JMI Journal of Medical Imaging

RESEARCH PAPER

Deep learning-based mediastinal lymph node assessment on PET/CT images without pixel-level annotations

Sofija Engelson,[a,b,*] Yannic Elser,[c] Malte Maria Sieren,[c,d] Jan Ehrhardt,[a,b] Julia Andresen,[a] Stefanie Schierholz,[e] Tobias Keck,[e,f] Daniel Drömann,[g] Jörg Barkhausen,[c] and Heinz Handels[a,b]

[a]University of Lübeck, Institute of Medical Informatics, Medical Image Computing and Artificial Intelligence, Lübeck, Germany
[b]German Research Center for Artificial Intelligence, AI in Medical Image and Signal Processing, Lübeck, Germany
[c]University Medical Center Schleswig-Holstein, Department of Radiology and Nuclear Medicine, Lübeck, Germany
[d]University Medical Center Schleswig-Holstein, Department of Interventional Radiology, Lübeck, Germany
[e]University Medical Center Schleswig-Holstein, Department of Surgery, Lübeck, Germany
[f]Fraunhofer Research Institution for Individualized and Cell-Based Medical Engineering IMTE, Lübeck, Germany
[g]University Medical Center Schleswig-Holstein, Department of Pulmonology, Lübeck, Germany

ABSTRACT. **Purpose:** *N*-staging, a critical component in cancer diagnostics, quantifies metastatic involvement of lymph nodes and plays an important role in guiding treatment decisions. Manual assessment of lymph nodes on PET/CT scans is time-consuming due to minimal contrast to surrounding tissue and strong heterogeneity of the lymph node's morphology. To streamline the *N*-staging process, we propose a deep learning-based algorithm that localizes lymph node stations through atlas-to-patient registration, classifies mediastinal lymph node stations as malignant or benign, and subsequently performs automated *N*-staging. Notably, our model is trained without any pixel-level annotations, i.e., using image-level classification labels only.

Approach: To address the challenge of training without annotations at the pixel level, we use prior knowledge of the lymph node station locations through atlas-to-patient registration and deduce pseudo-labels for lymph node station groups from the *N*-stage to enable weakly supervised network training.

Results: The proposed algorithm achieves an accuracy of 0.88 ± 0.02, a sensitivity of 0.72 ± 0.08, and a specificity of 0.90 ± 0.03 for lymph node station classification, which is significantly better than the performance of the standard threshold-based approach used for lymph node assessment in radiological images and an algorithm for PET lesion segmentation that was trained with segmentation masks. For automatic *N*-staging, the accuracy of 0.63 ± 0.04 is on par with an algorithm that was trained with segmentation masks.

Conclusions: The division of the problem setting into subtasks as well as the integration of prior knowledge enables better or comparable performance of models trained with and without segmentation masks.

 [DOI: 10.1117/1.JMI.13.1.014507]

Keywords: weakly supervised learning; mediastinal lymph nodes; *N*-staging; deep learning; image-level labels; priors

*Address all correspondence to Sofija Engelson, sofija.engelson@uni-luebeck.de

Paper 25286GRR received Jul. 28, 2025; revised Jan. 19, 2026; accepted Jan. 21, 2026; published Feb. 18, 2026.

1 Introduction

One of the most commonly used staging systems for cancer is the TNM classification of malignant tumors (TNM), wherein the *N*-staging sheds light on the infestation of metastases in regional lymph nodes. The four *N*-stages quantify the metastatic spread to the regional lymph nodes on an ordinal scale. The differentiation between malignant and benign lymph nodes (and the subsequent *N*-staging) can be determined in multiple ways. A histopathological examination after surgical resection or biopsy, e.g., through mediastinoscopy, of the lymph nodes is the most reliable way to learn the malignancy status.[1] Oftentimes, before the use of invasive procedures, imaging procedures such as CT or PET/CT pose a visualization for metastatic spread to the lymph nodes. Accurately identifying metastatic lymph nodes in CT is challenging due to minimal contrast with surrounding tissue and variations in lymph node shape, size, number, and location. The CT image shows morphological factors; PET/CT additionally visualizes abnormal metabolic activity and is, thus, considered to be the gold standard for image-based diagnosis.

