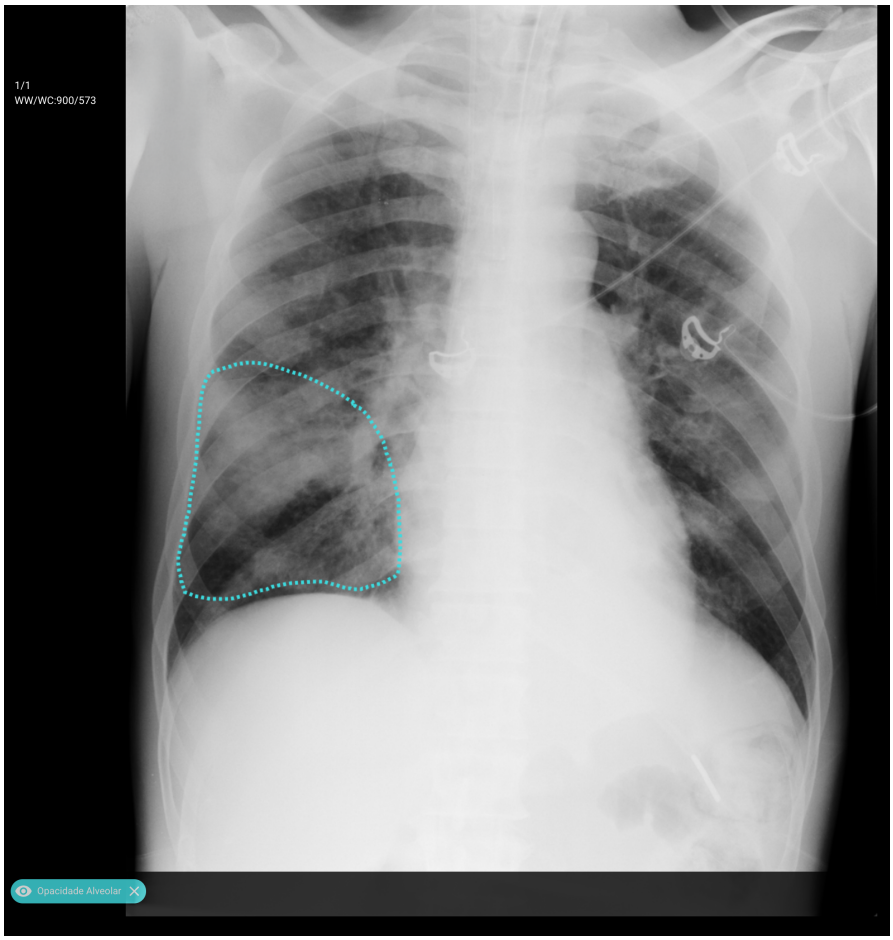

Figure 1. Images were proofed by a radiologist who manually drew a region of interest around the pathology and assigned a finding, as shown here for an alveolar opacity.

Table 1. Number of Labeled Cases per Data Source.

Source	Labeled Cases
NIH	111,783
Padchest	109,167
Private	49,866
Indiana	3783
Shenzhen	662
JSRT	246
Montgomery	138
Total Labeled Cases	**275399**

Table 2. Distribution of Selected Labels in the Training Dataset.

Pathology	Positive Cases	Negative Cases	Total Labeled
No Finding	125,046	150,266	275,312
Opacity	32,053	240,521	272,574
Cardiomegaly	15,915	255,944	271,859
Mass/Nodule	14,527	258,383	272,910
Effusion	5392	266,535	271,927
Pneumothorax	4521	267,408	271,929
Consolidation	2483	110,920	113,403

 doi:10.20944/preprints202510.1351.v1

2.1.2. Algorithm Architecture

The first stage employs a U-Net architecture with a MobileNetV2 encoder to generate a unified thoracic mask. This process uses a "tongue-shaped" mask that creates a single contiguous region of interest (ROI) encompassing both lungs and the intervening mediastinal and hilar regions. This approach simplifies binary segmentation and ensures that critical central structures are included in the analysis. Input images were resized to 512x512 pixels, and the model was trained using a composite Dice and Binary Focal loss function. Figure 2 shows an example.

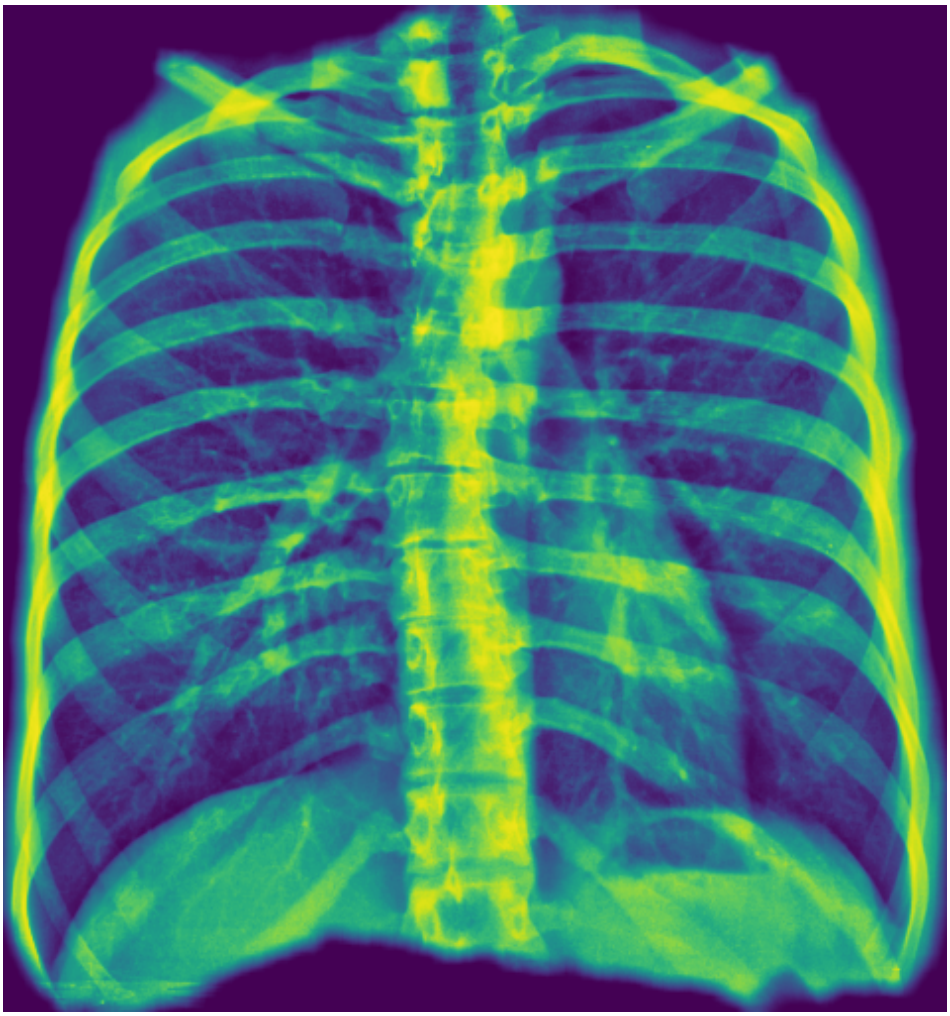

Figure 2. A chest radiography (CXR) after first-stage segmentation, showing the tongue-shaped mask.

