Am J Nucl Med Mol Imaging 2026;16(1):55-62
www.ajnmmi.us /ISSN:2160-8407/ajnmmi0168724

Original Article

CNN-based detection of pediatric lymphoma on whole body [^{18}F]FDG-PET/MRI

Shashi B Singh[1*], Yashas Ullas Lokesha[1*], Hongzhi Wang[2], Michael Joseph Barrow[1], Ricarda von Kruechten[1], Iryna Vasyliv[1], Amir Hossein Sarrami[1], Joy Tzung-Yu Wu[1], Lucia Baratto[1], Lisa Christine Adams[1], Hyun Gi Kim[1], Jason Wong[1], Tie Liang[1], Sergios Gatidis[1], Tanveer Syeda-Mahmood[2], Heike E Daldrup-Link[1,3]

[1]Department of Radiology, Stanford University School of Medicine, Stanford, CA 94305, USA; [2]IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA; [3]Department of Pediatrics - Hematology/Oncology, Lucile Packard Children's Hospital, Stanford University, Stanford, CA 94304, USA. [*]Equal contributors.

Received September 10, 2025; Accepted February 13, 2026; Epub February 15, 2026; Published February 28, 2026

Abstract: We assessed the performance of a deep convolutional neural network (CNN) in detecting pediatric lymphoma lesions on [^{18}F]FDG-PET/MRI. We evaluated CNN's sensitivity, specificity, percentage agreement, and processing time compared to the interpretations of a pediatric radiologist and a second-year radiology resident. In this retrospective study, a CNN was trained on annotated [^{18}F]FDG-PET/MRI scans from 53 pediatric lymphoma patients and tested on 30 additional scans. The CNN and two human readers recorded the presence of lesions in five anatomical regions. An additional pediatric radiologist and a nuclear medicine physician determined the reference standard. The sensitivity and specificity of the CNN were compared with those of human readers using the McNemar test, and the detection time of the CNN and human readers was compared using the Wilcoxon signed-rank test. The CNN demonstrated higher sensitivity (84.6%) and specificity (93.7%) than the radiology resident (69.2%, P=0.023; 81.5%, P<0.001), but lower than the pediatric radiologist (98.7%, P<0.001; 99.5%, P<0.001). The CNN achieved 83% agreement with the reference standard (95% CI: 79%-87%), higher than the resident's 63% (95% CI: 59%-69%) but lower than the pediatric radiologist's 94% (95% CI: 92%-97%). The median values and interquartile ranges for the time taken (in minutes) were 4 (3, 5) for the CNN, 8 (7, 10) for the pediatric radiologist, and 15 (9, 20) for the radiology resident. The sensitivity, specificity, and percentage agreement of the CNN were higher than those of a radiology resident but lower than those of a pediatric radiologist. The CNN readout was significantly faster compared to both human readers.

Keywords: Lymphoma, pediatric, artificial Intelligence (AI), convolutional neural network (CNN), positron emission tomography/magnetic resonance imaging (PET/MRI)

Introduction

Lymphoma is the third most common malignancy in children after leukemia and central nervous system tumors, accounting for 12% of overall cancer diagnoses in children and adolescents less than 20 years of age [1]. Pediatric lymphoma has an incidence rate of 27.3 per 1 million children and adolescents under 20 years of age, with approximately 38,052 cases diagnosed in the United States between 2003 and 2019 [2]. [^{18}F]FDG-PET/CT is currently the clinical standard for staging patients with lymphoma [3, 4]. However, [^{18}F]FDG-PET/MRI has recently enabled lymphoma staging with 80% less radiation exposure compared to [^{18}F]FDG-PET/CT [5]. To minimize radiation exposure of pediatric patients, a number of pediatric oncology centers have adopted [^{18}F]FDG-PET/MRI as their main staging procedure for pediatric lymphoma patients [6-10]. However, each [^{18}F]FDG-PET/MRI scan generates 30,000-50,000 images, making interpretation laborious, time-consuming, and occasionally inconsistent. The application of artificial intelligence (AI) algorithms has the potential to expedite and standardize the interpretation of pediatric lymphoma on [^{18}F]FDG-PET/MR images.

To date, significant attempts have been made to develop deep learning algorithms for the detection of lymphoma on PET/CT images, mainly using adult datasets [11-18]. However, limited efforts have been made to develop such algorithms for pediatric patients [19]. Thus far, only a few AI algorithms have been developed for the detection of lymphoma in children and young adults on whole-body PET images [20]. Tie et al. recently developed a longitudinally aware segmentation network (LAS-Net) for automatic quantification of serial PET/CT images of pediatric patients with Hodgkin lymphoma [21]. Similarly, Etchebehere et al. [22] tested an adult-based CNN algorithm developed by Sibille et al. [12] on the pediatric dataset to measure whole-body metabolic tumor volume (wbMTV) and whole-body total lesion glycolysis (wbTLG) on 102 baseline [^{18}F]FDG-PET/CT studies of pediatric lymphoma patients [22]. However, both algorithms in the Tie et al. [21] and Etchebehere et al. [22] studies are based on PET/CT, not PET/MRI - the preferred imaging technique for pediatric lymphoma. There is a scarcity of AI algorithms for pediatric lymphoma detection on whole-body PET/MRI scans. As shown in an earlier investigation, Wang et al. developed the first deep-learning algorithm for the automatic detection of pediatric lymphoma using [^{18}F]FDG-

https://doi.org/10.62347/RSDQ2273

Table 1. Patient demographics and diagnosis

Characteristics	Patient cohort (*N*=83)
Age (in years)	
Mean ± Standard deviation	18.73±6.37
Range	1-30
Sex	
Female	42 (50.60%)
Male	41 (49.40%)
Type of lymphoma	
Hodgkin lymphoma	20 (24%)
Non-Hodgkin lymphoma	53 (64%)
Posttransplant lymphoproliferative disorder	10 (12%)

PET/MRI [23]. The algorithm outperformed the state-of-the-art medical object detection method (nnDetection) [23]. However, the performance of the algorithm with respect to the reference standard and relative to human readers was not evaluated.

