Annals of Nuclear Medicine
https://doi.org/10.1007/s12149-023-01866-5

ORIGINAL ARTICLE

Convolutional neural network-based program to predict lymph node metastasis of non-small cell lung cancer using ^{18}F-FDG PET

Eitaro Kidera[1,2] · Sho Koyasu[2] · Kenji Hirata[3] · Masatsugu Hamaji[4] · Ryusuke Nakamoto[2] · Yuji Nakamoto[2]

Received: 6 May 2023 / Accepted: 11 September 2023

Abstract

Purpose To develop a convolutional neural network (CNN)-based program to analyze maximum intensity projection (MIP) images of 2-deoxy-2-[F-18]fluoro-D-glucose (FDG) positron emission tomography (PET) scans, aimed at predicting lymph node metastasis of non-small cell lung cancer (NSCLC), and to evaluate its effectiveness in providing diagnostic assistance to radiologists.

Methods We obtained PET images of NSCLC from public datasets, including those of 435 patients with available N-stage information, which were divided into a training set ($n = 304$) and a test set ($n = 131$). We generated 36 maximum intensity projection (MIP) images for each patient. A residual network (ResNet-50)-based CNN was trained using the MIP images of the training set to predict lymph node metastasis. Lymph node metastasis in the test set was predicted by the trained CNN as well as by seven radiologists twice: first without and second with CNN assistance. Diagnostic performance metrics, including accuracy and prediction error (the difference between the truth and the predictions), were calculated, and reading times were recorded.

Results In the test set, 67 (51%) patients exhibited lymph node metastases and the CNN yielded 0.748 predictive accuracy. With the assistance of the CNN, the prediction error was significantly reduced for six of the seven radiologists although the accuracy did not change significantly. The prediction time was significantly reduced for five of the seven radiologists with the median reduction ratio 38.0%.

Conclusion The CNN-based program could potentially assist radiologists in predicting lymph node metastasis by increasing diagnostic confidence and reducing reading time without affecting diagnostic accuracy, at least in the limited situations using MIP images.

Keywords Deep learning · Lymph node metastasis · Non-small cell lung cancer · Positron emission tomography

Introduction

Lung cancer is the leading cause of cancer-related death in the United States, with 237,000 new cases and 130,000 estimated deaths in 2022 [1]. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer, including more specific types, such as adenocarcinoma and squamous cell carcinoma [2]. Accurate diagnosis of hilar or mediastinal lymph node metastasis is essential to determine the treatment strategy for NSCLC, which includes surgery, radiation therapy, and chemotherapy [3]. The American College of Chest Physicians guidelines recommend positron emission tomography (PET) with 2-deoxy-2-[F-18]fluoro-D-glucose (FDG) for staging NSCLC [4]. However, the diagnostic performance of FDG-PET for hilar or mediastinal lymph node metastasis of NSCLC varied among reports, and the

✉ Sho Koyasu
sho@kuhp.kyoto-u.ac.jp; koyasusho@gmail.com

1 Department of Radiology, Kishiwada City Hospital, Kishiwada, Japan

2 Department of Diagnostic Imaging and Nuclear Medicine, Graduate School of Medicine, Kyoto University, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan

3 Department of Diagnostic Imaging, Graduate School of Medicine, Hokkaido University, Sapporo, Japan

4 Department of Thoracic Surgery, Kyoto University Hospital, Kyoto University, Kyoto, Japan

Published online: 27 September 2023

sensitivity and the specificity were not very high even with FDG-PET. According to Schimmer et al., FDG PET with or without integrated CT had a sensitivity of 58–94% and a specificity of 76–96% in detecting mediastinal lymph node metastasis [5]. Subsequently, low sensitivities and high specificities of FDG PET/CT in surgical cases were reported: Billé et al. reported a sensitivity of 54.2% and specificity of 91.9% [6], and Ose et al. reported a sensitivity of 50.0% and specificity of 94.5% [7]. Improving the diagnostic performance of FDG-PET in predicting lymph node metastasis is clinically important. For example, if the absence of lymph node metastasis could be accurately diagnosed using preoperative FDG PET scans, surgeons could select sublobar resection, such as segmentectomy and wedge resection with or without selective lymph node dissection, which may be associated with fewer postoperative complications [8, 9]. In contrast, if hilar or mediastinal lymph node metastasis was detected on preoperative FDG PET scans, induction chemotherapy or other therapies would be considered.

In recent years, machine learning methods have been introduced for oncologic FDG PET imaging [10]. More recently, deep learning methods using convolutional neural network (CNN) were introduced to analyze medical images, as well as to predict lymph node metastasis of NSCLC [11–13]. In this study, we focused on the use of maximum intensity projection (MIP) images from FDG PET. This may address the class imbalance problem of simply providing 2D slices of FDG PET to the CNN, that is, because most 2D slices of FDG PET do not include abnormal findings, CNN tends to classify more 2D slices as normal than actual. Kawauchi et al. developed a CNN-based system to classify FDG PET/CT examinations into benign, malignant, or equivocal FDG PET images, and reported accuracies of 99.4, 99.4, and 87.5% for benign, malignant, and equivocal images, respectively [14]. They preferred MIP images to 2D slices regarding the class imbalance problem, because most MIP images of malignant patients contain FDG accumulation. We also focused on diagnostic assistance for radiologists using the CNN. The frequent use of FDG-PET to examine malignant tumors has increased the burden on radiologists. Approximately two million PET scans are performed annually in the United States according to the Lawrence Berkeley National Laboratory [15]. In Japan, the total number of FDG PET studies in 2017 was 630 570, a 24.5% increase compared to the result in 2012 [16]. Machine learning methods are expected to assist radiologists beyond simply making diagnoses, as their workload increases [17].