The second stage uses an InceptionResNetV2 model, pre-trained on ImageNet, for pathology classification. The model's top layers were replaced to suit the specific task. A notable trade-off in development was resizing input images to a low resolution of 75x75 pixels to facilitate rapid training. The model was trained using class weights to counteract data imbalance.

To provide visual explanations for the pathologies identified, both Grad-CAM heatmaps and segmentation masks are generated and overlaid on the original CXR images. The Grad-CAM technique produces class activation maps, highlighting the image regions most influential in the model's prediction for a specific pathology. The resulting overlays—which can be a heatmap, bounding box, or contoured area—identify the most relevant regions, ensuring the system provides clear and interpretable visual evidence for its predictions (Figure 3).

 doi:10.20944/preprints202510.1351.v1

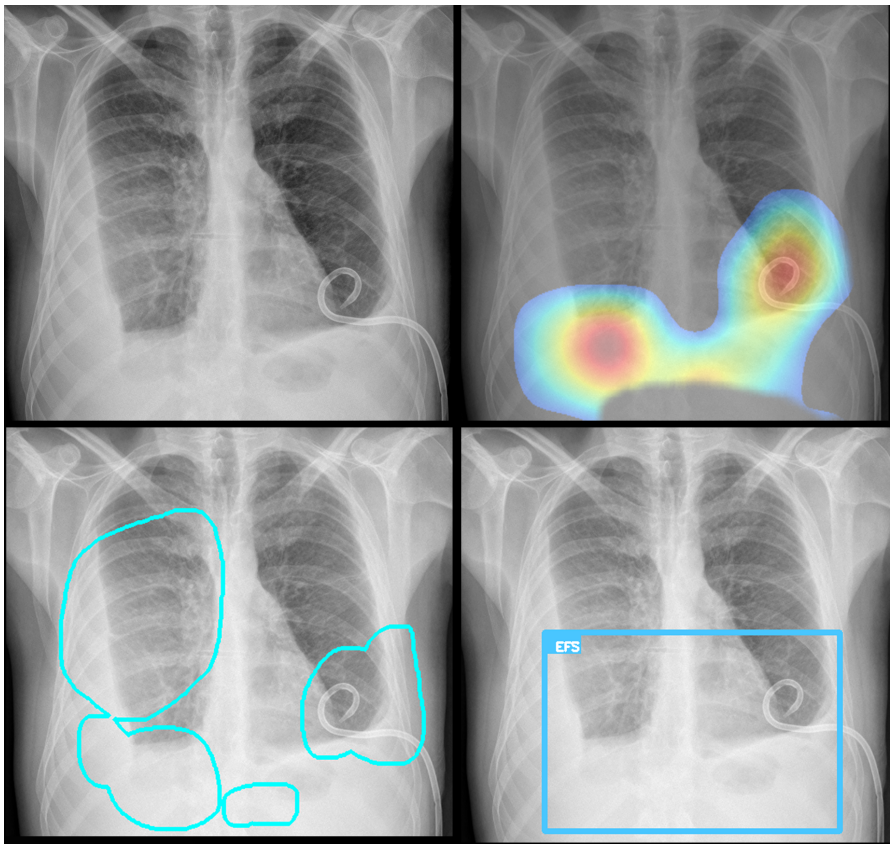

Figure 3. A true positive example of pleural effusion. The image shows three types of activation maps produced for explainability: a heatmap, a contoured line, and a bounding box. The original report also indicated suspected micronodules, a less significant finding in the overall context.

2.2. Validation Study Design

For validation, a retrospective, cross-sectional study was conducted on a dataset of 1,045 anonymized chest radiographs from HCFMUSP, over the past five years. Eligibility was restricted to adult CXR studies that included both frontal and lateral images and had an accompanying report authored by a senior radiologist. To ensure a comprehensive evaluation of the algorithm's performance on both normal and abnormal cases, a purposive sampling strategy was employed. The final cohort was constructed by selecting 568 examinations with reports describing a normal chest and 477 examinations with abnormal findings. The abnormal cases were identified by selecting reports that lacked a normal description and contained one or more of the following keywords: "pneumonia", "opacity", "cardiomegaly", "pneumothorax", "effusion", "nodule" or "mass". To create structured, binary labels for evaluation, these reports were processed using a large language model (LLM) assisted extraction method[14]. A prompt was crafted to structure the output to match the JSON format of the computer vision algorithm's results. To evaluate the normal and abnormal separation, the accuracy of each pathology classifier and the criticality ensemble were calculated. The radiographs were processed by the TRIA system that served the institution through its Picture Archiving and Communication System (PACS) using files in Digital Imaging and Communications in Medicine (DICOM) format.

2.3. Analysis

The evaluation was performed using a multi-step pipeline that began with threshold optimization for each individual pathology classifier. Receiver operating characteristic curves (ROCC) were generated, and optimal probability thresholds were selected using Youden's J statistic to balance sensitivity and specificity.

Following threshold optimization, four ensemble classification strategies were evaluated to combine the predictions of individual classifiers into a single normal/abnormal decision. The approach

 doi:10.20944/preprints202510.1351.v1

used in the commercial product is a rule-based method ("Possibility-based"). Other methods included a threshold-based approach, which flags an exam as abnormal if any pathology prediction exceeds its optimal threshold, and a weighted sum-based method, which uses a weighted sum of prediction probabilities, giving higher weight to the general "yesfinding" indicator. In addition, a hierarchical approach that prioritizes critical findings was also implemented.

The next step involved a comprehensive performance evaluation, where sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, F1 score, and area under the receiver operating characteristic curve (AUROC) were calculated for all models and ensembles. Exact binomial 95% confidence intervals (CI) were computed for all metrics, while AUROC CI were estimated via bootstrap resampling to ensure statistical rigor. Finally, an error analysis was performed by a senior radiologist who inspected the original images, reports, extracted data, and predictions for all false negatives from the six listed pathologies. Figure 4 depicts performed validation and analysis .