The purpose of our study was to assess the performance of a deep convolutional neural network (CNN) in detecting pediatric lymphoma lesions on whole-body [^{18}F]FDG-PET/MR images. We evaluated CNN's sensitivity, specificity, and processing time compared to two separate interpretations provided by two human readers at different levels of experience [a pediatric radiologist (RVK) and a second-year radiology resident (JTW)] and calculated the percentage agreement of CNN as well as each human reader with the reference standard.

Materials and methods

Subjects

In this retrospective study, we obtained approval from the local institutional review board (IRB 65444) to collect clinical [^{18}F]FDG-PET/MRI studies within a centralized image registry for downstream CNN development. Informed consent was waived due to the retrospective nature of the study. Between 06/21/2022 and 04/19/2024, we collected 83 whole-body baseline [^{18}F]FDG-PET/MRI scans from 83 children and young adults (42 female, 41 male), with a mean age of 18.73±6.37 (range: 1-30) years. Tumor histology consisted of 20 patients with Hodgkin lymphoma, 53 with non-Hodgkin lymphoma, and 10 patients with posttransplant lymphoproliferative disorder (**Table 1**). The imaging studies were retrieved, anonymized using Scientific and Educational Computation Technology Research Associates/Picture Archiving and Communication System (SECTRA/PACS), and transferred for further processing in compliance with the Health Insurance Portability and Accountability (HIPAA) requirements.

Image acquisition

All patients fasted for at least 4-6 hours prior to the injection of [^{18}F]FDG, given at a dose of 3-5 MBq/kg body weight. The serum glucose level was determined at the time of the [^{18}F]FDG injection, and all patients demonstrated a glucose level below 120 mg/dL. Approximately 60 minutes after tracer administration, whole-body PET/MRI was conducted on a 3T scanner (Signa GE Healthcare, Milwaukee, Wis), utilizing a 32-channel torso phased-array coil and an eight-channel, receive-only head coil. The PET data acquisition time was 3:30 minutes per bed position (covering 89 slices per bed) for a total of 5-9 bed positions. MRI sequences included axial diffusion-weighted imaging, Dixon sequences, and breath-hold fat-saturated T1-weighted gradient-echo sequences after gadolinium-chelate administration (Gadavist at a dose of 0.1 mmol Gd/kg) following the institution's standard PET/MR protocol.

Development of Deep-CNN

A pediatric radiologist (AHS) with 5 years of experience annotated lymphoma lesions on T1-weighted fat-saturated gadolinium-enhanced MR images using ITK-SNAP version 3.8.0 software (www.itksnap.org) for training the CNN [24]. An experienced computer scientist (HW) developed a multimodal deep CNN that delineated pediatric lymphoma on baseline [^{18}F]FDG-PET/MRI scans in two steps: First, hypermetabolic regions with SUV greater than 2.0 were detected on PET by applying non-maximum suppression as candidate lesions. The non-maximum suppression algorithm was empirically set to merge local maxima closer than 10 mm to avoid producing noisy overlapping candidates. However, this parameter could be further optimized through cross-validation using the training data. No shape information was considered for candidate generation. Next, to reduce false positives, the CNN classified true lesions using 3D image patches (patch size 64*64*64) extracted for each candidate (**Figure 1**).

For each modality, MRI and PET, the corresponding 3D patches were processed by their respective modality-specific encoder. The outputs of the modality-specific encoder are modality-specific features. The extracted MRI and PET features were then fused via weighted feature fusion [25, 26] using a learnable feature quality evaluator. The evaluator took the feature vector produced by the MRI encoder as input and produced a single value score as output to indicate the usefulness of the input feature vector for classification. Similarly, the evaluator also produced a single value score for the feature vector produced by the PET encoder. The two produced feature quality scores were then converted to modality weights via the SoftMax operation. The fused feature was obtained by a weighted average of the extracted MRI and PET features and was passed to a classification layer for final classification. Each modality-specific encoder had four convolutional layers, each with 32 filters and a rectified linear unit (ReLU) for activation. Each of the first three convolutional layers was followed by a max pooling layer with a pool size of 2. The spatial pyramid pooling approach [27] was applied to produce features from the last convolutional

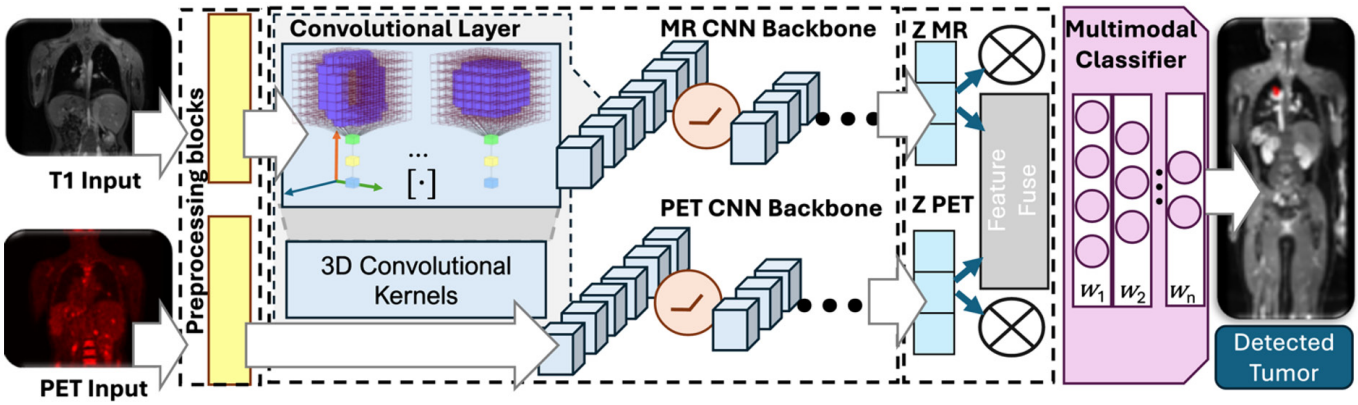

Figure 1. Overview of the proposed two-stage method for pediatric lymphoma detection on [^{18}F]FDG PET/MR scans. During preprocessing, hypermetabolic regions with SUV greater than 2.0 were detected on PET as candidates that may contain lesions by applying non-maximum suppression. To reduce false positives, the convolutional neural network classified true lesions using 3D image patches extracted for each candidate.

