AI Methods for the Diagnosis of Pneumonia on Chest Radiographs: Evaluation of Clinical Interpretability

Christopher L. Valdes, MSc[1], Karen Drukker, PhD[1], Lubomir M. Hadjiyski, PhD[2], Carol Wu, MD[3], Samuel G. Armato III, PhD[1]

1. Department of Radiology, The University of Chicago 2. Department of Radiology, University of Michigan 3. Department of Radiology, MD Anderson Cancer Center

2026 JULY 19–22 VANCOUVER, BC JOINT AAPM | COMP MEETING

Abstract

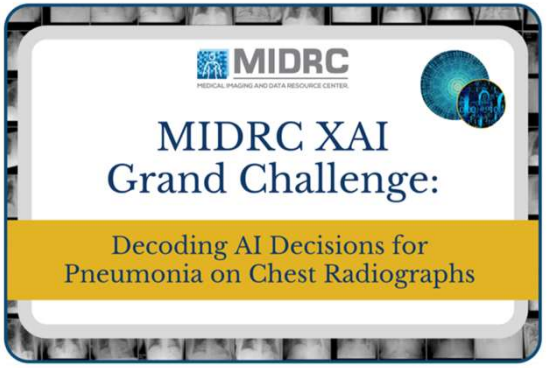
Purpose: To identify which performance metrics for artificial intelligence (AI) models detecting pneumonia on chest radiographs best align with radiologists' subjective assessments.

Methods: 100 chest radiographs from the MIDRC XAI Challenge test set were paired with reference probability maps derived from manual segmentations by three radiologists. Four top-performing models (A-D) were evaluated using the weighted log-loss (WLL), Dice similarity coefficient (DSC), and weighted DSC (wDSC). Model outputs were binned into quartiles for the wDSC. Models were ranked by each metric, and average values with 95% confidence intervals were calculated. Five radiologists independently ranked model outputs for clinical utility in pneumonia localization, and these rankings were compared with metric-based rankings.

Results: Radiologists ranked the models from best to worst as C, B, D, A. The WLL produced the ranking A, B, C, D. The DSC produced the ranking B, C, A, D. The wDSC produced the ranking C, B, A, D.

Conclusion: The wDSC aligned most closely with the rankings of the radiologists. Future studies will explore alternative weighting schemes and investigate alternative metrics.

Background



- Challenge participants developed AI models for pneumonia classification on chest radiographs and provided explainability maps showing regions that influenced outputs. We assessed the utility of these explainability maps.
- In the XAI Challenge, the weighted log-loss (WLL) metric was used to evaluate explainability maps.
- A subsequent reader study of 100 cases found that the winning model received the lowest ratings for clinical utility. This metric was initially selected to discourage participants from developing segmentation-based classification models and submitting segmentations as explainability maps.

Hypothesis

- In this study, we hypothesized that alternative metrics, such as the Dice similarity coefficient (DSC) and weighted DSC (wDSC), better align with radiologist assessments of clinical utility.

Methods

Figure 1: Chest radiograph demonstrating pneumonia used in this study.

- A subset of 100 cases from the Challenge test set was evaluated.
- Model outputs (probability maps) from top 4 performing models were compared with corresponding reference standards.
- The WLL, DSC, and wDSC were calculated for each output and reference standard pairing.
- For DSC calculations, pixel values of 85 or higher were incorporated. For the reference standard, pixels included within the contour of at least 1 radiologist.
- For wDSC calculations, pixel values from model outputs were binned into quartiles to represent each contour level (0-3).

$$L_{log}(y,p) = -w(y\log(p) + (1-y)\log(1-p))$$

Equation 1: Equation for the WLL. y represents the true pixel label while p represents the reference standard probability. Weights (w) of 1, 15, 30, and 45 were assigned to pixels with 0, 1, 2, and 3 radiologist-drawn contours from the reference standard, respectively.