We hypothesized that a CNN-based system analyzing MIP images of FDG-PET would improve the diagnostic performance for lymph node metastasis in NSCLC. In addition, we expected that the system would reduce the reading time of the radiologists. This study aimed to develop a CNN-based program analyzing MIP images of FDG PET to predict lymph node metastasis of NSCLC and to evaluate whether it could provide diagnostic assistance to radiologists.

Materials and methods

Patients

This study used FDG PET images obtained from The Cancer Imaging Archive (TCIA) [18], a public database of medical images. Due to the anonymous and public nature of the TCIA, ethics committee approval was not required under the regulations of our country.

We obtained data from six databases in TCIA: ACRIN-NSCLC-FDG-PET ($n = 232$), NSCLC Radiogenomics ($n = 153$), CPTAC-LSCC ($n = 7$), CPTAC-LUAD ($n = 3$), TCGA-LUSC ($n = 23$), and TCGA-LUAD ($n = 17$). The details of these databases are described elsewhere [19–26]. The inclusion criteria were as follows: (1) available FDG PET images obtained using dedicated PET scanners or combined PET/CT scanners and (2) available N stage information. The N stage diagnostic criteria were according to AJCC 5th Edition ($n = 232$), 6th Edition ($n = 16$), 7th Edition ($n = 16$), 8th Edition ($n = 5$), or unavailable ($n = 166$). This study focused on whether lymph node metastasis was present (N0 or N1–3). Therefore, it was unaffected by the differences among the diagnostic criteria. Regarding ACRIN-NSCLC-FDG-PET, the protocol for nodal metastasis diagnosis was based on clinical information including conventional imaging. The other five databases provided the pathological N stage.

The cohort was randomly divided into a training set (70%) and test set (30%).

A schematic representation of the patient selection protocol is illustrated in Fig. 1.

Data preprocessing

For each patient, we generated MIP images of the thoracic region from the PET images at increments of 10º rotations up to 360° (0°, 10°, …, 350°). The thoracic regions were identified by a board-certified radiologist with 7 years of experience. The SUV range of 0–5 was converted linearly to greyscale of 0–1 for every patient. The MIP images were resized to (224, 224) pixels and converted from greyscale to RGB to match the input shape of the pre-trained base model described below.

For each patient, the N stage was reclassified into two categories: N0 and N1–3.

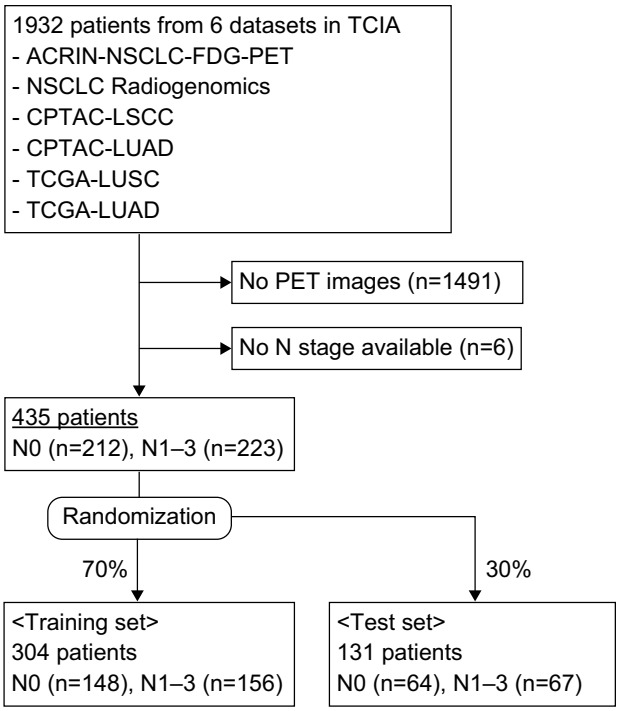

Fig. 1 Flow chart showing the patient selection protocol

Architecture of the CNN

We used the pre-trained ResNet-50 [27] as the base model. ResNet-50 is a 50-layer CNN using "residual" blocks to solve the gradient disappearance problem and has an input shape of (224, 224, 3) and 1000 output classes. We removed the original output layer and appended a fully connected layer comprising 256 neurons and an output layer of two categories corresponding to N0 and N1–3 (Fig. 2a). We also used the transfer-learning technique because of the relatively small size of the dataset. That is, the layers from the original ResNet-50 were initialized and frozen with weights from pre-training on ImageNet, and only the appended layers were trained on our dataset. Fivefold cross-validation was used to optimize hyperparameters, such as batch size and epochs.

To confirm that ResNet-50 was an appropriate base model for this study, we compared it to VGG-16 and DenseNet, which were used in previous studies dealing with NSCLC and FDG PET [11, 28]. The mean accuracy of fivefold cross-validation in the training set was 0.721, 0.715, and 0.716 for ResNet-50, VGG-16, and DenseNet, respectively. Because the differences were small and ResNet-50 was already used in a previous study dealing with MIP images of FDG PET [14], we selected ResNet-50 as the base model.

We applied Grad-CAM and guided Grad-CAM [29] to the CNN to identify which part of the MIP image the CNN focused on (Supplementary Fig. 1).

Prediction by the CNN

We trained the CNN using MIP images of patients in the training set to predict N0 or N1–3 for each MIP image. It should be noted that MIP images of 180° to 350° rotations were horizontal flips of MIP images of 0° to 170° rotations and could be considered data augmentation. To predict N0 or N1–3 for each patient, we introduced weighting of the MIP angles. For each patient, the CNN returned 36 values, representing the probabilities of N1–3 for 36 MIP angles (0°, 10°, …, 350°). The probability of N1–3 for the patient was calculated as a weighted average of the probabilities of N1–3 for 36 MIP angles with optimized weights. The optimized weights were identical throughout this study and calculated as follows: (1) out-of-fold predictions were obtained for all the MIP images in the training set using fivefold cross-validation, and a correlation matrix of probabilities of N1–3 for 36 MIP angles was created; (2) weights were optimized by minimizing the mean squared error (MSE) of the patient-based predictions defined above on the training set. The MSE was defined as $\frac{1}{n}\sum_{i=1}^{n}\left(P_i - T_i\right)^2$ where P_i is the probability of N1–3 calculated as a weighted average, and $T_i = 0$ for N0 and 1 for N1–3 for the ith patient.