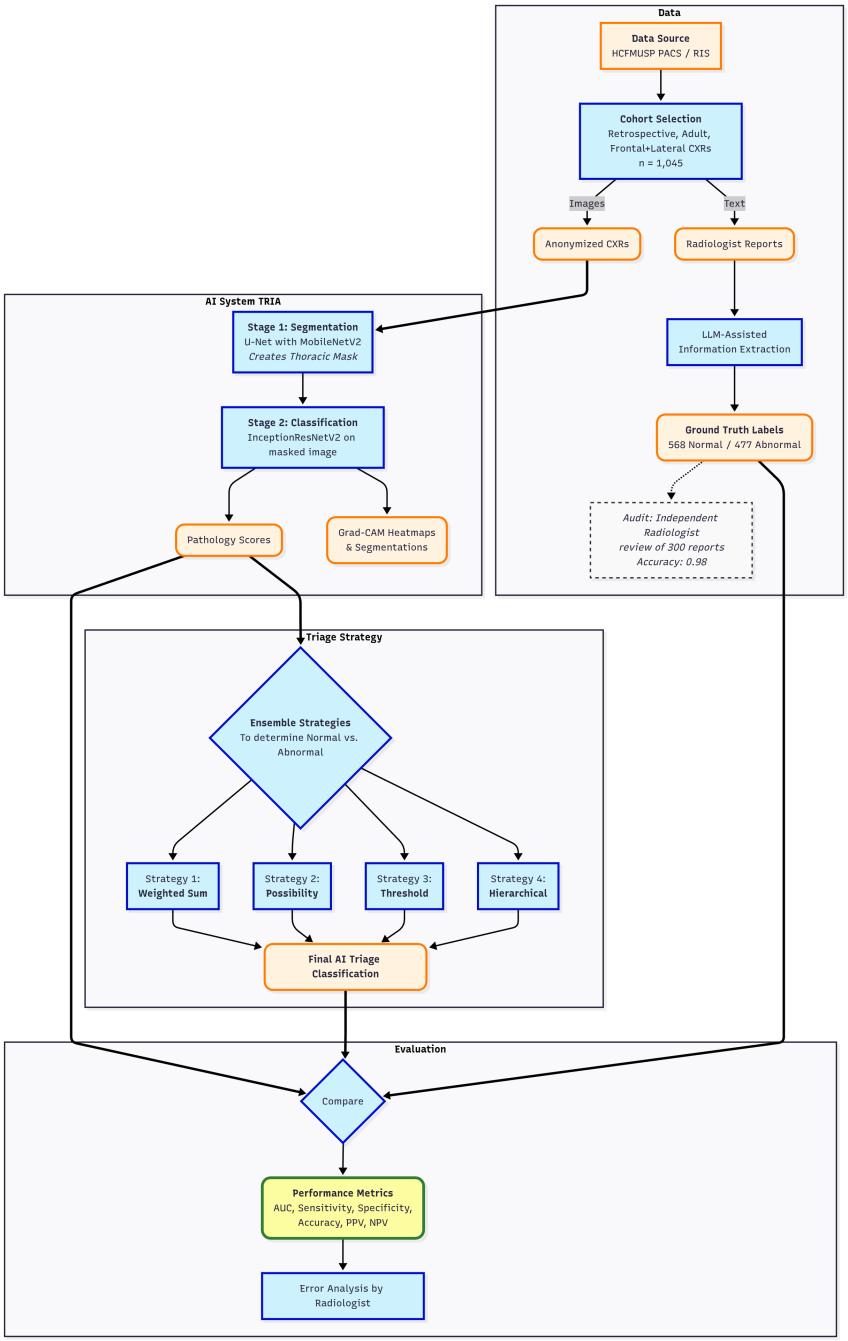

Figure 4. Study flowchart diagram.

 doi:10.20944/preprints202510.1351.v1

3. Results

3.1. Individual Classifier Performance

The general "yesfinding" model demonstrated strong performance with an AUROC of 0.911. For specific pathologies, the models for cardiomegaly, pneumothorax, and effusion showed excellent discriminative ability, achieving AUROC of 0.968, 0.955, and 0.935, respectively. In contrast, the model for bacterial pneumonia, while showing high sensitivity (0.829), had a relatively low specificity (0.656), indicating a high false-positive rate. Detailed performance metrics for each classifier are presented in Table 3.

Table 3. Performance Metrics of Individual Classifiers on the Test Set.

Pathology	Positive Cases	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	AUROC (95% CI)
Yesfinding	502	0.841 (0.806-0.870)	0.822 (0.789-0.851)	0.831 (0.807-0.852)	0.911 (0.893-0.927)
Cardiomegaly	199	0.960 (0.923-0.979)	0.881 (0.858-0.901)	0.895 (0.876-0.912)	0.968 (0.960-0.977)
Pneumothorax	52	0.904 (0.794-0.958)	0.879 (0.857-0.897)	0.880 (0.859-0.898)	0.955 (0.927-0.980)
Effusion	150	0.840 (0.773-0.890)	0.924 (0.905-0.939)	0.912 (0.894-0.928)	0.935 (0.907-0.959)
Opacity	264	0.848 (0.800-0.887)	0.733 (0.702-0.762)	0.761 (0.735-0.786)	0.865 (0.841-0.888)
Mass	56	0.786 (0.662-0.873)	0.862 (0.840-0.882)	0.858 (0.836-0.878)	0.839 (0.778-0.891)
Bacterial Pneumonia	41	0.829 (0.687-0.915)	0.656 (0.626-0.684)	0.662 (0.634-0.690)	0.775 (0.703-0.835)

Notes: AUROC = area under the receiver operating characteristic curve; CI = Confidence Interval.

3.2. Ensemble Model Performance for Abnormality Detection

To classify exams as normal or abnormal, four distinct ensemble strategies were evaluated. The weighted sum-based ensemble yielded the most balanced and superior overall performance. It achieved an accuracy of 0.854 (95% CI: 0.831-0.874), a sensitivity of 0.845 (95% CI: 0.810-0.875), a specificity of 0.861 (95% CI: 0.830-0.887), and an AUROC of 0.927 (95% CI: 0.911-0.940). This model provided the best trade-off between sensitivity and specificity. In contrast, the threshold-based and possibility-based methods achieved higher sensitivity (0.943 and 0.922, respectively) at a significant cost to specificity (0.609 and 0.685, respectively), resulting in a greater number of false positives. The comprehensive performance of all four ensemble methods is detailed in Table 4, with ROCC for the top models shown in Figure 5 and confusion matrices for all methods in Figure 6.

Table 4. Performance Comparison of Four Ensemble Strategies for Abnormality Detection.

Ensemble Method	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 Score	AUROC (95% CI)
Weighted Sum	0.854 (0.831-0.874)	0.845 (0.810-0.875)	0.861 (0.830-0.887)	0.836 (0.800-0.866)	0.869 (0.838-0.894)	0.840	0.927 (0.911-0.940)
Hierarchical	0.830 (0.806-0.851)	0.847 (0.812-0.876)	0.815 (0.781-0.845)	0.794 (0.756-0.827)	0.864 (0.832-0.890)	0.819	N/A
Possibility	0.793 (0.768-0.817)	0.922 (0.895-0.943)	0.685 (0.646-0.722)	0.711 (0.674-0.745)	0.913 (0.883-0.936)	0.803	N/A
Threshold	0.762 (0.735-0.787)	0.943 (0.919-0.961)	0.609 (0.568-0.648)	0.670 (0.633-0.704)	0.928 (0.897-0.950)	0.783	0.919 (0.901-0.935)

Notes: PPV = Positive Predictive Value; NPV = Negative Predictive Value.