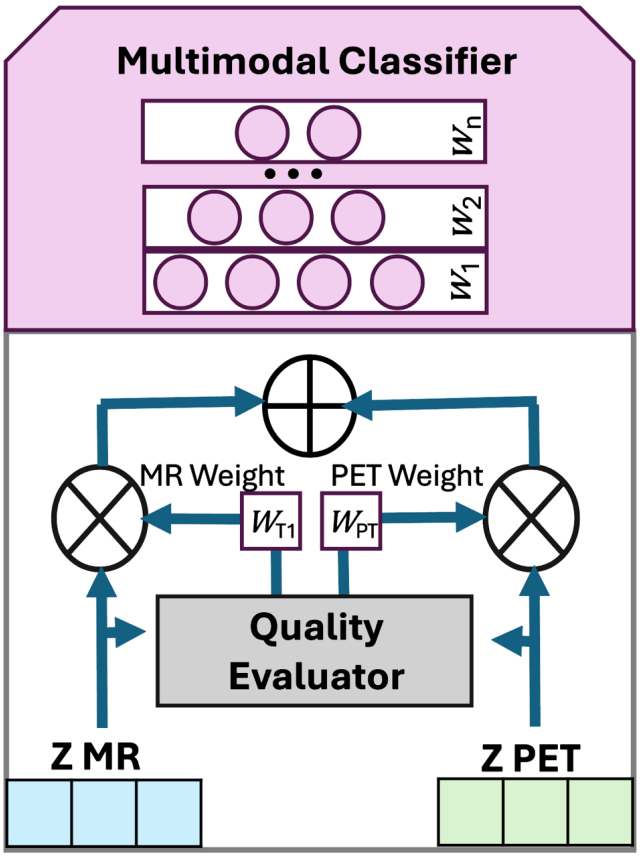

Figure 2. Patch-based tumor lesion classification. Features are extracted from PET and MRI patches using their respective modality-specific 3D CNN encoders. The extracted features are then fused via weighted average, where the fusion weights are produced by a feature quality evaluator.

layer. A regressor with two fully connected layers was applied for the feature quality evaluator used in weighted feature fusion. A fully connected layer with a sigmoid activation was used as the classification layer (**Figure 2**).

The 83 [^{18}F]FDG PET/MR images in our database were randomly assigned to separate data sets for training (*N*=53) and testing (*N*=30). Data partitions were disjoint at the patient level; no patient in the training dataset was included in the validation or testing datasets. To prevent data leakage, scans from previously enrolled patients were identified using unique medical record numbers before de-identification and were excluded from other partitions accordingly. The training data was produced by extracting lesion candidates and their corresponding image patches for all training subjects. The class label for each candidate was derived from expert annotations, where candidates were labeled as true lesions if identified as lesions and false lesions if labeled as non-lesions. A location-aware learning strategy [28] was applied for training. Instead of training a single classifier for the entire body region, a set of classifiers was trained to cover the entire body region, with each classifier focusing on one local anatomical region. For location-aware training, a whole-body anatomical template was created using all training PET scans with the unbiased template-building method [29] and the FLIRT affine registration [30]. Since human anatomy has good left-right symmetry around the longitudinal axis, the longitudinal coordinate in the template was used to index through different anatomical regions. To propagate the template longitudinal coordinate to an individual subject, the template was affine-registered to the subject, from which a template longitudinal coordinate was obtained for each voxel in the target subject.

Let M(μ; σ) be a location-aware classification model, where μ is the central longitudinal coordinate of the anatomical region covered by the model, and σ defines the spatial spread. To train M(μ; σ), all training candidates labeled as true lesions were applied. For training candidates labeled as false lesions, only those located within/near the anatomical region covered by M(μ; σ) were applied in training via a probabilistic sampling strategy. Let x^0 be image patches for one candidate labeled as a false lesion. x^0 was used for training model M(μ; σ) with probability $e^{-\frac{|a(x^0)-\mu|}{\sigma}}$, where $a(x^0)$ is the template longitudinal coordinate for x^0.

Overall, 14 anatomically localized models were trained to cover the entire body region, with σ=90 mm and μ=∈{150 *mm*, 240 *mm*, ..., 1320 *mm*}. These parameters were chosen empirically to make sure that each model has adequate training samples. However, these parameters can be optimized through cross-validation as well. During training, standard affine transform-based data augmentation with up to 30°C rotation and up to 20% scale variations was applied to augment the 3D patches.

For inference, given a patch x, its class label was predicted using the following ensemble rule:

$$p(l \mid x) = \sum_{j=1}^{n} w_j(x)\, p(l \mid x, M(\mu_j, \sigma))$$

where $\sum_{j=1}^{n} w_j(x) = 1$ with $w_j(x) \sim e^{-\frac{|a(x)-\mu_j|}{\sigma}}$.

On average, each validation subject contains five expert-annotated lesions. The lesion candidate generation step produced 192 lesion candidates for each validation subject, with 4.35 true positive and 178.95 false positive lesions. After applying the patch-based lesion classification approach, the final tumor detection performance for the 13 validation subjects is 76% sensitivity at a 10% false discovery rate (FDR).

A computer scientist recorded the time required by the CNN to process each MR image (later referred in the manuscript as "time taken by the CNN" or "processing time of CNN"), and an MD manually recorded the presence of lymphoma lesions identified by the CNN in five anatomical regions (head, neck, thorax, abdomen and pelvis, and extremities). The time taken by the CNN includes only the time required for CNN to process the image and does not include the time taken by the MD to manually record the presence of lymphoma lesions in five anatomical regions.

Image interpretation and reference standard

All images were interpreted using SECTRA/PACS and ITK-SNAP version 3.8.0 software. A pediatric radiologist (RVK) with 3 years of experience and a second-year radiology resident (JTW) after their first nuclear medicine rotation independently noted the presence and number of lymphoma lesions in five anatomical regions (head, neck, thorax, abdomen and pelvis, and extremities). Within each anatomical region, they distinguished between lymph nodal or extra-nodal lesions. In addition, we measured the time taken to detect lesions on each [^{18}F] FDG-PET/MR scan using the CNN and by each human reader, separately. The reference standard for true positive lesions was determined by a joint review by an expert pediatric radiologist (IV) with four years of experience and an expert nuclear medicine physician (LB) with eight years of experience. Disagreement resolution or adjudication between the experts was achieved by utilizing a combination of the final clinical report from all imaging studies, and a consensus was eventually reached in all cases.