After training the CNN and optimizing the weights using the training set, we predicted N0 or N1–3 for patients in the test set. A probability of N1–3 greater than 0.5 was considered a positive prediction.

Prediction by the radiologists

For each patient in the test set, four board-certified diagnostic radiologists (Exp1, Exp2, Exp3 and Exp4, with 13, 10, 20 and 13 years of experience, respectively) and three 2nd-year residents (Res1, Res2 and Res3) predicted N0 or N1–3 using the 36 MIP images. Screenshots of the software used in the experiments are shown in Supplementary Fig. 2. First, the radiologists examined the MIP images and predicted the probability of lymph node metastasis, which was entered using a slider with values ranging from 0 to 100 (Supplementary Fig. 2a). The 2D slices of PET, CT, and fusion were not used. The time required for prediction was also recorded. The time required to move the slider was not included to accurately measure the prediction time. Subsequently, the radiologists predicted N0 or N1–3 for each patient with CNN assistance, that is, referring to the CNN prediction for each patient (Supplementary Fig. 2b). The initial value of the slider is the probability that the radiologist predicted in the first step. Figure 2b illustrates the process diagram showing how the CNN and radiologists predicted N0 or N1–3 for each patient.

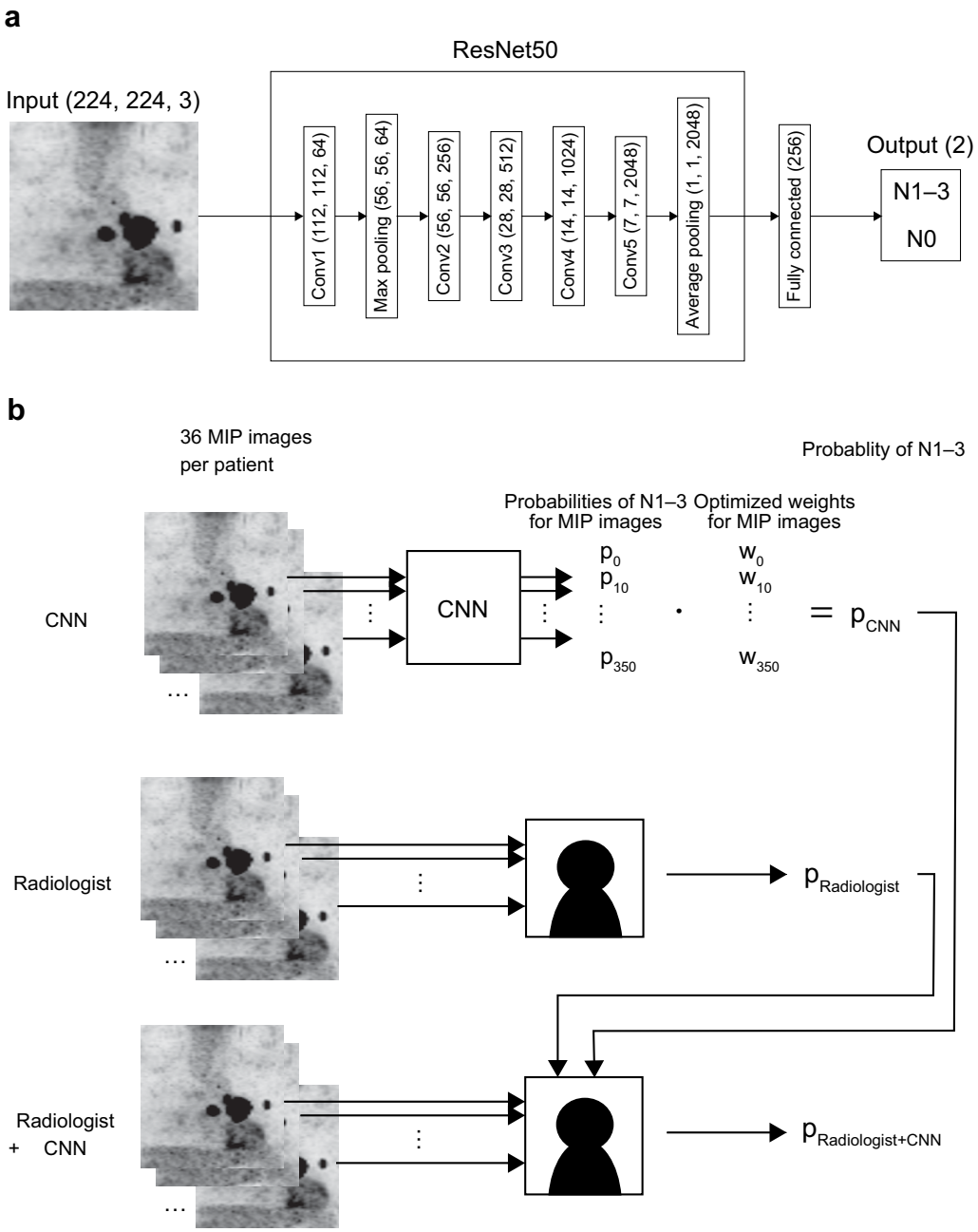

Fig. 2 **a** Functional architecture of the CNN. **b** Process diagram showing how the CNN and the radiologists predicted N0 or N1–3 for each patient

Statistical analysis

Accuracy, sensitivity, and specificity in the prediction of N0 or N1–3 for the test set were calculated for the CNN and the radiologists with and without the assistance of the CNN. For each radiologist, the accuracies were compared between the predictions with and without the assistance of the CNN using the binominal test, and the prediction errors and times were compared using the Wilcoxon signed-rank test. The prediction error was defined as $|P_i - T_i|$ where P_i is the predicted probability of N1–3, and $T_i = 0$ for N0 and 1 for N1–3 for the ith patient. It should be noted that a decrease of the prediction error can be interpreted as an increase of the diagnostic confidence.