Mountain and Dresler[2] define a map for mediastinal lymph nodes, in which single lymph nodes are grouped into lymph node stations (LNS). For *N*-staging, an even higher aggregation level is required, that is, multiple LNS form an LNS group. The LNS groups are defined as follows: 1, ipsilateral peribronchial and/or hilar lymph nodes; 2, ipsilateral mediastinal and/or subcarinal lymph nodes; and 3, contralateral mediastinal, contralateral hilar ipsilateral, or contralateral scalene, or supraclavicular lymph nodes (see Table 1 and Fig. 1).[3]

Manual assessment of lymph nodes based on CT or PET/CT imaging is resource-intensive because lymph nodes and LNS have to be located before determining their malignancy status. With the rise in popularity of AI and publicly available datasets in the area of medical imaging, there is potential of solving difficult tasks such as lymph node assessment with algorithms aiming to automate tumor staging and support decision-making regarding surgery and further treatment. Algorithms should follow two objectives: on the one hand, they should have a robust prediction performance, potentially learning from relevant features that are unknown to humans. On the other hand, to be useful in practice, algorithms should require as little human interaction as possible. Ideally, pixel-level annotation in the radiological images should not be a requirement to generate input data for the algorithm. Here, pixel-level annotation is defined as manual labeling of a location in the radiological image, that is a segmentation mask, a region of interest, or the center point of a lymph node or lymph node station. By contrast, image-level labels are oftentimes a by-product of clinical routine and are, thus, substantially easier to acquire than pixel-level annotation. In this study, image-level labels are the *N*-stage and a list of pathological LNS per patient according to image-based diagnosis, the location of which is not annotated in the images.

Table 1 Rule-set that derives the *N*-stage from (a) the pathological LNS and (b) the tumor side.[3] For each LNS group, at least one of the listed LNS needs to be pathological for the group to be counted as positive.

Tumor side	*N*-stage/ LNS group	Positive LNS	Negative LNS	LNS with unknown status
R	1	10R, 11R	All other	—
L	1	10L, 11L	All other	—
R	2	5, 6, 7, 8, 9	1L, 2L, 4L, 10L, 11L	10L, 11L
L	2	1L, 2L, 3A, 3P, 4L, 5, 6, 7, 8, 9	1R, 2R, 4R, 10R, 11R	10R, 11R
R	3	1L, 2L, 4L, 10L, 11L	—	All other
L	3	1R, 2R, 4R, 10R, 11R	—	All other

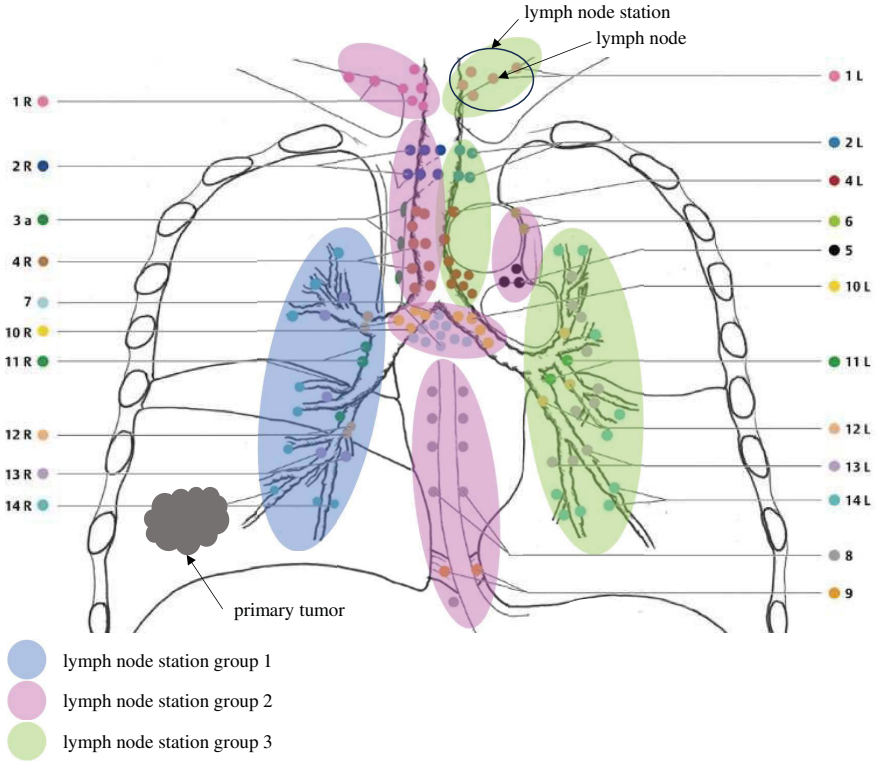


Fig. 1 Definition of lymph nodes, lymph node stations, and lymph node station groups given that the primary tumor is on the right side of the lung.

To the best of our knowledge, there exist no fully automated pipelines for *N*-staging of lung cancer patients on PET/CT, and the majority of existing approaches for subtasks such as binary lymph node classification require pixel-level annotations or evaluate the malignancy status of lymph nodes based solely on lymph node size. Our main contribution can be described as follows: we introduce a deep learning-based pipeline for binary classification of the mediastinal LNS status (malignant, benign) and *N*-stage prediction based on PET/CT trained solely on image-level labels. From a methodological perspective, we aim to mimic the radiologists' approach for *N*-staging and in this way guide the learning process by utilizing prior knowledge in two ways: (1) detecting the lymph node station locations through atlas-to-patient registration enables local patch-sampling and (2) transferring patient-level knowledge to the LNS-level permits weak supervision. The latter uses the patient's ground-truth *N*-stage (patient-level) to infer the pathology status of the corresponding LNS (LNS-level). We revise the problem of *N*-staging, which has taken a backseat in current research, by comparing multiple fully automated pipelines trained with and without segmentation masks. The aim of our comparison is to answer the question: To what extent can mediastinal lymph node assessment on PET/CT images be automated without the need for pixel-level annotations, given the currently available datasets and tools?