Statistical analysis

Lesion-level correspondence between CNN outputs and the reference standard was validated using concordance based on body-region-level detection. Bipartite classification was used to measure sensitivity and specificity. Each of the five anatomical regions (head, neck, thorax, abdomen and pelvis, and extremities) that had a lesion was assigned "1", and those that did not contain a lesion were assigned "0". When CNN or the human reader identified a lesion in a body region, and the reference standard also indicated the presence of a lesion in that region, it was considered a true positive (TP). When the CNN or human reader did not identify a lesion in a body region where the reference standard also indicated no lesion, it was considered a true negative (TN). When the CNN or human reader identified a lesion in a body region where the reference standard indicated no lesion, it was considered a false positive (FP). When the CNN or human reader did not identify a lesion in a body region where the reference standard indicated the presence of a lesion, it was considered a false negative (FN).

The sensitivity and specificity of the CNN were compared with those of each human reader using the McNemar test. To account for multiple comparisons, a Bonferroni correction was applied, setting the significance threshold at P=0.05/2=0.025. The agreement between CNN and human readers with the reference standard was analyzed using the percentage agreement analysis. CNN's median detection time was compared to that of each human reader using a Wilcoxon signed-rank test.

Results

The number of true positive regions for CNN, the pediatric radiologist, and the radiology resident was 66, 77, and 54, respectively. The number of false negative regions for CNN, the pediatric radiologist, and the radiology resident was 12, 1, and 24, respectively. The number of true negative regions for CNN, the pediatric radiologist, and the radiology resident was 207, 220, and 180, respectively. The number of false positive regions for CNN, the pediatric radiologist, and the radiology resident was 14, 1, and 41, respectively. The positive predictive value (PPV) of the CNN, pediatric radiologist, and radiology resident was 82.5%, 98.7%, and 56.8%, respectively, whereas the negative predictive value (NPV) was 94.5%, 99.5%, and 88.2%, respectively.

On qualitative analyses, we observed false positives in regions with physiologically high FDG uptake (e.g., bowel), benign lesions with increased FDG uptake (inflammatory lesions), and physiologic nodal uptake. The false negatives included regions with small lesions (<1 cm), regions with low FDG uptake (SUVmax<2), and most commonly

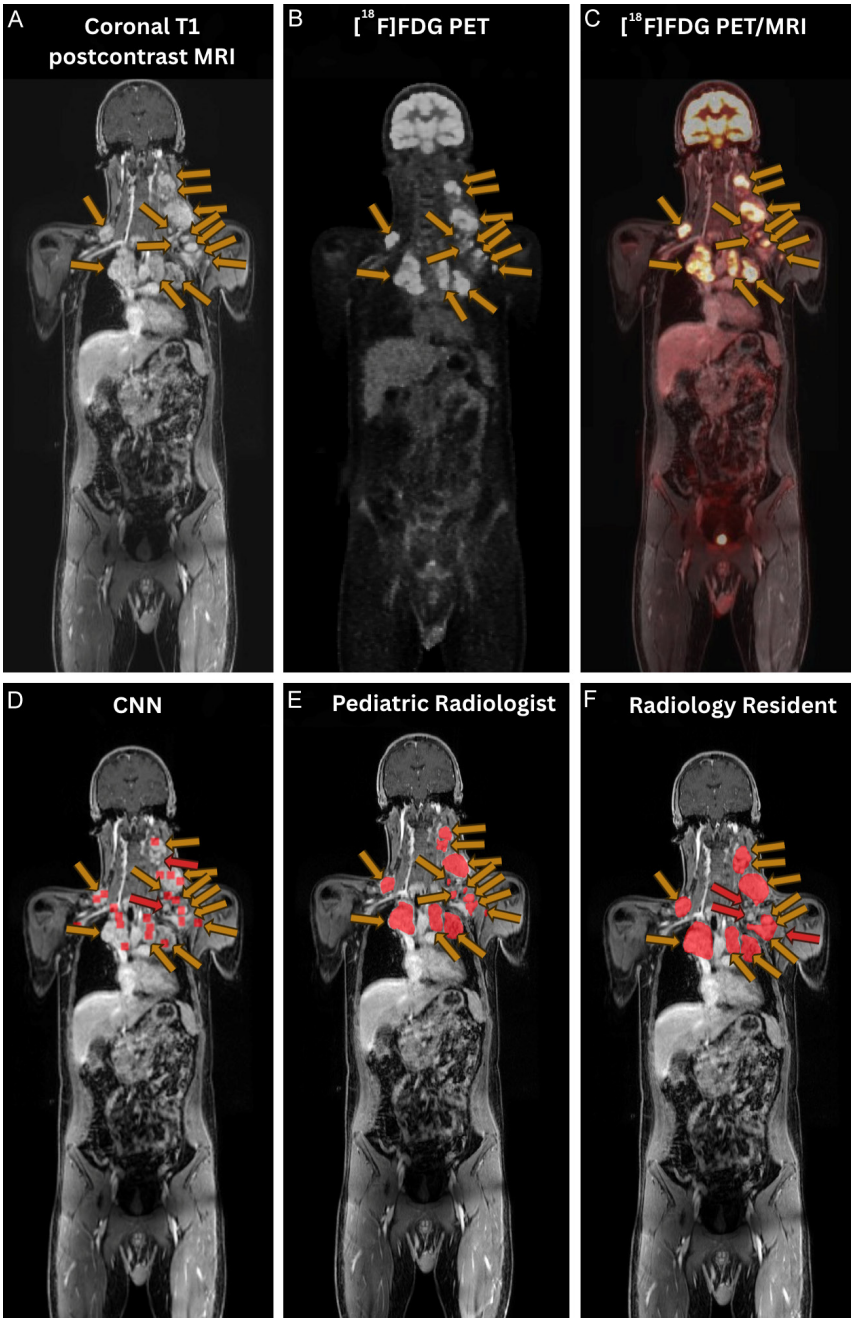

Figure 3. Comparison of tumor detection with CNN and by human readers with respect to the reference standard on [^{18}F]FDG PET/MR image in a 14-year-old patient with non-Hodgkin's lymphoma. A. Coronal contrast-enhanced T1-weighted gradient echo MRI demonstrating multiple enhancing lymphoma lesions (yellow arrows). B. Simultaneously acquired [^{18}F]FDG PET image showing increased radiotracer uptake within lymphoma lesions (yellow arrows). C. Whole-body [^{18}F]FDG PET/MR image illustrating increased radiotracer uptake of lymphoma lesions (yellow arrows). D. CNN-derived annotations of multiple lymphoma lesions (red squares) on the MRI scan; the yellow arrow indicates true positive lesions and the red arrows indicate false negative lesions. E. Annotations by a pediatric radiologist (red regions) with 3 years of experience, showing no missed lesions. F. Annotations by a second-year radiology resident (red regions); yellow arrow indicates true positive lesions, and the red arrows indicate false negative lesions.