Hardware/software environments

Training and testing of the CNN were performed under the following conditions: operating system, Windows 10 64bit; CPU, AMD Ryzen 7 3700X; RAM, DDR4-2666 32 GB; GPU, NVIDIA GeForce RTX 2070 SUPER 8 GB; framework, TensorFlow 2.4; language, Python 3.8. Statistical analysis was performed using R version 4.1 (R Foundation for Statistical Computing, Vienna, Austria).

Results

From six datasets in TCIA, we included 435 patients (men: 290, women: 145; mean age ± SD: 66.5 ± 9.5 years; range 37–87 years). The N stages were N0 ($n = 212$); N1 ($n = 37$); N2 ($n = 83$); and N3 ($n = 103$). The stages were stage I ($n = 136$); stage II ($n = 39$); stage III ($n = 246$); stage IV ($n = 6$); not available ($n = 8$). Table 1 presents the characteristics of the patients in the training and test sets. In the test set, 67 patients (51%) exhibited nodal metastasis (N1–3).

We trained the CNN and optimized its weights using the training set. Figure 3 shows the correlation matrix of the out-of-fold predictions for different MIP angles. The correlations between two angles which differed by nearly 90° were relatively low. Using the trained CNN and optimized weights, we predicted N0 or N1–3 for patients in the test set.

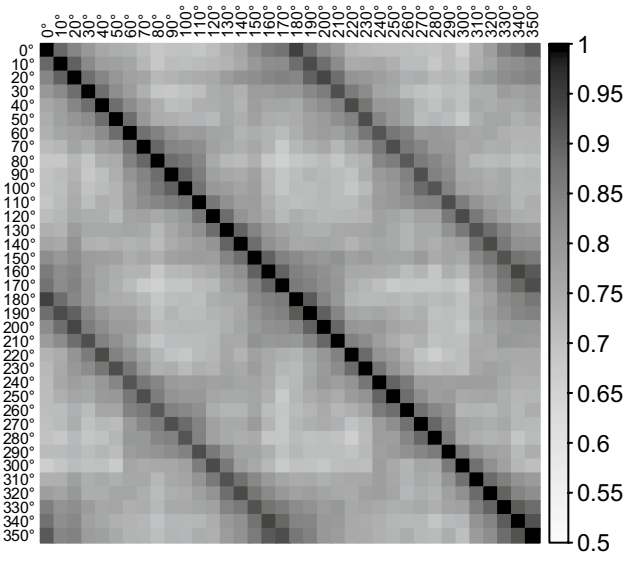

Fig. 3 Correlation matrix of out-of-fold predictions for different MIP angles

Table 2 shows the metrics for the predictions on the test set by CNN and the radiologists with and without the assistance of the CNN. The accuracy of CNN was 0.748. With the assistance of the CNN, the accuracies of the radiologists did not change significantly. The prediction error was reduced for all the seven radiologists and reduced with

Table 1 Patient characteristics of the training and test sets

	Training set ($n = 304$)	Test set ($n = 131$)	p value
Age, median (IQR)	68.0 (62.0–74.0)	67.0 (59.0–72.0)	0.065
Gender			0.18
Male, *n* (%)	209 (69%)	81 (62%)	
Female, *n* (%)	95 (31%)	50 (38%)	
N-stage, *n* (%)			0.53
N0	148 (49%)	64 (49%)	
N1	25 (8.2%)	12 (9.2%)	
N2	54 (18%)	29 (22%)	
N3	77 (25%)	26 (20%)	
Stage, *n* (%)			0.72
I	93 (31%)	43 (33%)	
II	30 (10%)	9 (7.0%)	
III	170 (57%)	76 (59%)	
IV	5 (1.7%)	1 (0.8%)	
Unknown	6	2	
Histology, *n* (%)			0.17
Adenocarcinoma	98 (63%)	40 (61%)	
Squamous cell carcinoma	47 (30%)	25 (38%)	
Other	11 (7.1%)	1 (1.5%)	
Unknown	148	65	

For continuous variables, data are presented as medians (IQR) and compared using Wilcoxon rank-sum test
For categorical variables, data are presented as *n* (%) and compared using Fisher's exact test

Table 2 Performance metrics of the predictions on the test set by the CNN and the radiologists with and without CNN assistance

	Accuracy			Sensitivity		Specificity		Error (mean)			Time (mean [second])		
CNN assistance	No	Yes	*p*	No	Yes	No	Yes	No	Yes	*p*	No	Yes	*p*
CNN	0.748			0.731		0.766		0.292					
Exp1	0.718	0.740	0.615	0.821	0.836	0.609	0.641	0.294	0.276	0.116	11.2	6.9	**< 0.001**
Exp2	0.679	0.733	0.733	0.821	0.806	0.531	0.656	0.370	0.322	**< 0.001**	6.3	3.2	**< 0.001**
Exp3	0.779	0.763	0.375	0.791	0.836	0.766	0.688	0.312	0.292	**0.034**	10.3	7.8	**0.002**
Exp4	0.756	0.748	0.471	0.731	0.791	0.781	0.703	0.323	0.300	**0.027**	9.4	12.0	**< 0.001**
Res1	0.710	0.763	0.630	0.821	0.776	0.594	0.750	0.418	0.344	**< 0.001**	27.2	11.9	**< 0.001**
Res2	0.802	0.771	0.357	0.731	0.731	0.875	0.813	0.377	0.357	**0.033**	8.7	5.3	**0.001**
Res3	0.687	0.756	0.629	0.657	0.746	0.719	0.766	0.402	0.339	**0.001**	28.6	35.5	**0.009**