2 Related Work

The computer-aided assessment of mediastinal lymph nodes including classification, detection, and segmentation has a long and extensive history of research. Primarily, binary lymph node segmentation on CT has gained attention recently.[4–6] Atlas-based and learning-based segmentation of lymph nodes on CT including station mapping has been researched by Refs. 7–9. However, several studies show that mediastinal lymph node assessment by size alone may not be reliable[10,11] and the PET scan adds valuable information about the malignancy status of lymph nodes.[12] Considering the superiority of image-based assessment on PET/CT against CT only, many studies have focused on the binary classification of mediastinal lymph nodes into

malignant and benign using PET/CT images. References 13–18 use diagnostic features extracted based on manual segmentations of the lymph nodes and primary tumor, e.g., standardized uptake value, size, tumor shape, lymphatic drainage route. Taralli et al.[13] and Yoo et al.[14] add nonclinical patient information, e.g., age, sex, smoking status, as features for model training. The authors used or compared multiple algorithms such as decision trees, logistic regression, SVMs, or small multilayer perceptrons. The balanced accuracy for those studies ranges from 0.70 to 0.86. For *N*-staging, the balanced accuracy is 0.77[18] and 0.99.[17] In the studies that aim for *N*-staging, the diagnostic features are fused on LNS group-level instead of LNS-level. Consequently, the localization of lymph nodes, LNS mapping, and merging information into LNS groups has to be done manually to generate input features. As an alternative or addition to using predefined diagnostic features, Refs. 14, 16, 19, 20 apply tools to extract radiomics features based on segmentation masks of the lymph nodes. However, the diagnostic features are still necessary to build the best-performing model.

In the category of deep learning-based methods, Wang et al.[16] train a 2.5D convolutional neural network (CNN) on extracted image patches centered at the lymph node center. The authors report that the balanced accuracy of 0.86 for the CNN model is not significantly different from the results of traditional machine learning models tested in their study. All studies presented above require manual segmentation or manual ROI annotation during training and testing. To address this issue, Wallis et al.[21] propose a multistep approach for lymph node assessment without the need for annotations during inference. First, they trained a 2D U-Net to generate lymph node candidates and then extracted volumes that were classified into benign or malignant in the second step. For the generation of candidates during training, the authors used pixel-level annotations, i.e., 15 mm spheres centered at the lymph node center, to extract the ROIs. The authors report a sensitivity of 0.87 and 0.41 false positives per patient. A summary of the literature can be reviewed in Table 8 in Appendix A. However, the comparability between the related work and the presented methods in this paper is limited due to the fact that the existing approaches only cover a subtask of *N*-staging and require high manual annotation efforts for input data generation.

3 Methods

Similar to the process of radiologists, the task of *N*-staging was split into two subtasks—identifying pathological LNS, followed by the prediction of the *N*-stage. For the first step, the binary classification of the LNS, we propose to use prior knowledge about the approximate LNS locations by registering publicly available atlases with annotated LNS[22] to our data. The resulting segmentation masks of the LNS are used for targeted input patch-sampling. Precisely, the segmentation masks from registration of multiple atlases were used to generate a weight map (see Sec. 3.2.2 for detailed description) to further extract equally sized patches according to the weight map probabilities. Weighted patch-sampling, instead of cropping a bounding box around the LNS segmentation mask itself or its center, has the advantage of accounting for the uncertainty from atlas registration and allows for maintaining the information about surrounding anatomy, potentially involving the primary tumor. The proposed steps to generate the LNS input patches can be reviewed in Fig. 2. After preprocessing as described in Sec. 3.2, we trained a six-layer CNN to output a probability for malignancy given a patch approximately centered on any mediastinal LNS. In each convolutional block, the feature maps were normalized using instance normalization and spatially reduced by max-pooling. A ReLU was used as an activation function, and we added dropout layers to prevent overfitting. The station index was encoded and concatenated as additional information to be processed by the final classification layers. An overview of the proposed network structure is displayed in Fig. 3.

For the second step, we compared three training strategies, which mainly differ in the way the subtask of *N*-staging is solved. The options were as follows:

1. Training an AI-based LNS classifier followed by rule-based *N*-staging (TS 1).
2. Training both the LNS classifier and *N*-staging simultaneously (TS 2).
3. Freezing the pretrained LNS classifier and refining the last network layers for *N*-staging (TS 3).