 doi:10.20944/preprints202510.1351.v1

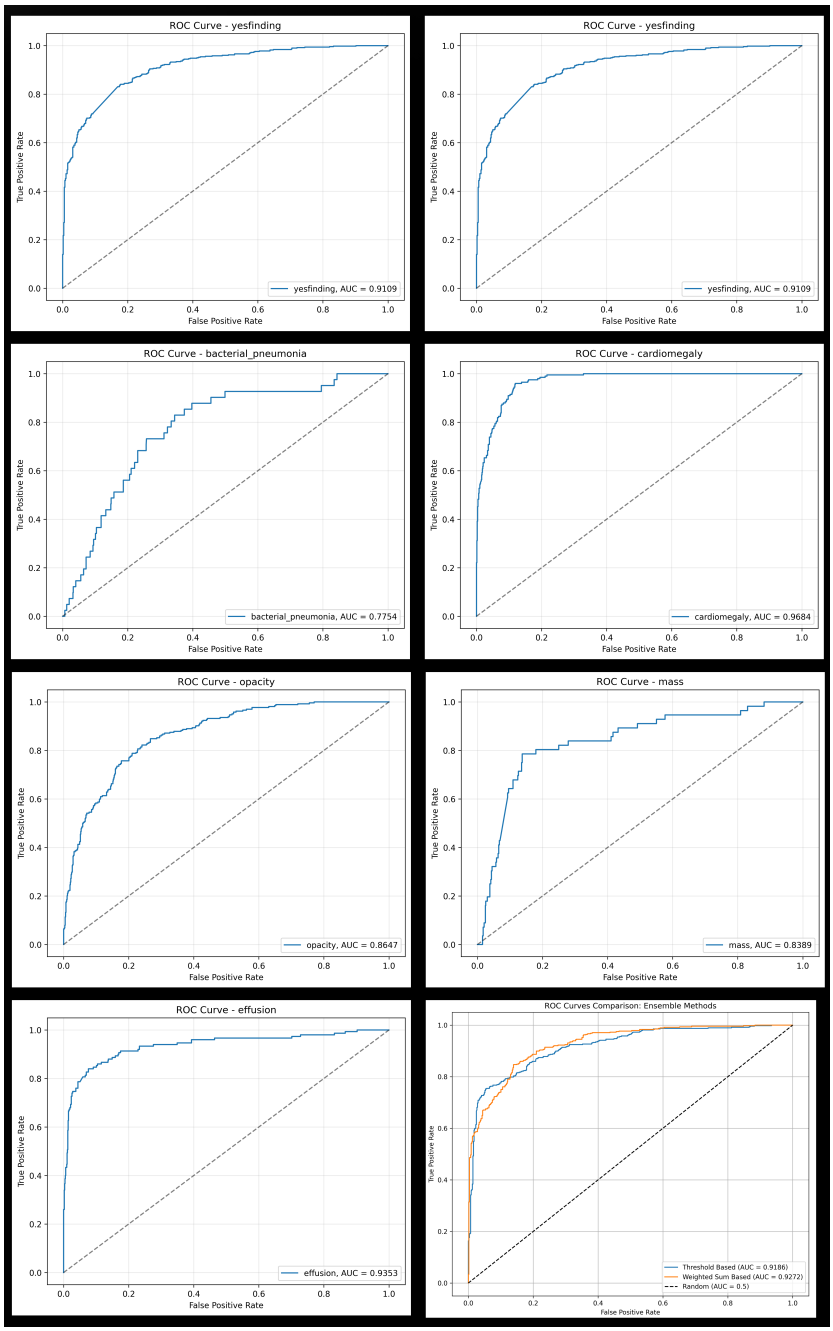

Figure 5. Comparison of receiver operating characteristic curves for the Threshold-Based and Weighted Sum-Based ensemble methods.

 doi:10.20944/preprints202510.1351.v1

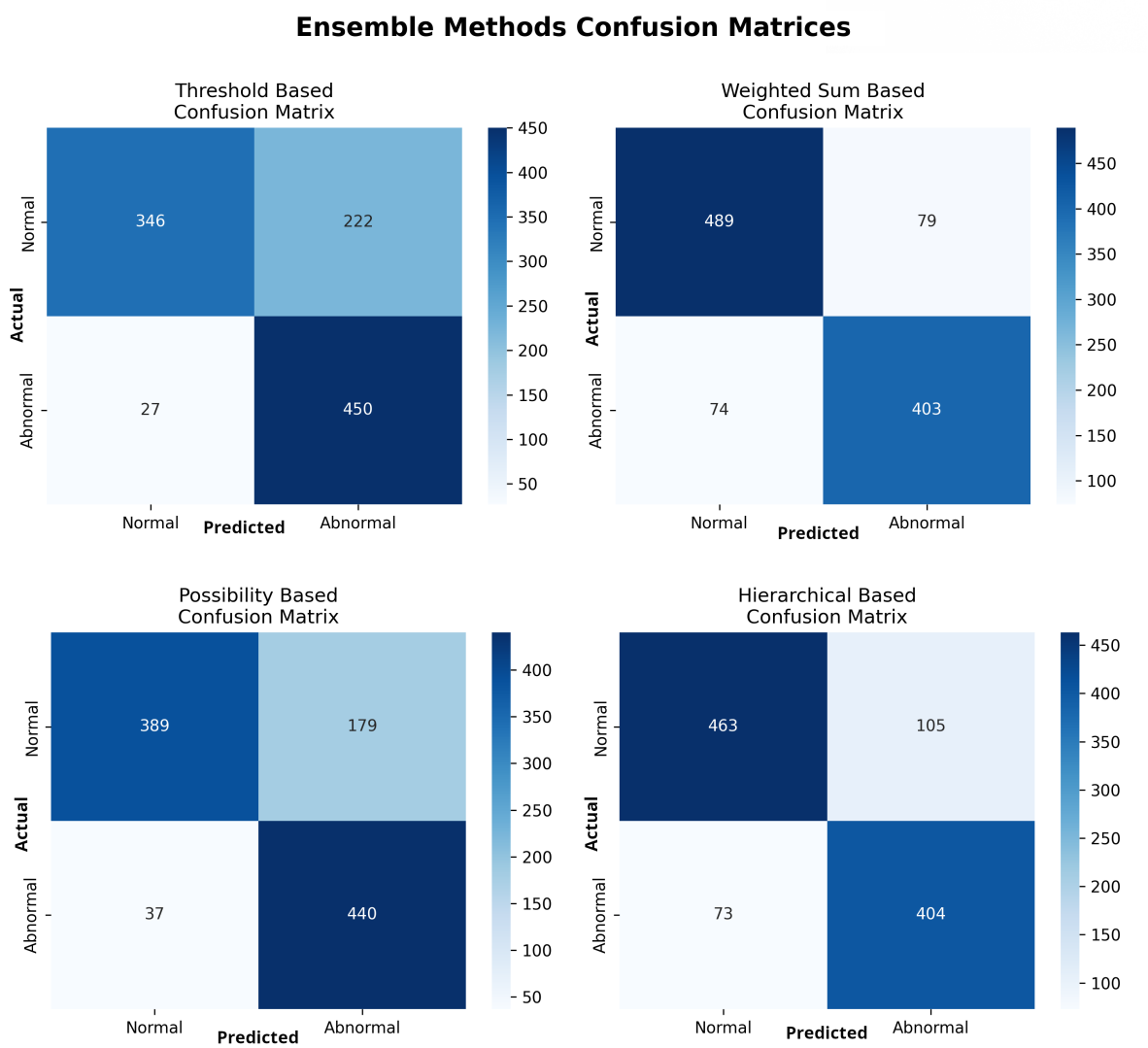

Figure 6. Confusion matrices for the four ensemble methods evaluated for abnormality detection.