Bold font indicates statistical significance ($p < 0.05$)

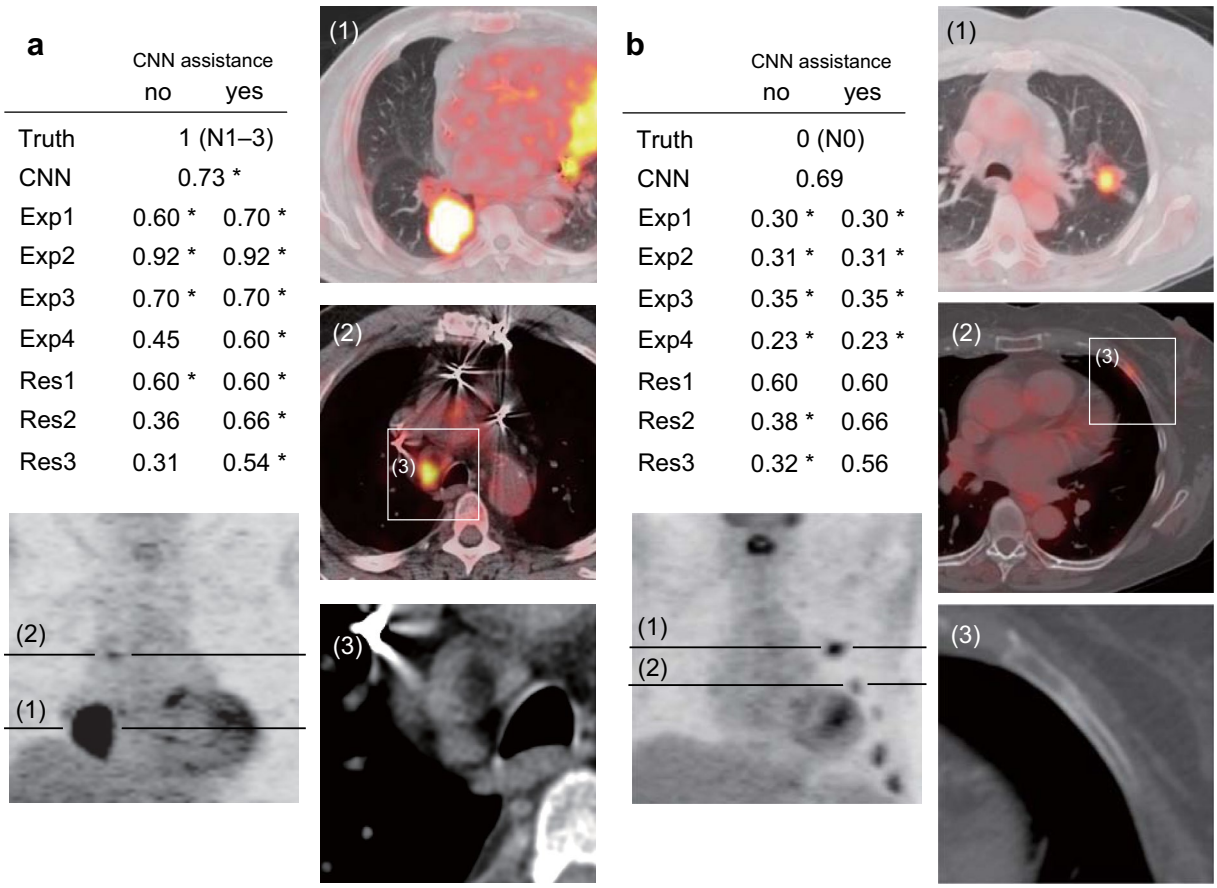

Fig. 4 Representative images with the predictions by the CNN and the radiologists with and without the assistance of the CNN. **a** The CNN and four of the seven radiologists classified the patient as N1–3 correctly. The other three radiologists misclassified the patient as N0 but corrected the prediction with the CNN assistance. **b** The CNN misclassified the patient as N1–3 but all the board-certified radiologists (Exp1–4) correctly classified the patient as N0 without being influenced by the CNN. They determined that the FDG uptake foci were rib fractures from their knowledge and experience. On the other hand, one of the residents (Res1) misclassified the patient as N1–3, and the other two residents (Res2,3) mistakenly changed the prediction from N0 to N1–3 with the CNN assistance. The stage of the patient was IA, confirming the absence of bone metastases. *Accurate prediction

statistical significance for six of the seven radiologists. The prediction time was significantly reduced for five of the seven radiologists, and the median reduction ratio of the prediction time of all seven radiologists was 38.0%. Figure 4 depicts the representative images.

Discussion

We successfully developed a CNN-based program to predict lymph node metastasis in NSCLC. In addition, the prediction error was significantly reduced for six of the seven radiologists and the prediction time was significantly reduced for five of the seven radiologist with CNN assistance. Although no statistical significance was demonstrated for the accuracy, the significant reduction in the prediction error, which means the increase of the diagnostic confidence, could have another clinical value. Our results suggest that the CNN-based program could be a promising tool to support radiologists in predicting lymph node metastasis in NSCLC. In particular, improved diagnostic performance may contribute to individualized surgical management of clinical stage I NSCLC, such as lobectomy versus sublobar resection and systematic lymph node dissection versus selective lymph node dissection.