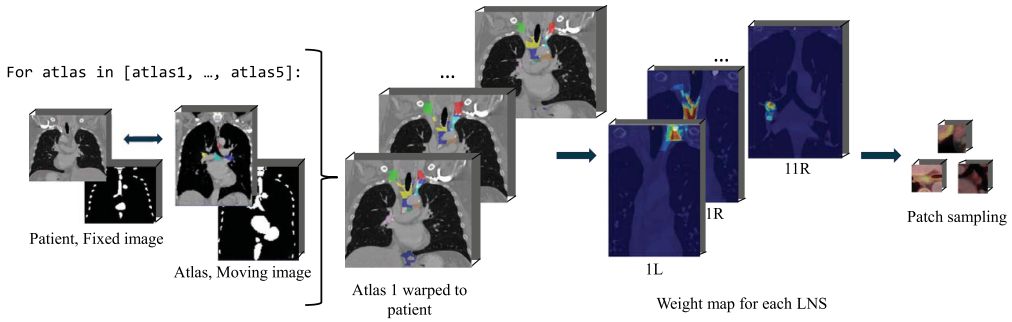


Fig. 2 Proposed approach for the generation of input patches for LNS classification. We used the segmentation masks of a selected range of anatomical structures from the TotalSegmentator algorithm to carry out multi-atlas registration of the mediastinal LNS to the training data. Based on the resulting LNS segmentations, a weight map is calculated for each LNS. Given the probabilities of the weight map, patches were extracted on LNS-level.

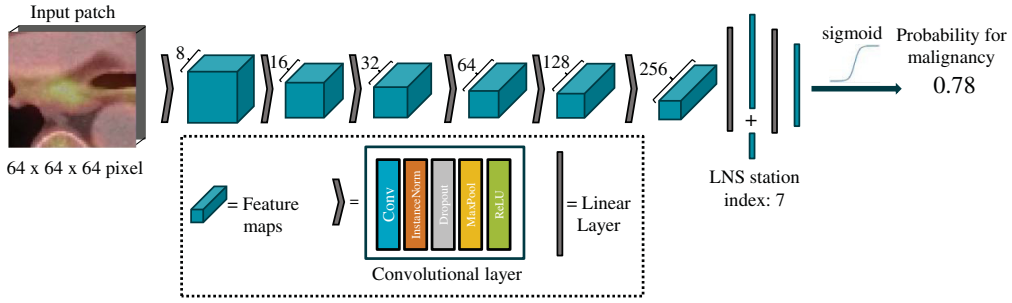


Fig. 3 Proposed six-layer CNN for LNS classification.

For TS 1, we trained the CNN-based LNS classifier as described above. By iteratively applying the network to all the LNS of a patient, we get a list of pathological LNS per patient. The rules described in Table 1 were applied to determine the *N*-stage. Alternatively, we investigated training an end-to-end network (TS 2), which further processes the logits from the LNS classification and the encoded primary tumor side in two linear layers to lastly output the *N*-stage. Here, patches from all the LNS of one patient are processed by *L* LNS classifiers, whereas LNS classifier weights were shared. The weights from TS 1 of the LNS classifier were used for transfer learning in TS 2 and TS 3. The advantage of using an end-to-end algorithm for *N*-staging (TS 2 and 3) is that more training data becomes available, i.e., cases where the *N*-stage is known, but the list of pathological LNS is unavailable. Leveraging the fact that *N*-staging errors can be backpropagated, we added pseudo-labels into the loss calculation by transferring implicit knowledge from patient- to LNS-level (see Sec. 3.3.1 for more details). More precisely, pseudo-labels were generated exclusively from the known *N*-stage in combination with medical prior knowledge describing the relationship to the malignancy status of LNS. We consider the introduction of pseudo-labeling and the architectural design a form of weak supervision, as prior knowledge about *N*-staging is used to guide network training and divergence from causal mechanisms is penalized. TS 3 is architecturally the same as TS 2, but the weights of the LNS classifiers remain frozen. Only the last network layers are finetuned for *N*-staging. The overviews of the proposed workflows for TS 1, TS 2, and TS 3 are presented in Figs. 4–6, respectively.

3.1 Data

Two internal datasets were retrospectively collected at the University Medical Center Schleswig-Holstein (UKSH). UKSH-1 comprises data from 353 patients, whereas UKSH-2 includes 42 patients, all treated as part of routine clinical care. In addition to imaging data, nonclinical and diagnostic information was extracted from the clinical information system. Preoperative imaging–based diagnoses and postoperative histopathological diagnoses were

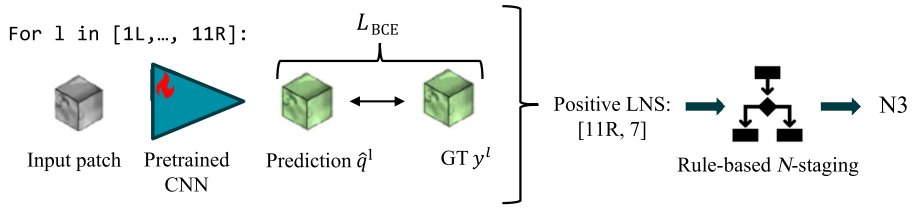


Fig. 4 Overview of training strategy 1. Using the LNS classifier to predict the malignancy status of all LNS of the patient, a list of pathological LNS is generated, to which the rule-based *N*-staging can be applied.