Comparison of the sensitivity and specificity of CNN to those of human readers

The CNN demonstrated higher sensitivity, at 84.6%, and specificity, at 93.7%, compared to the radiology resident, who achieved a sensitivity of 69.2% (P=0.023) and a specificity of 81.5% (P<0.001). However, the CNN's performance was lower than that of the pediatric radiologist, who attained a sensitivity of 98.7% (P<0.001) and a specificity of 99.5% (P<0.001) (**Figure 3**). Similar results were obtained for sub-analyses of regions with lymph nodal lesions and extra-nodal lesions. For sub-analysis of regions with lymph nodal lesions, the CNN's sensitivity (84.4%) was lower than that of the pediatric radiologist (100%, *P*=0.016), but its specificity (97.1%) was higher than that of the resident (79%, *P*<0.001). For sub-analysis of regions with extra-nodal lesions, CNN's sensitivity (84.8%) was higher than that of the resident (45.5%, *P*=0.002), and specificity (90.6%) was lower than that of the pediatric radiologist (100%, *P*<0.001) (Supplementary Table 1).

Percentage agreement of CNN or human readers with the reference standard

The CNN demonstrated 83% agreement (95% CI=79%-87%) with the reference standard for detecting regions with lymphoma lesions on [^{18}F]FDG-PET/MR images. By comparison, the percentage agreement between human readouts and the reference standard was 94% (95% CI=92%-97%) for the pediatric radiologist and 63% (95% CI=59%-69%) for the radiology resident (Supplementary Table 2). The percentage agreement between CNN and the reference standard was 85% (95% CI=78%-90%) for the detection of regions with lymph nodal lesions and 81% (95% CI=74%-87%) for regions with extra-nodal lesions. By comparison, the percentage agreement between human readouts and the reference standard was 91% (95% CI=86%-95%) for regions with lymph nodal and 0.97% (95% CI=93%-99%) for regions with extra-nodal lesions for the pediatric radiologist and 61% (95% CI=52%-69%) for regions with lymph nodal and 67% (95% CI=59%-75%) for regions with extra-nodal lesions for the radiology resident.

in the abdomen and pelvis. Among the 12 false negatives, 5 were in the abdomen and pelvis, 2 in the head, 2 in the neck, 2 in the thorax, and 1 in the extremity.

Comparison of the time spent by CNN and human readers for detecting lymphoma lesions

The median values and interquartile ranges (25th, 75th percentile) for the time taken (in minutes) were 8 (7, 10) for the pediatric radiologist, 15 (9, 20) for the radiology resident, and 4 (3, 5) for the CNN. The CNN readout was significantly faster compared to each human reader (P<0.001, in both cases).

Discussion

The sensitivity, specificity, and percentage agreement of the CNN algorithm for detecting regions with pediatric lymphoma lesions on [^{18}F]FDG-PET/MRI were higher compared to those of the radiology resident but lower compared to those of the pediatric radiologist. The CNN detected lesions faster than the radiology resident as well as the pediatric radiologist individually.

Our CNN demonstrated higher sensitivity and specificity compared to the 3D U-Net-based models developed by Blanc-Durand et al. [11] and Jiang et al. [18]. Specifically, our model achieved a sensitivity of 84.6% (95% CI: 74.7%-91.8%) and specificity of 93.7% (95% CI: 89.6%-96.5%), surpassing Blanc-Durand et al.'s voxel-level sensitivity of 75%±22% and specificity of 79%±20%, as well as Jiang et al.'s voxel-level sensitivities of 0.83±0.22 (median 0.93, IQR: 0.75-0.99) in the training cohort and 0.81±0.27 (median 0.91, IQR: 0.77-0.99) in the validation cohort. The superior performance of our CNN likely resulted from several architectural and methodological differences. The integration of multimodal PET/MRI data through learnable weighted feature fusion possibly provided complementary metabolic and anatomical information for lesion detection. In addition, our two-stage lesion detection strategy, involving SUV-guided tumor candidate generation and location-aware classification across 14 anatomical regions, allowed effective reduction of false positives. Moreover, our lesion-level evaluation metric permitted true-positive detection based on partial lesion overlap, whereas the voxel-level metrics used by Blanc-Durand et al. and Jiang et al. imposed stricter criteria, penalizing even minor segmentation inaccuracies. Furthermore, unlike Jiang et al., we also assessed specificity, and unlike both Blanc-Durand et al. and Jiang et al., we assessed processing time and percentage agreement analysis with respect to the reference standard, offering a comprehensive evaluation of clinical applicability.

In contrast to Blanc-Durand et al. and Jiang et al., the DeepSSTL model by Leung et al. achieved higher sensitivity than our CNN, demonstrating a median true-positive rate (TPR) of 87% for lymphoma using voxel-level segmentation versus our tumor-level sensitivity of 84.6% [14]. Their voxel-level segmentation model utilized an nnU-Net architecture, which is consistently recognized for its excellent performance in PET/CT lesion segmentation challenges. Moreover, the DeepSSTL model incorporated semisupervised transfer learning trained on a large, comprehensively annotated [^{18}F]FDG-PET/CT dataset from adults with multiple types of cancers [14]. In comparison, our CNN relied on a simpler 3D CNN classifier for candidate generation and classification using a relatively smaller pediatric lymphoma PET/MRI dataset. Regarding specificity, our CNN explicitly demonstrated a high tumor-level specificity of 93.7% (95% CI: 89.6%-96.5%), whereas Leung et al. did not directly report specificity. Although their model achieved a low false discovery rate (FDR) of 13% for lymphoma, corresponding to a positive predictive value (PPV) of 87%, along with true negative rate (TNR) and negative predictive value (NPV) of 1.00 across all patients with multiple cancer types, the absence of an explicitly quantified specificity metric limits precise quantitative comparison with our algorithm. Leung et al. did not study the percentage agreement analysis and processing time of the algorithm, either.

