Evaluating AI Performance in Pediatric Lymphoma Detection on [^{18}F]FDG PET/MRI

Shashi B. Singh, MBBS[1], Hongzhi Wang, PhD[2], Iryna Vasyliv, MD[1], Vanessa Ricarda Sophie von Kruechten, MD[1], Hyun Gi Kim, MD[1], Lucia Baratto, MD[1], Joy Tzung-Yu Wu, MBChB[1,2], Amir Hossein Sarrami, MD[1], Lisa Christine Adams, MD[1], Tie Liang, PhD[1], Tanveer Syeda-Mahmood, PhD[2], Heike E. Daldrup-Link, MD, PhD[1]

[1]Department of Radiology, Stanford University School of Medicine, Stanford, CA; [2]IBM Almaden Research Center, San Jose, CA

PURPOSE

- We developed a 3D convolutional neural network (CNN) for detecting lymphoma in children on [^{18}F]FDG-PET/MR scans
- This study evaluates the performance of this new AI algorithm and three human readers compared to a reference standard of joined review by a radiologist and a nuclear medicine expert

METHODS

- Our 3D CNN comprises a two-step method: (1) flagging tumor candidates on PET with SUV>2.0 (2) removal of false positives using patches extracted from PET and MRI via a multi-modal fusion method (**Figure 1**)
- The CNN was trained on 53 annotated baseline [^{18}F]FDG-PET/MR images until the algorithm reached 200 epochs of training using the Adam optimizer
- The algorithm was then tested on 30 non-annotated [^{18}F]FDG-PET/MR images
- An MD researcher noted the number of lymphoma lesions flagged by the AI algorithm in five anatomical regions (head, neck, thorax, abdomen and pelvis, and extremities)
- For comparison, three human researchers with different levels of experience flagged and documented lymphoma lesions in five anatomical regions (head, neck, thorax, abdomen and pelvis, and extremities). All human researchers were trainees or radiologists not routinely reading whole body scans.
- The reference standard for true positive lesions was determined by a joined read by an expert pediatric radiologist and an expert nuclear medicine physician
- The agreement and 95% confidence interval (CI) between the AI algorithm or human readers and the reference standard was determined using percent agreement analysis
- The established benchmark in the literature for agreement evaluation is >0.6, indicating substantial agreement, and >0.8, indicating perfect agreement [1]