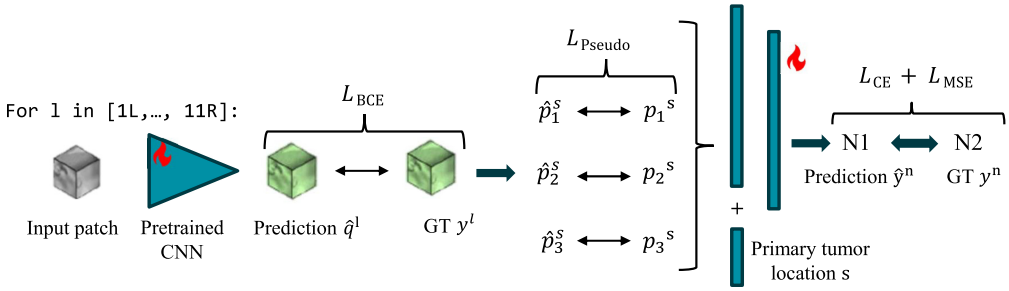


Fig. 5 Overview of training strategy 2. The CNN-based LNS classifier predicts the malignancy status for all LNS of a patient. The *N*-stage is determined by training two linear layers, which process the logits of LNS classification and the primary tumor side. The proposed workflow is trained end-to-end.

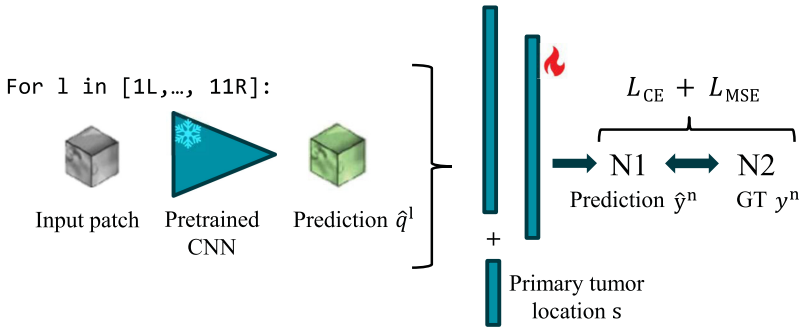


Fig. 6 Overview of training strategy 3. For all LNS, a pretrained CNN is used to predict the probability of malignancy. The logits of LNS classification concatenated with the primary tumor side are used to train two linear layers for *N*-staging.

obtained from radiology and pathology reports, respectively. In addition to the UKSH datasets, the LungPETCTDx dataset[23] was used for training TS 2 and 3. LungPETCTDx contains PET/CT images with available *N*-stage and segmentation masks of the primary tumor. However, it lacks LNS-level information and therefore cannot be used for TS 1 training. Dataset characteristics are summarized in Table 2. The distribution of pathological LNS based on image-derived diagnoses is highly imbalanced, with stations 4L/4R, 7, and 11L/11R occurring more frequently than other stations (Fig. 7).

3.2 Preprocessing

PET voxel intensities were converted to standardized uptake values (SUV) following Gatidis et al.[24] All images were resampled to an isotropic spacing of $1 \times 1 \times 1$ mm^3. CT intensities were scaled to the range [0,1] using the soft tissue window (−200 to 400 Hounsfield units), whereas SUV values in the range between 0 and 4 were linearly scaled to [0,1] without clipping.

Table 2 Dataset characteristics. Images are cropped to the mediastinum. AC, adenocarcinoma; SCC, squamous cell carcinoma. pN is the pathological *N*-stage based on histological examination. cN is the clinical *N*-stage based on the manual assessment of preoperative PET/CT scans. The dash means that this information has not been evaluated.