The processing time of our CNN was notably longer than that of Blanc-Durand et al.'s algorithm [11], with a median of 4 minutes (IQR: 3-5 minutes) per case compared to approximately 30 seconds. Our model incorporated multimodal PET and MRI data, requiring the integration of both metabolic and anatomical inputs, which likely increased preprocessing time and memory usage. Although their algorithm used PET/CT (also a dual imaging modality), MRI typically includes a greater number of image slices than CT, resulting in higher data volume and greater computational complexity. Our algorithm employed a two-stage architecture that first performed SUV-guided candidate generation, followed by region-wise classification using a 3D CNN, which inherently introduces additional computational steps and complexity. The use of anatomical location-aware classifiers and weighted feature fusion further added to the computational burden. In contrast, Blanc-Durand et al. utilized a streamlined end-to-end 3D U-Net architecture (nnU-Net) that performed direct voxel-wise segmentation on resampled PET/CT images with minimal preprocessing and no candidate filtering stage, enabling faster inference [11].

Our study has several strengths. To our knowledge, this is the first study to evaluate the sensitivity, specificity, percentage agreement, and processing time of an AI algorithm for detecting regions with lymphoma on whole-body [^{18}F]FDG-PET/MRI scans of pediatric and young adult populations and compare its performance against two human readers with different levels of experience. In addition, we assessed our CNN's performance in different anatomical regions (head, neck, thorax, abdomen and pelvis, and extremities) and locations (lymph nodes and extra-nodal). Our study is the first to conduct sub-analyses for regions with lymph nodal and extra-nodal lesions separately. Although our CNN was not trained to distinguish nodal vs extra-nodal diseases, we believe it is crucial to evaluate whether the AI algorithm's lesion detection ability may vary in these locations, as the tissue

characteristics of the nodal and extra-nodal lesions are different.

We recognize several limitations in our study. The study involved a small sample size. Pediatric lymphoma datasets are relatively scarce compared to adults, and PET/MRI technology remains limited in availability relative to PET/CT. The CNN was unable to identify all the lesions; detection of all the lesions would have been ideal for clinical application. Detecting tumors on whole-body PET/MRI is inherently challenging due to extensive anatomical coverage, variability in lesion characteristics and locations, physiologic tracer uptake mimicking pathology, and pediatric-specific challenges such as motion artifacts. The limited sample size and performance of CNN could possibly be addressed by employing data augmentation strategies in future studies. Our CNN can only flag lesions, but was not designed to segment the lesions. Adding a segmentation tool could be used for treatment monitoring and radiation therapy planning. Several other authors have developed algorithms for the automatic segmentation of lymphoma on PET/CT scans in adult patients [11, 13, 14, 16, 18], and such tools could be added to our CNN as deemed appropriate. It may also be feasible to integrate an existing segmentation network, such as nnU-Net, into our CNN framework to produce a segmentation mask. The size and shape information provided by the segmentations may be valuable for this downstream task and could be a future extension of our current work. We tested CNN only on an internal dataset and only on baseline [^{18}F] FDG-PET/MRI scans before the start of therapy. Similarly, the small sample size of the test dataset posed a limitation to performing subgroup analyses to evaluate the CNN's performance across demographic variables such as age group, gender, and lymphoma types. Future work includes validating the CNN on multi-institutional datasets, testing it on post-treatment scans, adding an algorithm for lesion segmentation, and adding an algorithm for automated tumor volume and SUV measurements. Future studies with a larger sample size can also test the performance of the model across demographic variables such as age group, gender, and lymphoma types.

In conclusion, the sensitivity and specificity of our CNN-based deep-learning algorithm for automated detection of regions with pediatric lymphoma on whole body [^{18}F] FDG-PET/MRI were significantly higher compared to those of a second-year radiology resident but significantly lower compared to those of a pediatric radiologist with three years of experience. Similarly, CNN outperformed the second-year radiology resident for the detection of regions with lymphoma in terms of percent agreement but was less accurate than the pediatric radiologist. The CNN detected lesions faster than both human readers individually. While it cannot replace expert review, it has the potential to expedite diagnosis, facilitate same-day imaging and clinic visits, and support trainees. With lesion-detection performance intermediate between that of a radiology resident and an experienced pediatric radiologist, the CNN is best positioned to serve as a supervised decision-support tool rather than a stand-alone reader. To facilitate clinical workflow, it could serve as a second reader by providing an optional overlay of candidate regions during interpretation to prompt re-examination of subtle findings. It may also support education and quality assurance by allowing trainees to compare preliminary reads with CNN outputs to identify potentially missed lesions, and by enabling a post-dictation discrepancy check of regions with suspicious but missed lesions that may require confirmatory review. These potential clinical workflow integrations may leverage CNNs' rapid lesion-flagging capability while maintaining radiologist oversight to ensure accuracy and accountability for clinical decision-making. Further research is needed to validate its clinical integration, generalizability, and applicability to other malignancies and adult patients.

Acknowledgements

This work was supported by grants from the National Cancer Institute (Grant number R01CA269231), the Stanford Center for Artificial Intelligence in Medicine & Imaging (AIMI), and the Human-Centered Artificial Intelligence (HAI). In addition, M.B. was supported by a supplement grant to R01CA269231.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Heike E Daldrup-Link, Department of Radiology, Division of Pediatric Radiology, Stanford University School of Medicine, Stanford, CA 94305, USA. Tel: 650-497-8376; Fax: 650-498-9865; E-mail: heiked@stanford.edu

References

[1] Sultan I, Alfaar AS, Sultan Y, Salman Z and Qaddoumi I. Trends in childhood cancer: incidence and survival analysis over 45 years of SEER data. PLoS One 2025; 20: e0314592.