In this study, we used MIP images rather than 2D or 3D images. Since none of the six datasets from TCIA included annotations, that is, data were unavailable regarding which lymph nodes were involved by the tumor. Hence, we could not label 2D slices as positive or negative. Even if annotations had been provided, MIP images might have been better than 2D images in addressing the class imbalance problem, as discussed by Kawashima et al. [14]. Furthermore, we considered MIP images superior to 2D images in that each MIP image contained information about lesion distribution, for example, unilateral or bilateral. We did not use 3D images for three reasons. First, because our dataset was too small ($n = 304$) for a large number of 3D model parameters, overfitting was inevitable. Second, 3D models require excessive memory and can only be executed in a limited environment. Third, pre-trained 3D models are less readily available than pre-trained 2D models, such as ResNet pre-trained with ImageNet. Girum et al. used MIP images instead of 3D images for the same reasons as our former two reasons. They successfully developed an artificial intelligence (AI) based on the CNN estimating total metabolic tumor volume and tumor dissemination of diffuse large B cell lymphoma from MIP images of FDG PET/CT and confirmed those values as prognostic biomarkers [30]. Finally, we chose ResNet-50 as the base 2D model for two reasons. First, for ResNet-50, weights pre-trained with ImageNet were available. Second, Kawauchi et al. successfully classified MIP images of FDG PET to benign, malignant and equivocal ones with a model based on ResNet-50 [14]. To match the input shape of ResNet-50, (224, 224, 3), we converted greyscale MIP images to RGB images with 224×224 pixels.

We introduced 36 MIP angles instead of one or a few MIP angles, and weighting of MIP angles for the following two reasons. First, due to the small size of our dataset ($n = 435$) for the CNN, it was needed to increase the number of input data. Second, we hypothesized that different MIP angles would yield different diagnostic performances. For example, in lateral views (90° and 270°), uptake in mediastinal lymph nodes might be less distinguishable from primary tumors or physiological mediastinal uptake than in frontal views (0° and 180°), whereas hilar lymph nodes might be easier to distinguish from medially located primary tumors. In the correlation matrix of out-of-fold predictions for different MIP angles (Fig. 3), it was observed that the correlations between the two angles that differed by nearly 90° were relatively low.

We reclassified the N stage into two classes: N0 and N1–3. It is clinically important to distinguish not only between N0 and N1, but also between N1 and N2, and between N2 and N3. However, due to the small size of our dataset ($n = 435$) and the especially small number of N1 cases ($n = 37$, 8.5%), the class imbalance problem would have been inevitable if we used four classes. Therefore, we reduced the number of classes from four to two in this study, but a study with a larger dataset and four classes is needed in future.

Several studies have analyzed FDG PET images with CNN to predict lymph node metastasis in NSCLC. Tau et al. used 2D-CNN to analyze segmented primary tumors in FDG PET images of 264 patients with newly diagnosed NSCLC and achieved a sensitivity of 0.74, specificity of 0.84, and accuracy of 0.80 in fivefold partitioning of 223 patients with histopathologic N category available [11]. It is noteworthy that they achieved equivalent or better results than our results without analyzing the lymph nodes themselves, but it should also be noted that their results were achieved by cross-validation in a single cohort, whereas our results were achieved in a pre-split test cohort. Further, they used a single PET/CT scanner while we collected data from various PET scanners with or without CT. Ouyang et al. analyzed FDG PET/CT images with a 2D-CNN to predict occult lymph node metastasis in patients with clinical N0 adenocarcinoma [12]. Their PET-alone model yielded a sensitivity of 0.75, specificity of 0.63, and accuracy of 0.65; their PET/CT combined model yielded a sensitivity of 0.88, specificity of 0.80, and accuracy of 0.81 in the internal validation set ($n = 60$). We could not simply compare their results with ours because they focused on clinical N0 adenocarcinoma only. Nevertheless, their PET-alone model performed worse than ours in terms of specificity and accuracy, whereas their PET/CT combined model performed better than ours. It is

reasonable to conclude that using both PET and CT data led to better results than using PET data only. However, we did not use CT data because our study design focused on MIP images of FDG PET, and because our datasets included data from dedicated PET scanners. Wallis et al. used 3D-CNN to analyze mediastinal lymph nodes in FDG PET/CT images of 134 NSCLC patients from one scanner and reported a sensitivity of 0.87 and a false-positive rate per patient of 0.41 in the test set ($n = 29$), using the classifications of one experienced radiologist as reference [13]. They also validated the model in a second cohort ($n = 71$) from another scanner, and the corresponding results were 0.53 and 0.24, respectively, without transfer learning, and 0.88 and 0.69, respectively, with transfer learning. They achieved a high sensitivity at the cost of a high false-positive rate in their first cohort, but the reference standard was a single radiologist, while five of the six datasets we used provided pathological N categories. The low sensitivity without transfer learning and the high false-positive rate with transfer learning in their second cohort might suggest insufficient robustness of their model.

The accuracy of our CNN-based program was higher than that of four of the seven radiologists (Table 2). However, this result did not mean that the CNN could outperform the radiologists; only one of the seven radiologists showed lower sensitivities than the CNN, which meant that the radiologists could detect more lymph node metastases. In addition, radiologists might be able to confidently negate lymph node metastasis using their knowledge and experience in some situations, as illustrated in Fig. 4b.

Our CNN-based program significantly reduced the reading time for five of the radiologists with the median reduction radio of 38.0% for all the seven radiologists. Rodríguez-Ruiz et al. reported the breast cancer detection performance of radiologists with mammography unaided versus supported by an AI system based on a CNN [31]. With AI support, the overall reading time did not change significantly, but the reading time in low-suspicion examinations decreased by 11%, whereas that in high-suspicion examinations increased by 2%, which indicated that the reading time would decrease by 4.5%, assuming a real screening scenario. CNN-based programs can improve diagnostic performance and reduce reading time by optimizing the readings of radiologists.