Characteristics	UKSH-1	UKSH-2	LungPETCTD[23]
Number of patients	353	42	102
Sex (male/female)	214/139	—	58/44
Age (min/max/mean)	43/103/72	—	39/90/62
Pathological LNS status (malignant/benign)	231/5770	—	—
Clinical LNS status (malignant/benign)	347/5552	340/374	—
pN (0/1/2/3)	223/45/82/3	—	—
cN (0/1/2/3)	196/41/81/35	0/0/9/33	64/25/4/9
Cancer type (AC/SCC/other)	214/105/34	—	—
Number of scans	511	78	175
Size in *X*-axis (mean ± std)	180 ± 15	145 ± 33	200 ± 21
Size in *Y*-axis (mean ± std)	163 ± 20	127 ± 32	185 ± 25
Slice number (mean ± std)	163 ± 20	127 ± 32	295 ± 84
In-plane resolution in mm (mean ± std)	0.98 ± 0.02	1.25 ± 0.27	0.80 ± 0.06
Slice spacing in mm (mean ± std)	2.92 ± 0.49	2.99 ± 0.51	0.88 ± 0.49

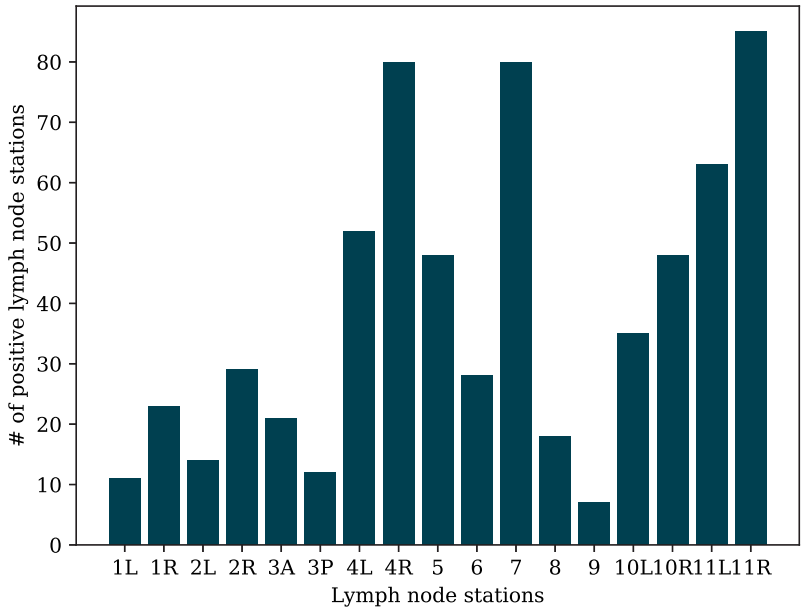


Fig. 7 Distribution of positive LNS in dataset UKSH-1 according to the preoperative image-based diagnosis of physicians.

3.2.1 *Atlas registration for lymph node station detection*

The registration pipeline comprised sequential rigid, affine, and deformable registration, with the latter implemented using a GPU-based version of ITK's VariationalRegistration module.[25] To improve robustness, rigid and affine registrations were guided by segmentation masks of thoracic bones and other anatomical structures, which are less likely to be affected by tumor-induced or

other pathologies, generated with TotalSegmentator.[26] Registration parameters were optimized by maximizing the Dice score between the fixed-image segmentation masks and the warped masks of the moving image for the selected anatomical structures. The registration algorithm is available at https://github.com/MICAI-IMI-UzL/LNQ2023. To account for the anatomical differences of the atlas patients and the provided segmentation masks, multi-atlas segmentation was performed using five atlases published by Lynch et al.,[22] which provide mediastinal LNS segmentations according to the international lymph node map of the seventh TNM edition.[27] All atlases were registered to the training images, and the warped labels were fused via majority voting.

3.2.2 *Patch sampling strategy*

The patch sampling was executed based on an LNS-dependent weight map derived from the LNS segmentation masks from the multi-atlas registration. The weight map $W^l = (w^l_{ijk})$ at pixel index $(i, j, k) \in \Omega$, where Ω is the image space, is calculated for each LNS $l \in L$ as follows:

$$w^l_{ijk} = \begin{cases} 1 - d^l_{ijk} & m^l_{ijk} = 1 \\ 0 & m^l_{ijk} = 0 \end{cases} \tag{1}$$

where $M^l = (m^l_{ijk})$, $(i, j, k) \in \Omega$ is the binary mask of the segmented LNS l and $D^l = (d^l_{ijk})$, $(i, j, k) \in \Omega$ is a Euclidean distance map calculated from the centroid of the LNS mask l to all other points in the image. The weight map prioritizes patch sampling at the centroid of the LNS indicated by atlas-based LNS localization while probabilistically including surrounding regions with decreasing likelihood. The patch size was $64 \times 64 \times 64$ pixels. During inference, patches were extracted at the LNS centroid corresponding to the maximum weight map value.

3.3 Model Training

The biggest challenges of this use case are the heterogeneity of the image data and the severe class imbalance. The CT images have varying technical parameters (e.g., resolution, field of view, reconstruction kernel). The anatomy of the mediastinum in combination with pathologies is strongly individual for each patient. In addition, the appearance of the different lymph nodes and LNS differs significantly in size, shape, and location. To account for the heterogeneity of the CT images, we used random affine and elastic deformation for augmentation.