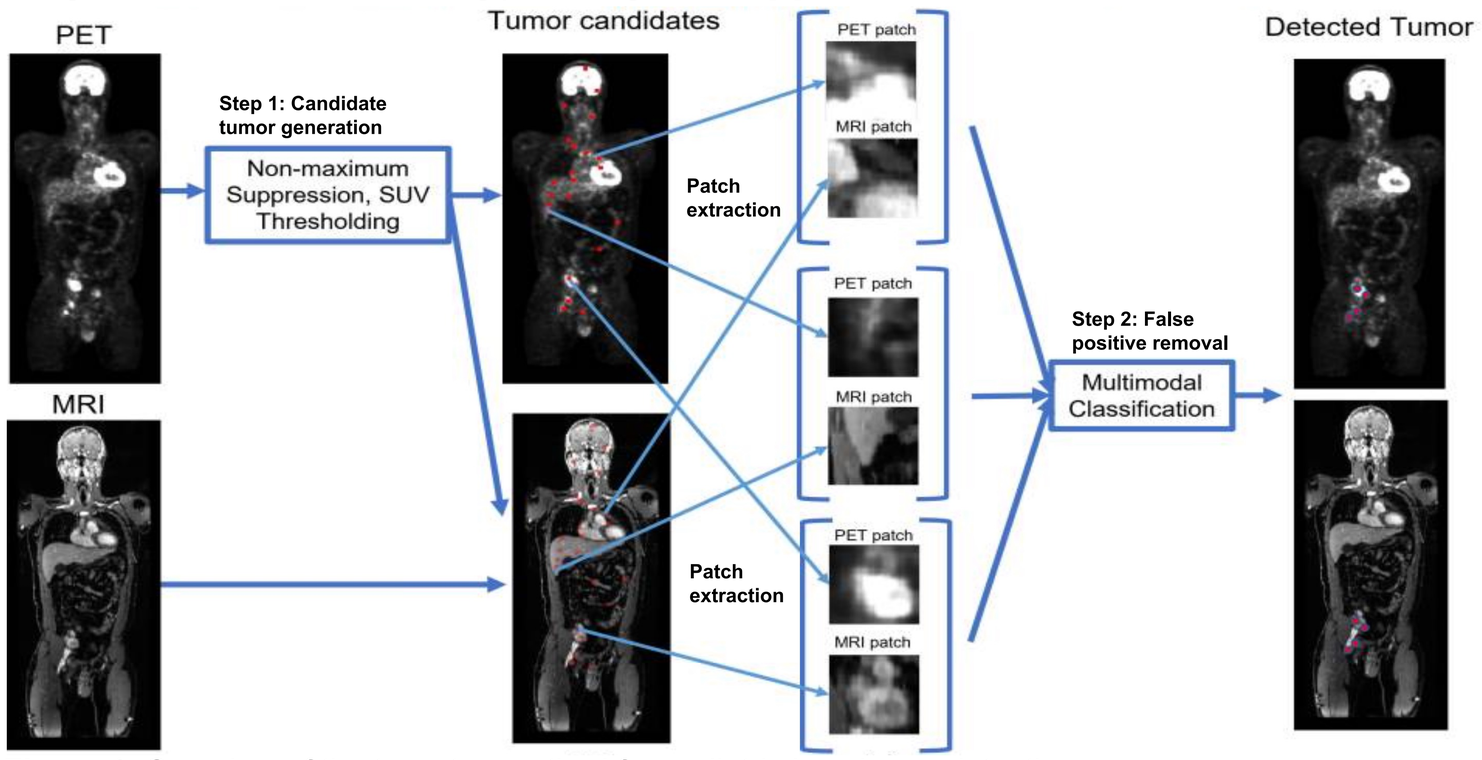


Figure 1. Overview of the two-step method for pediatric lymphoma detection

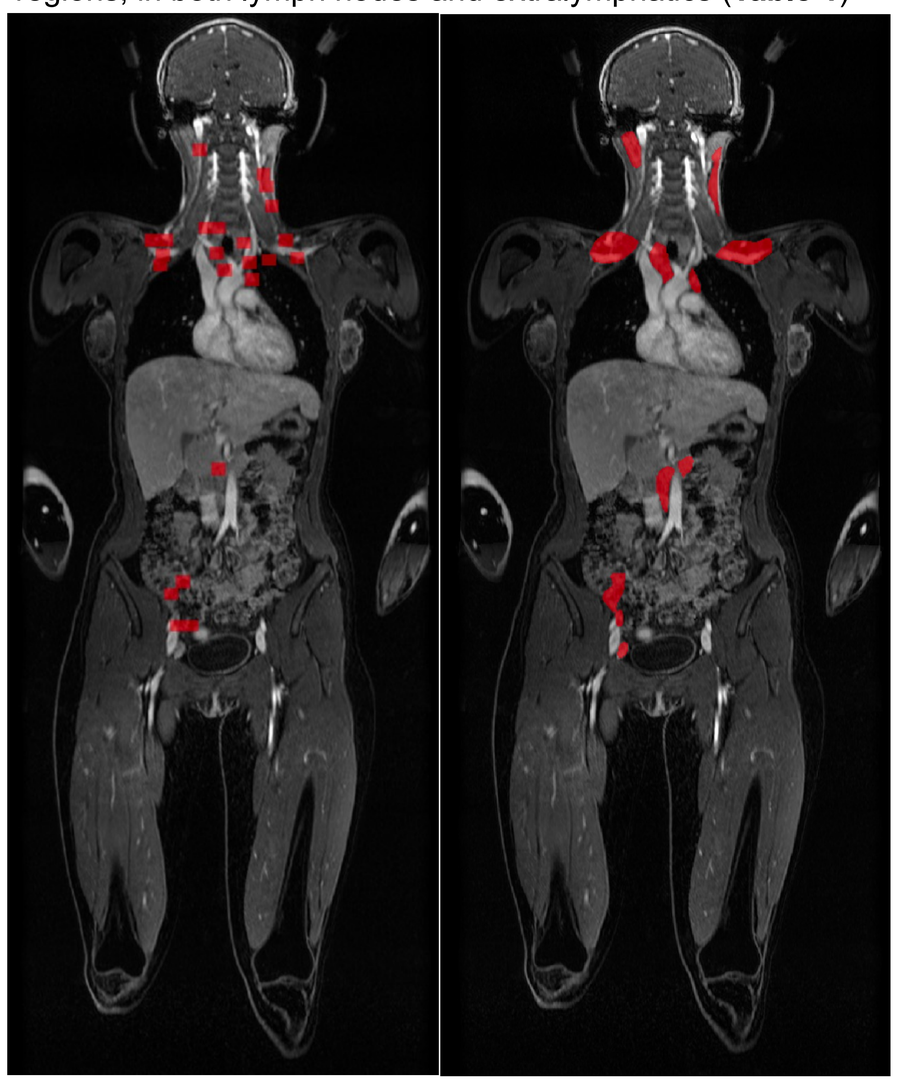
RESULTS

Region	Location	Truth vs AI: percent agreement (95% CI)	Truth vs reader 1: percent agreement (95% CI)	Truth vs reader 2: percent agreement (95% CI)	Truth vs reader 3: percent agreement (95% CI)
Head	Lymph node	1.00 (0.88-1.00)	0.97 (0.83-1.00)	0.93 (0.78-0.99)	0.97 (0.83-1.00)
Head	Extralymphatics	0.93 (0.78-0.99)	0.80 (0.61-0.92)	0.87 (0.69-0.96)	0.43 (0.25-0.63)
Neck	Lymph node	0.70 (0.51-0.85)	0.43 (0.25-0.63)	0.43 (0.25-0.63)	0.23 (0.10-0.42)
Neck	Extralymphatics	0.97 (0.83-1.00)	0.90 (0.73-0.98)	0.83 (0.65-0.94)	0.87 (0.69-0.96)
Thorax	Lymph node	0.73 (0.54-0.88)	0.47 (0.28-0.66)	0.50 (0.31-0.69)	0.43 (0.25-0.63)
Thorax	Extralymphatics	0.77 (0.58-0.90)	0.67 (0.47-0.83)	0.53 (0.34-0.72)	0.47 (0.28-0.66)
Abdomen and pelvis	Lymph node	0.80 (0.61-0.92)	0.57 (0.37-0.75)	0.67 (0.47-0.83)	0.50 (0.31-0.69)
Abdomen and pelvis	Extralymphatics	0.70 (0.51-0.85)	0.73 (0.54-0.88)	0.60 (0.41-0.77)	0.77 (0.58-0.90)
Extremities	Lymph node	1.00 (0.88-1.00)	1.00 (0.88-1.00)	1.00 (0.88-1.00)	0.93 (0.78-0.99)
Extremities	Extralymphatics	0.73 (0.54-0.88)	0.87 (0.69-0.96)	0.80 (0.61-0.92)	0.83 (0.65-0.94)

Table 1. Percent agreement between the AI algorithm or human readers and the reference standard

RESULTS

- The percent agreement between the AI readout and the reference standard was 0.83 (95% CI = 0.79-0.87 (**Figure 2**)