3.3. Error Analysis

An analysis of the algorithm's false negatives reveals a nuanced performance profile, where a specific miss did not always equate to a complete system failure. Across the board, a substantial number of cases with a false negative for one specific pathology were nevertheless correctly identified as abnormal by the general "yesfinding" classifier or had other co-occurring true positive pathologies. For example, out of 38 false negatives for 'opacity,' the algorithm correctly identified the exam as abnormal in 25 cases. In many instances of a missed finding, the algorithm successfully detected other pathologies within the same study. Misclassification was another notable source of error, where an existing pathology was detected but incorrectly labeled, such as opacities being predicted as masses or nodules (Figure 7). Radiologist review of these false-negative cases frequently noted findings that were "subtle," "small," "doubtful," "seen on lateral image only," or "non-specific." In other cases, the algorithm correctly identified a more critical finding, like a large pleural effusion, while missing a secondary, less significant pathology (Figure 3). Figures 10–12 illustrate false negative cases for opacity, pneumothorax, and nodule, respectively. This highlights the inherent limitations and inevitable failures of this kind of system.

 doi:10.20944/preprints202510.1351.v1

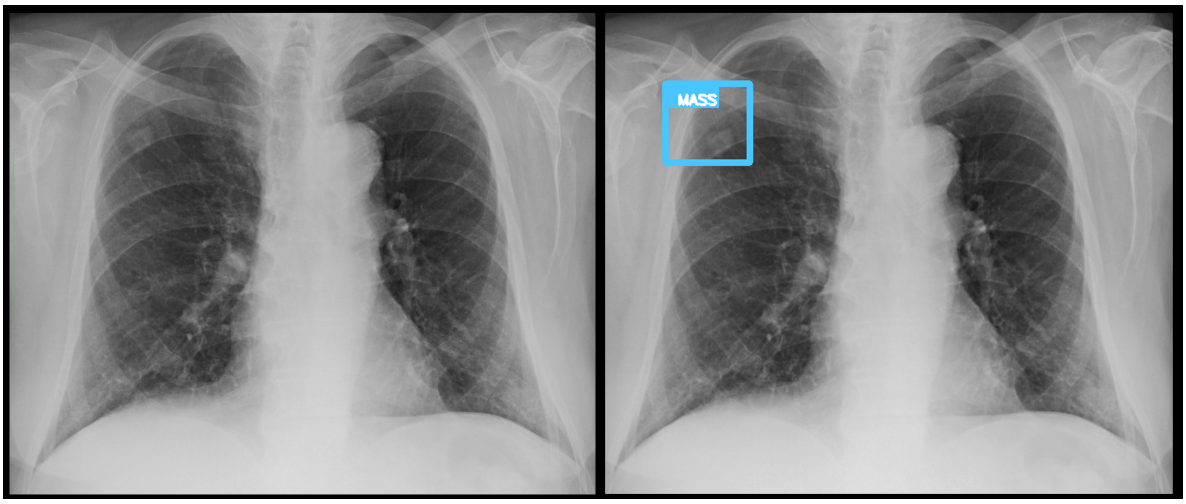

Figure 7. A true positive for a nodule/mass. In this case, the report indicated an opacity, however it was classified by the algorithm as a mass.

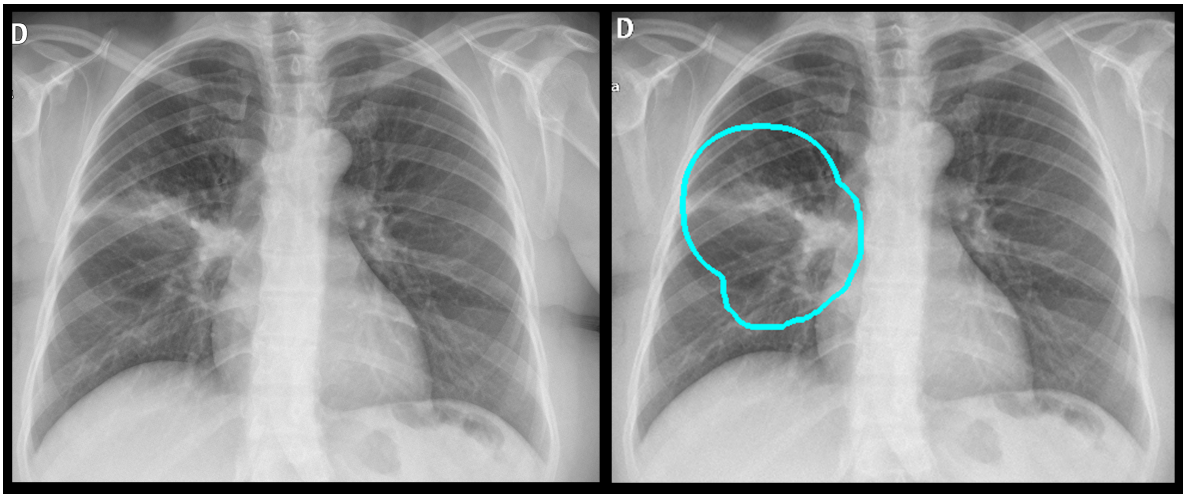

Figure 8. A true positive for opacity and consolidation.