Multiple strategies were used to address the problem of class imbalance: in the train-validation-test-split, the patients were stratified based on the combination of the N-stage and the sum of pathological LNS per patient. During model training, the positive class in binary classification or minority classes in N-staging were oversampled to balance the class distribution. For evaluation, we rely on metrics such as balanced accuracy because traditional classification accuracy might not identify poor performance for minority classes.

Model training was performed using PyTorch 2.2.1, a NVIDIA A100-SXM4-40GB GPU. The Adam optimizer (initial learning rate = 5×10^{-4}, weight decay = 0.001) was employed, with adaptive learning rate scheduling. The batch size was set to 64 for TS 1 and 16 for TS 2 and 3. Training was conducted for a maximum of 150 epochs, with early stopping applied if the monitored validation metric (balanced accuracy at LNS-level) did not improve within 30 epochs.

3.3.1 *Pseudo-labeling & loss calculation*

To mitigate the absence of pixel-level annotations, patient-level N-stage labels and medical prior knowledge were used to derive pseudo-labels for LNS groups and incorporated into the loss to enforce consistency with N-staging rules. Certain N-stages mandate or exclude pathological involvement of specific LNS groups (Tables 1 and 3). LNS groups with unknown pseudo-labels were excluded from the pseudoloss calculation.

The loss function for the end-to-end training of TS 2 is a combination of cross-entropy (CE) loss and mean squared error (MSE) loss for N-staging, as well as a binary cross-entropy loss (BCE) and a pseudoloss for LNS classification

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{CE}}(y^n, \hat{y}^n) + \lambda_2 \mathcal{L}_{\mathrm{MSE}}(y^n, \hat{y}^n) + \lambda_3 \mathcal{L}_{\mathrm{Pseudo}}(\boldsymbol{p}^s, \hat{\boldsymbol{p}}^s) + \lambda_4 \sum_{l=1}^{L} \mathcal{L}_{\mathrm{BCE}}(y^l, \hat{q}^l), \tag{2}$$

Table 3 Translation of the rule-set underlying *N*-staging described in Table 1 into pseudo-labels. The pseudo-labels are determined column-wise: There is one pseudo-label per LNS group, that is, three pseudo-labels per patient. The pseudo-labels marked with **?** are unknown and cannot be derived solely based on the ground-truth *N*-stage. Example: If the *N*-stage is one, all LNS in LNS groups 2 and 3 must be benign, whereas at least one of the LNS in LNS group 1 needs to be pathological.

LNS group	LNS for tumor side = R	LNS for tumor side = L	*N*1	*N*2	*N*3
1	10R, 11R	10L, 11L	1	?	?
2	1R, 2R, 3A, 3P, 4R, 5, 6, 7, 8, 9	1L, 2L, 3A, 3P, 4L, 5, 6, 7, 8, 9	0	1	?
3	1L, 2L, 4L, 10L, 11L	1R, 2R, 4R, 10R, 11R	0	0	1

where y^n and $\hat{y}^n$ are the ground truth and predicted N-stage, respectively. Given the ordinal nature of N-staging, the combination of the CE with the MSE loss penalizes larger deviations from the ground truth N-stage more heavily, similar to Nielsen et al.'s approach for therapy response assessment in cerebral ischemic stroke patients.[28] Given the N-stage and primary tumor side s, $\boldsymbol{p}^s \in \mathbb{R}^3$ with $p_i^s \in \{0,1\}$ denotes a vector of pseudo-labels (corresponding to the last three columns in Table 3), whereas $\hat{\boldsymbol{p}}^s \in \mathbb{R}^3$ with $\hat{p}_i^s \in [0,1]$ represents the predicted probabilities for the LNS groups. For the LNS group i, $\hat{p}_i^s$ is the maximum probability of malignancy of all classification results of this group, e.g., $\hat{p}_i^s = \max_j \hat{q}_j^s$, where $\hat{q}$ is a classification probability predicted by all LNS classifiers indexed by j that belong to group i. The pseudoloss maximizes or minimizes the class probabilities for LNS groups with or without positive LNS, respectively

$$\mathcal{L}_{\text{Pseudo}}(\boldsymbol{p}^s, \hat{\boldsymbol{p}}^s) = \sum_{i=1}^{3} |p_i^s - \hat{p}_i^s|. \tag{3}$$

To extend the calculation to batches, we further introduced a weighting factor that is inversely proportional to the frequency of pseudo-label entries. The $\mathcal{L}_{\text{BCE}}$ is responsible for maximizing agreement between the predicted malignancy status of LNS l, denoted $\hat{q}^l$, and the ground truth malignancy status y^l. $\mathcal{L}_{\text{BCE}}$ is calculated over all L LNS of one patient, if the ground truth on LNS-level y^l is available, i.e., this term is not calculated for the patients of LungPETCTDx. The loss weights were chosen to be $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.02$, and $\lambda_4 = 1$, so that all loss terms have approximately the same value range at the beginning of training.