This study had several limitations. Some of these were derived from the nature of the datasets. First, for one of our datasets (ACRIN-NSCLC-FDG-PET), the N stage was determined by clinical information, including conventional imaging, and not by histopathology. However, the "ground truth" issue is not limited to our research [32]. Kelly et al. reported in their meta-analysis that only 9% (46/535) of studies of AI in radiology determined the ground truth from pathologic reports [33]. Even when limited to FDG PET and CNN studies, several studies have used radiologic reports as the ground truth for malignancy of different organs (lung [13], head and neck [34], breast [35]). With regard to patients with NSCLC, not all patients undergo systematic lymph node dissection, and biopsy of all lymph nodes is impractical. Therefore, imaging findings can be a reasonable alternative to the ground truth. The other five datasets provided the pathological N stage. Second, we could not utilize the positions of the lymph node metastases because none of the datasets provided annotations. Therefore, we could not investigate the reasons of misclassification of the CNN or the radiologists. A study with datasets providing annotations will be needed in future. Third, the datasets included FDG-PET images from various PET scanners with or without combined CT, and detailed imaging protocols were unavailable. However, the accuracy of 0.748 yielded from such heterogeneous images might imply the robustness of our model, although we did not perform external validation due to the small size of the datasets. Fourth, this was a retrospective study with a limited number of patients using a public dataset. A prospective, multi-institutional study with a larger cohort is required to confirm our results.

Finally, we should emphasize that in the clinical practice of FDG PET, radiologists do not only make diagnoses from MIP images but also from PET 2D slices, CT, and fusion images. However, impressions from MIP images, which summarize the 3D distribution of FDG, can help radiologists diagnose and avoid oversight and misdiagnoses.

Conclusion

The CNN-based program could potentially improve the diagnostic performance of radiologists for lymph node metastasis in NSCLC by increasing the diagnostic confidence and reducing the reading time without affecting the diagnostic accuracy, at least in the limited situations using MIP images. The CNN-based program could be a promising tool to support radiologists in predicting lymph node metastasis in NSCLC.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12149-023-01866-5.

Acknowledgements The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We would like to express our sincere gratitude to the five readers, Kanae K. Miake, Tomomi W. Nobashi, Mahoto Juuo, Toshiya Takamura, Daiki Toda, Department of Diagnostic Radiology, Kyoto University Hospital, for their invaluable contributions to this research as participants in the reading experiment.

Author contributions EK: data collection, statistical analysis, drafting of the article. SK: data collection, conception and design, interpretation of data, drafting of this article. KH: conception and design, interpretation of data, revision for important intellectual content. MH: revision for important intellectual content. RN: data collection, revision

for important intellectual content. YN: conception and design, final approval of this article. All authors have read and approved the manuscript before submission.

Funding This study was financially supported by JSPS KAKENHI (Grant number 22K15879).