[2] Siegel DA, King JB, Lupo PJ, Durbin EB, Tai E, Mills K, Van Dyne E, Buchanan Lunsford N, Henley SJ and Wilson RJ. Counts, incidence rates, and trends of pediatric cancer in the United States, 2003-2019. J Natl Cancer Inst 2023; 115: 1337-1354.

[3] Seam P, Juweid ME and Cheson BD. The role of FDG-PET scans in patients with lymphoma. Blood 2007; 110: 3507-3516.

[4] Gallamini A and Borra A. Role of PET in lymphoma. Curr Treat Options Oncol 2014; 15: 248-261.

[5] Schafer JF, Gatidis S, Schmidt H, Guckel B, Bezrukov I, Pfannenberg CA, Reimold M, Ebinger M, Fuchs J, Claussen CD and Schwenzer NF. Simultaneous whole-body PET/MR imaging in comparison to PET/CT in pediatric oncology: initial results. Radiology 2014; 273: 220-231.

[6] Heacock L, Weissbrot J, Raad R, Campbell N, Friedman KP, Ponzo F and Chandarana H. PET/MRI for the evaluation of

patients with lymphoma: initial observations. AJR Am J Roentgenol 2015; 204: 842-848.
[7] Afaq A, Fraioli F, Sidhu H, Wan S, Punwani S, Chen SH, Akin O, Linch D, Ardeshna K, Lambert J, Miles K, Groves A and Kayani I. Comparison of PET/MRI with PET/CT in the evaluation of disease status in lymphoma. Clin Nucl Med 2017; 42: e1-e7.
[8] Sher AC, Seghers V, Paldino MJ, Dodge C, Krishnamurthy R, Krishnamurthy R and Rohren EM. Assessment of sequential PET/MRI in comparison with PET/CT of pediatric lymphoma: a prospective study. AJR Am J Roentgenol 2016; 206: 623-631.
[9] Kirchner J, Deuschl C, Schweiger B, Herrmann K, Forsting M, Buchbender C, Antoch G and Umutlu L. Imaging children suffering from lymphoma: an evaluation of different (18)F-FDG PET/MRI protocols compared to whole-body DW-MRI. Eur J Nucl Med Mol Imaging 2017; 44: 1742-1750.
[10] Picardi M, Cavaliere C, Della Pepa R, Nicolai E, Soricelli A, Giordano C, Pugliese N, Rascato MG, Cappuccio I, Campagna G, Cerchione C, Vigliar E, Troncone G, Mascolo M, Franzese M, Castaldo R, Salvatore M and Pane F. PET/MRI for staging patients with Hodgkin lymphoma: equivalent results with PET/CT in a prospective trial. Ann Hematol 2021; 100: 1525-1535.
[11] Blanc-Durand P, Jegou S, Kanoun S, Berriolo-Riedinger A, Bodet-Milin C, Kraeber-Bodere F, Carlier T, Le Gouill S, Casasnovas RO, Meignan M and Itti E. Fully automatic segmentation of diffuse large B cell lymphoma lesions on 3D FDG-PET/CT for total metabolic tumour volume prediction using a convolutional neural network. Eur J Nucl Med Mol Imaging 2021; 48: 1362-1370.
[12] Sibille L, Seifert R, Avramovic N, Vehren T, Spottiswoode B, Zuehlsdorff S and Schafers M. (18)F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. Radiology 2020; 294: 445-452.
[13] Yousefirizi F, Klyuzhin IS, O JH, Harsini S, Tie X, Shiri I, Shin M, Lee C, Cho SY, Bradshaw TJ, Zaidi H, Benard F, Sehn LH, Savage KJ, Steidl C, Uribe CF and Rahmim A. TMTV-Net: fully automated total metabolic tumor volume segmentation in lymphoma PET/CT images - a multi-center generalizability analysis. Eur J Nucl Med Mol Imaging 2024; 51: 1937-1954.
[14] Leung KH, Rowe SP, Sadaghiani MS, Leal JP, Mena E, Choyke PL, Du Y and Pomper MG. Deep semisupervised transfer learning for fully automated whole-body tumor quantification and prognosis of cancer on PET/CT. J Nucl Med 2024; 65: 643-650.
[15] Zhou Z, Jain P, Lu Y, Macapinlac H, Wang ML, Son JB, Pagel MD, Xu G and Ma J. Computer-aided detection of mantle cell lymphoma on (18)F-FDG PET/CT using a deep learning convolutional neural network. Am J Nucl Med Mol Imaging 2021; 11: 260-270.
[16] Yuan C, Zhang M, Huang X, Xie W, Lin X, Zhao W, Li B and Qian D. Diffuse large B-cell lymphoma segmentation in PET-CT images via hybrid learning for feature fusion. Med Phys 2021; 48: 3665-3678.
[17] Sadik M, Barrington SF, Tragardh E, Saboury B, Nielsen AL, Jakobsen AL, Gongora JLL, Urdaneta JL, Kumar R and Edenbrandt L. Metabolic tumour volume in Hodgkin lymphoma-A comparison between manual and AI-based analysis. Clin Physiol Funct Imaging 2024; 44: 220-227.
[18] Jiang C, Chen K, Teng Y, Ding C, Zhou Z, Gao Y, Wu J, He J, He K and Zhang J. Deep learning-based tumour segmentation and total metabolic tumour volume prediction in the prognosis of diffuse large B-cell lymphoma patients in 3D FDG-PET images. Eur Radiol 2022; 32: 4801-4812.
[19] Zhang Y, Deng Y, Zou Q, Jing B, Cai P, Tian X, Yang Y, Li B, Liu F, Li Z, Liu Z, Feng S, Peng T, Dong Y, Wang X, Ruan G, He Y, Cui C, Li J, Luo X, Huang H, Chen H, Li S, Sun Y, Xie C, Wang L, Li C and Cai Q. Artificial intelligence for diagnosis and prognosis prediction of natural killer/T cell lymphoma using magnetic resonance imaging. Cell Rep Med 2024; 5: 101551.
[20] Singh SB, Sarrami AH, Gatidis S, Varniab ZS, Chaudhari A and Daldrup-Link HE. Applications of artificial intelligence for pediatric cancer imaging. AJR Am J Roentgenol 2024; 223: e2431076.
[21] Tie X, Shin M, Lee C, Perlman SB, Huemann Z, Weisman AJ, Castellino SM, Kelly KM, McCarten KM, Alazraki AL, Hu J, Cho SY and Bradshaw TJ. Automatic quantification of serial PET/CT images for pediatric hodgkin lymphoma using a longitudinally aware segmentation network. Radiol Artif Intell 2025; 7: e240229.
[22] Etchebehere E, Andrade R, Camacho M, Lima M, Brink A, Cerci J, Nadel H, Bal C, Rangarajan V, Pfluger T, Kagna O, Alonso O, Begum FK, Mir KB, Magboo VP, Menezes LJ, Paez D and Pascual TN. Validation of convolutional neural networks for fast determination of whole-body metabolic tumor burden in pediatric lymphoma. J Nucl Med Technol 2022; 50: 256-262.
[23] Wang H, Sarrami A, Wu JT, Baratto L, Sharma A, Wong KCL, Singh SB, Daldrup-Link HE and Syeda-Mahmood T. Multimodal pediatric lymphoma detection using PET and MRI. AMIA Annu Symp Proc 2024; 2023: 736-743.
[24] Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC and Gerig G. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006; 31: 1116-1128.
[25] Shim MS, Zhao C, Li Y, Zhang X, Zhang W and Li P. Robust deep multi-modal sensor fusion using fusion weight regularization and target learning. arXiv preprint arXiv: 2019; 1901: 10610.
[26] Tashu TM, Hajiyeva S and Horvath T. Multimodal emotion recognition from art using sequential co-attention. J Imaging 2021; 7: 157.
[27] He K, Zhang X, Ren S and Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 2015; 37: 1904-1916.
[28] Huo Y, Xu Z, Xiong Y, Aboud K, Parvathaneni P, Bao S, Bermudez C, Resnick SM, Cutting LE and Landman BA. 3D whole brain segmentation using spatially localized atlas network tiles. Neuroimage 2019; 194: 105-119.
[29] Joshi S, Davis B, Jomier M and Gerig G. Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 2004; 23 Suppl 1: S151-160.
[30] Jenkinson M, Bannister P, Brady M and Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 2002; 17: 825-841.