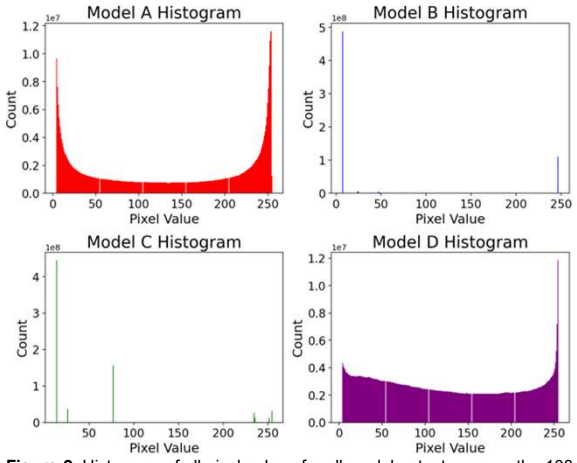


Figure 2. Histogram of all pixel values for all model outputs across the 100 cases (not including values less than 4). For DSC comparisons, pixel values greater than 85 were considered to contain pneumonia.

$$DSC = \frac{2|A \cap B|}{|A| + |B|}$$

Equation 2: Equation for Dice Similarity Coefficient (DSC). A is the area of the model output, and B is the area covered by the reference standard.

$$wDSC = \sum_{k=1}^{3} w_k DSC_k$$

Equation 3: Equation for weighted DSC. w_k is equal to 0.33 for each contour level, which is how many radiologists contoured that region in the reference standard. DSC_k is the DSC of the contour level.

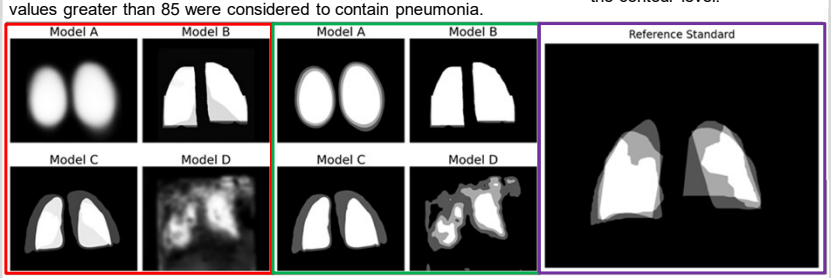


Figure 3. Left: Sample model outputs from the winning models of the Challenge. **Middle:** Binned model outputs used for wDSC comparisons. **Right:** Associated reference standard for this chest radiograph.

Results

Model	WLL	DSC	wDSC
A	0.15 [0.13, 0.17]	0.59 [0.55, 0.63]	0.49 [0.45, 0.53]
B	0.29 [0.23, 0.36]	0.66 [0.61, 0.71]	0.57 [0.54, 0.61]
C	0.39 [0.33, 0.43]	0.64 [0.59, 0.68]	0.66 [0.62, 0.70]
D	0.42 [0.38, 0.46]	0.38 [0.34, 0.41]	0.34 [0.31, 0.38]

Table 1: Average metric values for each model across all cases. 95% confidence intervals are within brackets. No statistical comparisons were performed.

Ranking	Radiologists	WLL	DSC	wDSC
1st	C	A	B	C
2nd	B	B	C	B
3rd	D	C	A	A
4th	A	D	D	D

Table 2: Rankings produced by the radiologists from the observer study and each metric. The wDSC achieved the most similar ranking to the radiologists.

Future Directions

- Investigate alternative WLL weighting schemes (quadratic, exponential, logarithmic).
- Investigate background handling strategies, including semantic segmentation-based approaches and increase penalties for false-positive explanations.
- Explore alternative approached for quantifying model attention, including ROC-based analyses.
- Extend analysis to all 1,939 test cases from the original challenge.

Conclusion

- The wDSC captured clinical relevance better than the DSC and WLL.
- Future challenges must carefully consider selected metrics to ensure clinical relevance in resultant model outputs.

Acknowledgments

MIDRC MEDICAL IMAGING AND DATA RESOURCE CENTER

Thank you to all radiologists who participated in the observer study!

References

Information about the challenge results can be found at the following link:

www.midrc.org/xai-challenge-2024

Additional information about the challenge can be found at the github repository (see link above for details).