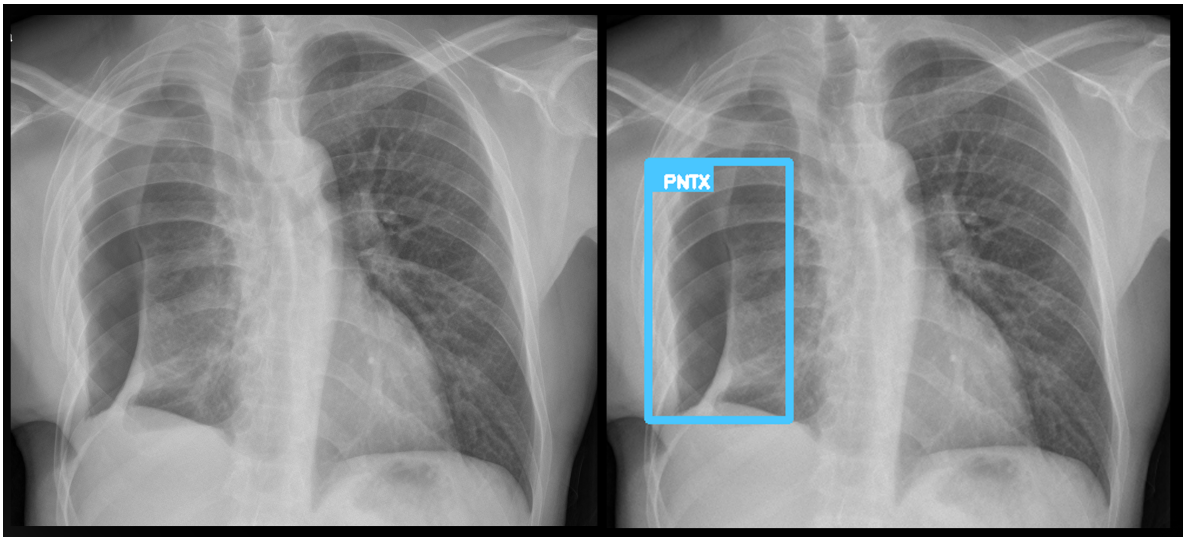

Figure 9. A true positive for pneumothorax. The algorithm successfully detected pneumothorax without a pleural tube, indicating it is not overfitted to post-drainage cases.

 doi:10.20944/preprints202510.1351.v1

Figure 10. A false negative example of a small, missed nodule (black arrow).

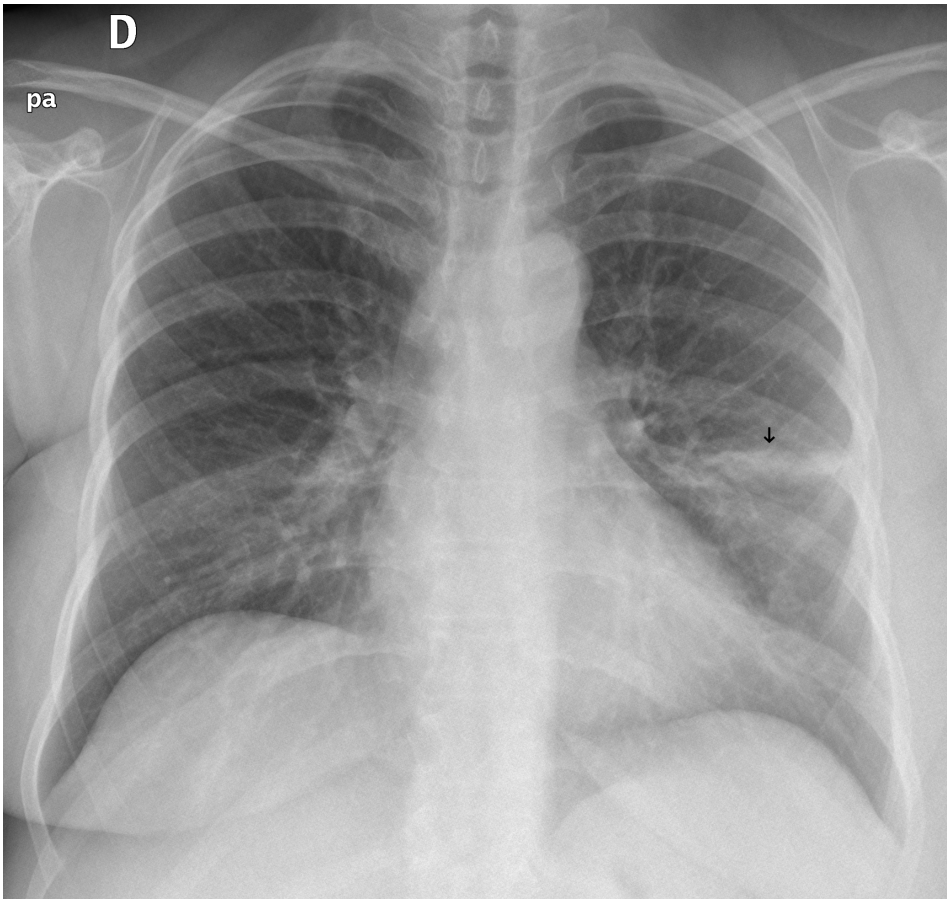

Figure 11. A false negative example of an alveolar opacity (black arrow).

 doi:10.20944/preprints202510.1351.v1

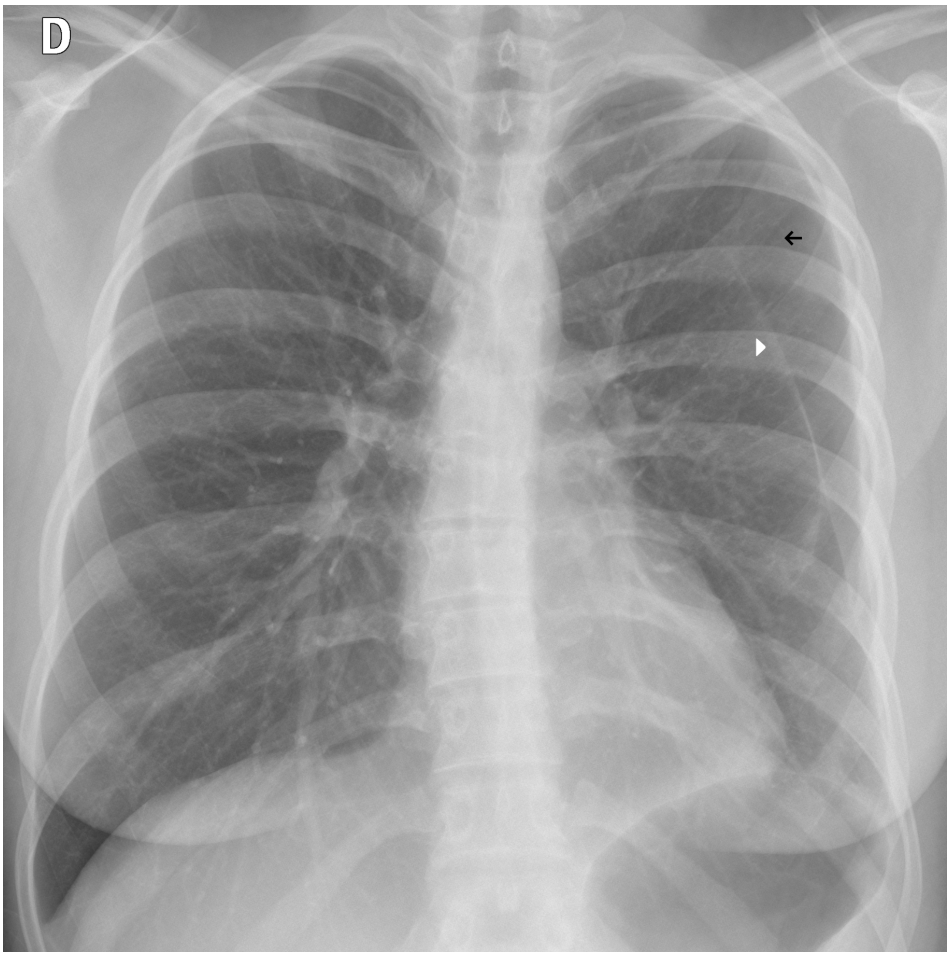

Figure 12. A false negative example of a missed loculated pneumothorax. The presence of pulmonary markings (black arrow) beyond the pleural line (white arrowhead) may have misled the algorithm.