Supplementary Table 1. Sensitivity and specificity for detecting nodal and extra-nodal lesions by the CNN, pediatric radiologist, and the radiology resident. All the numbers are indicated in percentage, and the numbers in brackets indicate the confidence interval

		CNN	Pediatric radiologist	Radiology resident
Sensitivity % (95% Confidence interval)	Nodal lesions	84.4 (70.9-92.4)	100*	86.7 (73.7-93.8)
	P-value (CNN Vs human reader for nodal lesions)	N/A	0.016	1
	Extra-nodal lesions	84.8 (68.8-93.4)	97 (79.2-99.6)	45.5 (29.5-62.4)*
	P-value (CNN Vs human reader for extra-nodal lesions)	N/A	0.13	0.002
	Nodal + Extra-nodal lesions	84.6 (74.7-91.8)	98.7 (93.1-100)*	69.2 (57.8-79.2)*
	P-value (CNN Vs human reader for nodal + extra-nodal lesions)	N/A	<0.001	0.023
Specificity % (95% Confidence interval)	Nodal lesions	97.1 (91.4-99.1)	99 (93.1-99.9)	79 (67.0-87.5)*
	P-value (CNN Vs human reader for nodal lesions)	N/A	0.5	<0.001
	Extra-nodal lesions	90.6 (83.2-94.9)	100*	83.8 (76.6-89.0)
	P-value (CNN Vs human reader for extra-nodal lesions)	N/A	<0.001	0.18
	Nodal + Extra-nodal lesions	93.7 (89.6-96.5)	99.5 (97.5-100)*	81.5 (75.8-86.4)*
	P-value (CNN Vs human reader for nodal + extra-nodal lesions)	N/A	<0.001	<0.001

* indicates statistically significant differences when human readers' sensitivity and specificity were compared with CNN.

Supplementary Table 2. Percentage agreement between the CNN and human readers with the reference standard in all five body regions, subdivided among two locations - nodal and extra-nodal lesions

Region	Location	Truth vs CNN: percent agreement (95% CI)	Truth vs Pediatric radiologist: percent agreement (95% CI)	Truth vs Radiology resident: percent agreement (95% CI)
Head	Nodal lesions	1.00 (0.88-1.00)	1.00 (0.88-1.00)	0.97 (0.83-1.00)
Head	Extra-nodal lesions	0.93 (0.78-0.99)	0.97 (0.83-1.00)	0.43 (0.25-0.63)
Neck	Nodal lesions	0.70 (0.51-0.85)	0.77 (0.58-0.90)	0.23 (0.10-0.42)
Neck	Extra-nodal lesions	0.97 (0.83-1.00)	1.00 (0.88-1.00)	0.87 (0.69-0.96)
Thorax	Nodal lesions	0.73 (0.54-0.88)	0.93 (0.78-0.99)	0.40 (0.23-0.59)
Thorax	Extra-nodal lesions	0.77 (0.58-0.90)	1.00 (0.88-1.00)	0.47 (0.28-0.66)
Abdomen and pelvis	Nodal lesions	0.80 (0.61-0.92)	0.87 (0.69-0.96)	0.50 (0.31-0.69)
Abdomen and pelvis	Extra-nodal lesions	0.70 (0.51-0.85)	0.97 (0.83-1.00)	0.77 (0.58-0.90)
Extremities	Nodal lesions	1.00 (0.88-1.00)	1.00 (0.88-1.00)	0.93 (0.78-0.99)
Extremities	Extra-nodal lesions	0.70 (0.51-0.85)	0.93 (0.78-0.99)	0.83 (0.65-0.94)
Overall percent agreement	Nodal lesions	0.85 (0.78-0.90)	0.91 (0.86-0.95)	0.61 (0.52-0.69)
	Extra-nodal lesions	0.81 (0.74-0.87)	0.97 (0.93-0.99)	0.67 (0.59-0.75)
	Nodal + Extra-nodal lesions	0.83 (0.79-0.87)	0.94 (0.92-0.97)	0.63 (0.59-0.69)