Data availability Raw images and clinical data are available in The Cancer Imaging Archive. Processed images and CNN models are available from the corresponding author on request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72:7–33. https://doi.org/10.3322/caac.21708.
2. Nicholson AG, Tsao MS, Beasley MB, Borczuk AC, Brambilla E, Cooper WA, et al. The 2021 WHO classification of lung tumors: impact of advances since 2015. J Thorac Oncol. 2022;17:362–87. https://doi.org/10.1016/j.jtho.2021.11.003.
3. Ettinger DS, Wood DE, Akerley W, Bazhenova LA, Borghaei H, Camidge DR, et al. Non-small cell lung cancer, version 1.2015. J Natl Compr Canc Netw Version 12015. 2014;12:1738–61. https://doi.org/10.6004/jnccn.2014.0176.
4. Silvestri GA, Gonzalez AV, Jantz MA, Margolis ML, Gould MK, Tanoue LT, et al. Methods for staging non-small cell lung cancer: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013;143(5):e211S – e250. https://doi.org/10.1378/chest.12-2355.
5. Schimmer C, Neukam K, Elert O. Staging of non-small cell lung cancer: clinical value of positron emission tomography and mediastinoscopy. Interact Cardiovasc Thorac Surg. 2006;5:418–23. https://doi.org/10.1510/icvts.2006.129478.
6. Billé A, Pelosi E, Skanjeti A, Arena V, Errico L, Borasio P, et al. Preoperative intrathoracic lymph node staging in patients with non-small-cell lung cancer: accuracy of integrated positron emission tomography and computed tomography. Eur J Cardiothorac Surg. 2009;36:440–5. https://doi.org/10.1016/j.ejcts.2009.04.003.
7. Ose N, Sawabata N, Minami M, Inoue M, Shintani Y, Kadota Y, et al. Lymph node metastasis diagnosis using positron emission tomography with 2-[18F] fluoro-2-deoxy-D-glucose as a tracer and computed tomography in surgical cases of non-small cell lung cancer. Eur J Cardiothorac Surg. 2012;42:89–92. https://doi.org/10.1093/ejcts/ezr287.
8. Wo Y, Li H, Zhang Y, Peng Y, Wu Z, Liu P, et al. The impact of station 4L lymph node dissection on short-term and long-term outcomes in non-small cell lung cancer. Lung Cancer. 2022;170:141–7. https://doi.org/10.1016/j.lungcan.2022.06.018.
9. Zhao Y, Mao Y, He J, Gao S, Zhang Z, Ding N, et al. Lobe-specific lymph node dissection in clinical stage IA solid-dominant non-small-cell lung cancer: a propensity score matching study. Clin Lung Cancer. 2021;22:e201–10. https://doi.org/10.1016/j.cllc.2020.09.012.
10. Sadaghiani MS, Rowe SP, Sheikhbahaei S. Applications of artificial intelligence in oncologic 18F-FDG PET/CT imaging: a systematic review. Ann Transl Med. 2021;9:823. https://doi.org/10.21037/atm-20-6162.
11. Tau N, Stundzia A, Yasufuku K, Hussey D, Metser U. Convolutional neural networks in predicting nodal and distant metastatic potential of newly diagnosed non-small cell lung cancer on FDG PET images. AJR Am J Roentgenol. 2020;215:192–7. https://doi.org/10.2214/AJR.19.22346.
12. Ouyang ML, Zheng RX, Wang YR, Zuo Z, Gu L, Tian Y, et al. Deep learning analysis using 18F-FDG PET/CT to predict occult lymph node metastasis in patients with clinical N0 lung adenocarcinoma. Front Oncol. 2022. https://doi.org/10.3389/fonc.2022.915871.
13. Wallis D, Soussan M, Lacroix M, Akl P, Duboucher C, Buvat I. An [18F]FDG-PET/CT deep learning method for fully automated detection of pathological mediastinal lymph nodes in lung cancer patients. Eur J Nucl Med Mol Imaging. 2022;49:881–8. https://doi.org/10.1007/s00259-021-05513-x.
14. Kawauchi K, Furuya S, Hirata K, Katoh C, Manabe O, Kobayashi K, et al. A convolutional neural network-based system to classify patients using FDG PET/CT examinations. BMC Cancer. 2020;20:227. https://doi.org/10.1186/s12885-020-6694-x.
15. Seeing more with PET scans: scientists discover new chemistry for medical images | Berkeley lab—news center. https://newscenter.lbl.gov/2017/07/27/new-chemistry-pet-scans-medical-imaging/. Accessed 18 Dec 2022
16. Nishiyama Y, Kinuya S, Kato T, Kayano D, Sato S, Tashiro M, et al. Nuclear medicine practice in Japan: a report of the eighth nationwide survey in 2017. Ann Nucl Med. 2019;33:725–32. https://doi.org/10.1007/s12149-019-01382-5.
17. Waller J, O'Connor A, Rafaat E, Amireh A, Dempsey J, Martin C, et al. Applications and challenges of artificial intelligence in diagnostic and interventional radiology. Pol J Radiol. 2022;87:e113–7. https://doi.org/10.5114/pjr.2022.113531.
18. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–57. https://doi.org/10.1007/s10278-013-9622-7.
19. Machtay M, Duan F, Siegel BA, Snyder BS, Gorelick JJ, Reddin JS, et al. Prediction of survival by [18F]fluorodeoxyglucose positron emission tomography in patients with locally advanced non-small-cell lung cancer undergoing definitive chemoradiation therapy: results of the ACRIN 6668/RTOG 0235 trial. J Clin Oncol. 2013;31:3823–30. https://doi.org/10.1200/JCO.2012.47.5947.
20. Kinahan P, Muzi M, Bialecki B, Herman B, Coombs L. Data from the ACRIN 6668 Trial NSCLC-FDG-PET. The Cancer Imaging Arch. 2019. https://doi.org/10.7937/tcia.2019.30ilqfcl.
21. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, et al. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data–methods and preliminary results. Radiology. 2012;264:387–96. https://doi.org/10.1148/radiol.12111607.
22. Bakr S, Gevaert O, Echegaray S, et al. Data for NSCLC radiogenomics collection. Cancer Imaging Arch. 2017. https://doi.org/10.7937/K9/TCIA.2017.7hs46erv.
23. National Cancer Institute clinical proteomic tumor analysis consortium (CPTAC). The clinical proteomic tumor analysis consortium lung squamous cell carcinoma collection (CPTAC-LSCC). 13th version. The Cancer Imaging Arch. 2018. https://doi.org/10.7937/K9/TCIA.2018.6EMUB5L2
24. National Cancer Institute clinical proteomic tumor analysis consortium (CPTAC). The clinical proteomic tumor analysis consortium lung adenocarcinoma collection (CPTAC-LUAD). 11th version. Cancer Imaging Arch. 2018. https://doi.org/10.7937/K9/TCIA.2018.PAT12TBS
25. Kirk S, Lee Y, Kumar P, et al. The cancer genome Atlas lung squamous cell carcinoma collection (TCGA-LUSC). 4th version

. The Cancer Imaging Arch. 2016. https://doi.org/10.7937/K9/TCIA.2016.TYGKKFMQ
26. Albertina B, Watson M, Holback C, et al. Radiology Data from the Cancer Genome Atlas Lung adenocarcinoma [TCGA-LUAD] collection. The Cancer Imaging Arch. 2016. https://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of IEEE comput soc conf comput vis pattern recognit; 2016;2016-December;p. 770–8. https://doi.org/10.1109/CVPR.2016.90
28. Han Y, Ma Y, Wu Z, et al. Histologic subtype classification of non-small cell lung cancer using PET/CT images. Eur J Nucl Med Mol Imaging. 2020;48(2):350–60. https://doi.org/10.1007/S00259-020-04771-5.
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2016. https://doi.org/10.1007/s11263-019-01228-7.
30. Girum KB, Rebaud L, Cottereau AS, Meignan M, Clerc J, Vercellino L, et al. 18F-FDG PET maximum-intensity projections and artificial intelligence: a win-win combination to easily measure prognostic biomarkers in DLBCL patients. J Nucl Med. 2022;63:1925–32. https://doi.org/10.2967/jnumed.121.263501.
31. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology. 2019;290:305–14. https://doi.org/10.1148/radiol.2018181371.
32. Lebovitz S, Levina N, Lifshitz-Assa H. Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. MIS Q. 2021;45:1501–26. https://doi.org/10.25300/MISQ/2021/16564.
33. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol. 2022;32:7998–8007. https://doi.org/10.1007/s00330-022-08784-6.
34. Chen L, Zhou Z, Sher D, Zhang Q, Shah J, Pham NL, et al. Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. Phys Med Biol. 2019;64: 075011. https://doi.org/10.1088/1361-6560/ab083a.
35. Leal JP, Rowe SP, Stearns V, Connolly RM, Vaklavas C, Liu MC, et al. Automated lesion detection of breast cancer in [18F] FDG PET/CT using a novel AI-Based workflow. Front Oncol. 2022;12:1007874. https://doi.org/10.3389/fonc.2022.1007874.