4. Discussion

The intended clinical application for the TRIA system is as a triage tool to prioritize worklists, a common and valuable use case for radiology AI[3,15]. This study provides a necessary external validation of the system for this purpose. The evaluation methodology used is a standard approach in AI validation, as highlighted in systematic reviews and other external evaluations[15–17].

A key methodological strength of the underlying AI model is the tongue-shaped segmentation mask. By including the mediastinum and hilar regions in the ROI, the model analyzes these diagnostically critical areas, which are often excluded by traditional lung-only segmentation. This approach creates a more clinically relevant area for analysis that may improve the detection of central pathologies.

However, a significant trade-off in the model's design is the use of a very low input resolution (75x75 pixels) for classification. While this allows for computational efficiency, it inherently risks the loss of fine-grained details necessary for detecting subtle pathologies like small nodules or early interstitial disease. This was reflected in the error analysis, where false negatives were often subtle cases. The high false-positive rate for conditions like opacity and bacterial pneumonia suggests the model may be oversensitive to non-specific patterns, highlighting a discrepancy between pure pattern recognition and a radiologist's clinical interpretation, which incorporates a higher threshold for significance.

The set of pathologies chosen to build the tool reflects the importance of critical findings for emergency physicians unassisted by radiology specialists. The limited number of specific pathologies favors computational efficiency, while the general normal/abnormal classifier ("yesfinding") provides a safeguard for unlisted pathologies, as depicted by the ensemble results.

 doi:10.20944/preprints202510.1351.v1

Our validation dataset, comprising 1,045 curated chest radiographs (CXRs), provides a statistically robust sample size. This cohort yields 95% confidence intervals with margins of error of approximately ±3–4% for key performance metrics, thereby exceeding typical requirements for medical imaging validation. Although this enriched sampling ensures that the algorithm is evaluated across a wide range of conditions, it does not reflect true disease prevalence and contrasts with studies employing large, consecutive prospective cohorts from routine clinical practice, which provide a more realistic benchmark of real-world performance[18,19]. Ground truth was established through LLM–assisted extraction of findings from reports authored by senior radiologists, a pragmatic strategy for large-scale datasets. A subset of 300 reports was independently reviewed by a radiologist, yielding an accuracy of 0.9838 (95% CI: 0.9783–0.9883), thereby supporting the reliability of the adopted approach. Nevertheless, more rigorous standards for ground truth determination exist, such as consensus interpretation by a panel of expert radiologists or correlation with reference imaging modalities, including computed tomography [18,20–22].

A comparative analysis reveals a landscape of varied performance among commercial AI solutions. In this study, the possibility-based ensemble (used in the commercial product) obtained high sensitivity, while the weighted sum ensemble demonstrated robust and balanced performance for abnormality stratification, achieving an AUROC of 0.927 with a sensitivity of 84.5% and a specificity of 86.1%. This balanced profile contrasts with other AI triage tools, such as the one evaluated by Blake et al., which was optimized for very high sensitivity (99.6%) at the expense of lower specificity (67.4%)[15]. Another algorithm evaluated in a primary care setting showed excellent specificity (92.0%) but significantly lower sensitivity (47.0%)[23]. These differences represent fundamental design choices: one prioritizes minimizing missed findings, accepting more false positives, while the other seeks to balance sensitivity and specificity to reduce workload more efficiently. The overall AUROC for TRIA is comparable to high-performing algorithms evaluated in other studies, which often achieve AUROC in the 0.92 to 0.94 range[24,25].

The error analysis of TRIA, which noted that false negatives were often subtle or challenging cases, is a consistent finding across the field. For instance, Plesner et al. similarly found that the performance of four different commercial tools declined for smaller-sized findings and on complex radiographs with multiple pathologies[18]. Our study identifies the model's very low input resolution as a potential factor contributing to missing fine-grained details. Despite this, while the model may misclassify or miss a specific finding, its ability to flag an exam as generally abnormal remains high, ensuring that complex cases are still prioritized for radiologist review. The variability highlighted in Table 5 also shows that algorithm performance is highly dependent on the clinical task and validation setting, suggesting that a single set of metrics may not fully capture an algorithm's utility.

This study has some limitations. First, this evaluation is retrospective, and performance can decrease when transitioning from a curated analysis to a prospective implementation[16]. Second, the validation was conducted at a single center, which may limit generalizability. Finally, the validation dataset was enriched with significant pathologies and therefore does not reflect the typical prevalence of findings in a real-world population, which may influence performance metrics.

Future work should include prospective, multi-center validation studies to assess the algorithm's performance in real-time clinical workflows and to measure its impact on downstream outcomes, such as time-to-diagnosis and workload reduction. Further refinement of the classification model, particularly by adding more pathologies and exploring higher input resolutions, could enhance its ability to detect more subtle findings and improve its clinical utility.

 doi:10.20944/preprints202510.1351.v1

Table 5. Comparative Performance Metrics of AI Algorithms for Chest Radiograph Abnormality Detection.

Study / Algorithm	Type	AUROC	Sens. (%)	Spec. (%)
This Study				
TRIA (Possibility)	R	N/A	92.2	68.5
TRIA (Weighted Sum)	R	0.927	84.5	86.1
Blake et al. 2023[15]				
qXR	R	N/A	99.6	67.4
Vasilev et al. 2023[16]				
Lunit INSIGHT	R	0.940	90.0	89.0
Lunit INSIGHT	P	0.840	77.0	81.0
Catalina et al. 2024[23]				
ChestEye	P	N/A	47.0	92.0
Arzamasov et al. 2024[25]				
System 1 (qXR)	R	0.921	N/A	N/A
System 2 (Lunit INSIGHT)	R	0.932	N/A	N/A
de Camargo et al. 2025[21]				
Algorithm (LAM)	R	0.938	36.3	99.5
Qin et al. 2019[24]				
Lunit INSIGHT	R	0.940	95.0†	80.0†
qXR	R	0.940	95.0†	72.0†

Notes: N/A = Not Available or not reported as a primary endpoint; R = Retrospective; P = Prospective; LAM = Lung Abnormality Model. †Metrics from Qin et al. represent the operating point required to achieve a sensitivity of at least 95%.

5. Conclusion

This independent retrospective validation demonstrates that the TRIA AI algorithm achieves robust and accurate performance in discriminating between normal and abnormal chest radiographs. The actual commercial version has excellent sensitivity and provides confidence for clinical implementation as a reliable tool for triaging CXRs. The strong and balanced performance of weighted sum ensemble strategy provides a viable alternative when more specificity is warranted. The tool could help manage reporting backlogs, prioritize urgent cases for radiologist review, and potentially support earlier patient intervention. This study provides a foundation for future prospective and comparative research to further evaluate the algorithm's clinical utility.