- The percent agreements between human readouts and the reference standard were 0.74 (95% CI = 0.69-0.79) for reader 1, 0.72 (95% CI = 0.66-0.77) for reader 2, and 0.64 (95% CI = 0.59-0.70) for reader 3
- Only AI met the 0.6 criteria of substantial agreement in all five regions, in both lymph nodes and extralymphatics (**Table 1**)



A. AI readout **B. Reference standard**

Figure 2. Comparison of AI-assisted detection of lymphoma lesions with expert human readers in a 14-year-old teenager with Hodgkin's lymphoma. A. Coronal contrast-enhanced T1-weighted gradient echo MRI scan. The AI algorithm flagged multiple lymphoma lesions in the neck, the bilateral supraclavicular regions, the mediastinum, the para-aortic, and the pelvis region (red squares). B. Coronal contrast-enhanced T1-weighted gradient echo MRI scan. Joined annotation of multiple lymphoma lesions by an expert pediatric radiologist and nuclear medicine physician (red regions).

CONCLUSION

- The AI algorithm showed substantial agreement with the reference standard for lymphoma lesion detection
- Our newly developed AI algorithm can improve the diagnosis of pediatric lymphomas on [^{18}F]FDG-PET/MR

REFERENCE

[1] Landis,J.R, Koch,G.G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, Vol.33,No.1,pp.159-174.

ACKNOWLEDGEMENT

This work was supported by grants from the
- Stanford Center for Artificial Intelligence in Medicine & Imaging (AIMI) and the Human-Centered Artificial Intelligence (HAI)
- National Cancer Institute (Grant number R01CA269231)