Author Contributions: Conceptualization, A.E. and A.C.; Methodology, A.C. and M.W.; Software, M.W., A.Y. and A.C.; Validation, I.D., M.S., and M.E.; Formal Analysis, M.W. and A.C.; Investigation, A.C., I.D., M.S., and M.E.; Resources, I.D., M.S., and M.E.; Data Curation, M.W., I.D., M.S., and M.E.; Writing—Original Draft Preparation, A.C.; Writing—Review and Editing, all authors; Supervision, M.S., A.C. and A.E.; Project Administration, A.C.; Funding Acquisition, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC was funded by NeuralMed (A2 Tecnologia Ltda)

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of HCFMUSP (protocol code CAAE: 85444924.9.0000.0068).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study, which involved the analysis of fully anonymized pre-existing medical data.

Data Availability Statement: The processed datasets generated and analysed during the current study are publicly available, however the images and the full reports are confidential and sensitive health data from HCFMUSP and still not released.

Acknowledgments: The authors would like to thank the Department of Radiology at HCFMUSP for their support and for providing the data for this validation study.

 doi:10.20944/preprints202510.1351.v1

Conflicts of Interest: A.C. and A.E. are founders and stakeholders of of NeuralMed, the company that manufactures the TRIA system and funded the study. A.C., A.E., M.W., and A.Y. are employees of NeuralMed. The other authors declare no conflicts of interest. The funders had a role in the design of the study; in the data analysis and interpretation; in the writing of the manuscript, and in the decision to publish the results.

References

1. Smith-Bindman, R.; Miglioretti, D.L.; Johnson, E.; et al. Use of Diagnostic Imaging Studies and Associated Radiation Exposure for Patients Enrolled in Large Integrated Health Care Systems, 1996-2010. *JAMA* **2012**, *307*, 2400–2409.
2. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118.
3. Annarumma, M.; Withey, S.J.; Bakewell, R.J.; Pesce, E.; Goh, V.; Montana, G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* **2019**, *291*, 196–202.
4. Anderson, P.G.; Tarder-Stoll, H.; Alpaslan, M.; Keathley, N.; Levin, D.L.; Venkatesh, S.; et al. Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. *Sci. Rep.* **2024**, *14*, 25151.
5. Windecker, D.; Baj, G.; Shiri, I.; et al. Generalizability of FDA-Approved AI-Enabled Medical Devices for Clinical Use. *JAMA Netw. Open* **2025**, *8*, e258052.
6. Agência Nacional de Vigilância Sanitária. *Consulta de Registro de Software como Dispositivo Médico - A2 Tecnologia Ltda.*; 2023. Available online: https://consultas.anvisa.gov.br/#/saude/25351120732202361/?nomeProduto=tria (accessed on 10 October 2025).
7. Agência Nacional de Vigilância Sanitária. *Dispõe sobre as boas práticas de fabricação de produtos médicos e produtos para diagnóstico de uso in vitro e dá outras providências. RDC Nº 657/2022*; Ministério da Saúde, 2022. Available online: https://www.in.gov.br/web/dou/-/resolucao-rdc-n-657-de-24-de-marco-de-2022-389712613 (accessed on 5 October 2025).
8. Brasil. *Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709)*; Diário Oficial da União, 14 August 2018. Available online: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm (accessed on 5 October 2025).
9. Jaeger, S.; Candemir, S.; Antani, S.; Wang, Y.J.; Lu, P.; Thoma, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475–477.
10. Demner-Fushman, D.; et al. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310.
11. Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.; et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am. J. Roentgenol.* **2000**, *174*, 71–74.
12. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471.
13. Bustos, A.; Pertusa, A.; Salinas, J.M.; de la Iglesia-Vayá, M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **2020**, *66*, 101797.
14. Gemini Team. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*; Google, 2025. Available online: https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf (accessed on 5 October 2025).
15. Blake, S.R.; Das, N.; Tadepalli, M.; Reddy, B.; Singh, A.; Agrawal, R.; et al. Using Artificial Intelligence to Stratify Normal versus Abnormal Chest X-rays: External Validation of a Deep Learning Algorithm at East Kent Hospitals University NHS Foundation Trust. *Diagnostics* **2023**, *13*, 3408.
16. Vasilev, Y.; Vladzymyrskyy, A.; Omelyanskaya, O.; Blokhin, I.; Kirpichev, Y.; Arzamasov, K. AI-Based CXR First Reading: Current Limitations to Ensure Practical Value. *Diagnostics* **2023**, *13*, 1430.
17. Harris, M.; Qi, A.; Jeagal, L.; Torabi, N.; Menzies, D.; Korobitsyn, A.; et al. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. *PLOS ONE* **2019**, *14*, e0221339.
18. Plesner, L.L.; Müller, F.C.; Brejnebøl, M.W.; Laustrup, L.C.; Rasmussen, F.; Nielsen, O.W.; et al. Commercially Available Chest Radiograph AI Tools for Detecting Airspace Disease, Pneumothorax, and Pleural Effusion. *Radiology* **2023**, *308*, e231236.

 doi:10.20944/preprints202510.1351.v1

19. Niehoff, J.H.; Kalaitzidis, J.; Kroeger, J.R.; Schoenbeck, D.; Borggrefe, J.; Michael, A.E. Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays. *Sci. Rep.* **2023**, *13*, 3680.
20. Khader, F.; Han, T.; Müller-Franzes, G.; Huck, L.; Schad, P.; Keil, S.; et al. Artificial Intelligence for Clinical Interpretation of Bedside Chest Radiographs. *Radiology* **2023**, *307*, e220510.
21. de Camargo, T.F.O.; Ribeiro, G.A.S.; da Silva, M.C.B.; et al. Clinical validation of an artificial intelligence algorithm for classifying tuberculosis and pulmonary findings in chest radiographs. *Front. Artif. Intell.* **2025**, *8*, 1512910.
22. van Leeuwen, K.G.; Schalekamp, S.; Rutten, M.J.C.M.; et al. Comparison of Commercial AI Software Performance for Radiograph Lung Nodule Detection and Bone Age Prediction. *Radiology* **2024**, *310*, e230981.
23. Catalina, Q.M.; Vidal-Alaball, J.; Fuster-Casanovas, A.; Escalé-Besa, A.; Ruiz Comellas, A.; Solé-Casals, J. Real-world testing of an artificial intelligence algorithm for the analysis of chest X-rays in primary care settings. *Sci. Rep.* **2024**, *14*, 5199.
24. Qin, Z.Z.; Sander, M.S.; Rai, B.; Titahong, C.N.; Sudrungrot, S.; Laah, S.N.; et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci. Rep.* **2019**, *9*, 15000.
25. Arzamasov, K.; Vasilev, Y.; Zelenova, M.; Pestrenin, L.; Busygina, Y.; Bobrovskaya, T.; et al. Independent evaluation of the accuracy of 5 artificial intelligence software for detecting lung nodules on chest X-rays. *Quant. Imaging Med. Surg.* **2024**, *14*, 5288–5303.