4 Results

In the following, we present results on the performance of the atlas-based LNS mapping and quantitative results for the proposed and comparison methods for LNS classification and N-staging.

4.1 Quality Assessment of Atlas-Based LNS Mapping

One of the key factors determining the upcoming results is the quality of the atlas-based LNS mapping. To evaluate the quality of the atlas-to-patient registration, we assessed the overlap of anatomical structures of the mediastinum (trachea, esophagus, pulmonary artery, and several parts of the heart) segmented by the TotalSegmentator algorithm[26] and the warped segmentation masks from atlas registration. The average Dice score for this eight-class segmentation task ranges from 0.85 ± 0.04 to 0.89 ± 0.05 depending on the atlas patient used as the moving image. However, the warped LNS segmentation masks vary depending on which atlas patient was used as the moving image (see Fig. 11 in Appendix C).

To further evaluate the localization performance of LNS mapping, we assessed the LNS mapping accuracy using the datasets provided by Refs. 29 and 30. These datasets contain mediastinal lymph node segmentations of 104 patients and also include labels for the LNS per lymph node. Thus, for each segmented lymph node, we assign the LNS with maximal weight map probability, whereas the weight map is generated as defined in Eq. (1). Similar to Hoffman et al.,[31]

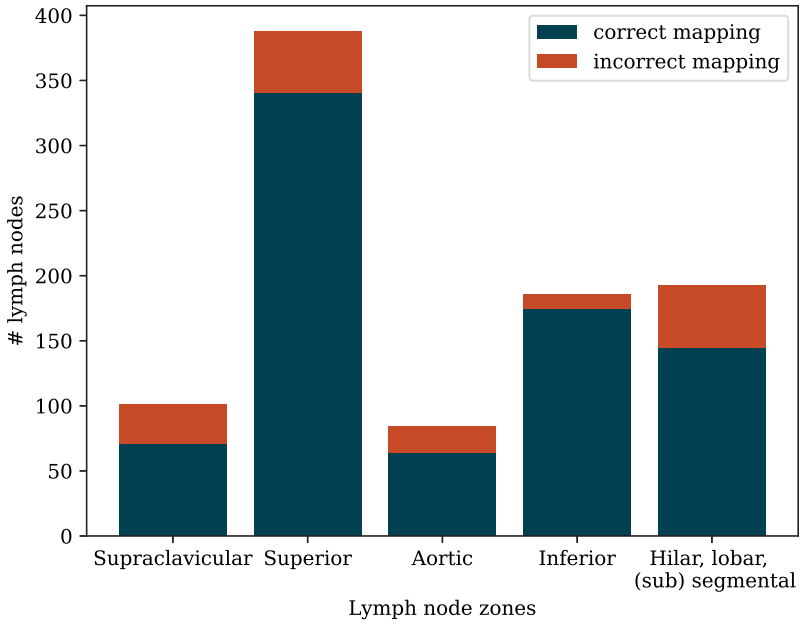


Fig. 8 Quality assessment of the atlas-based LNS mapping per lymph node zones. Correctly mapped lymph nodes are marked in blue, whereas the lymph nodes that were not correctly mapped to its ground-truth lymph node zone are marked in orange.

we report the performance for LNS mapping per lymph node zone in Fig. 8. The overall accuracy for atlas-based lymph node zone mapping is 84% averaged over all lymph nodes.

Furthermore, we conducted a sensitivity analysis to assess the impact of the atlas-based LNS mapping on the prediction performance of our proposed method. The experiment was designed as follows: we sampled three out of the five available atlases[22] for multi-atlas registration to create the weight map for patch sampling described in Sec. 3.2.2. Then, the LNS classifier trained for TS 1 was used to predict the malignancy status for all LNS of the test patients based on the patch extracted at the highest value of the weight map, that is the LNS centroid. For all possible multi-atlas combinations used for weight map generation, the performance differences were insignificant according to a McNemar's test (p value >0.05).

4.2 Baseline Methods

The gold standard for detecting pathological lymph nodes in clinical practice based on PET images is to identify the maximum SUV in a manually determined volume of interest. If the so-called SUV_{max} exceeds a predefined threshold t, the LNS is set to be pathological. Even though using the SUV instead of the PET values directly already introduces standardization across patients, each patient's physiological uptake is different and, therefore, the choice of the threshold can strongly influence the accuracy of detecting pathological lesions. For mediastinal lymph node segmentation, physicians usually use a threshold of 2.5;[32,33] however, Hellwig et al.[34] suggested that in their study the highest diagnostic accuracy was achieved with a threshold of 4.5. For our thresholding algorithm, we replicate this threshold-based approach by calculating the maximum value of the SUV-image of each patient within the segmentation mask of each LNS generated by atlas registration. Then, the SUV_{max} values are stored in a list and compared with a threshold t. Using the same train-test-splits as the AI-based methods, we find the optimal threshold $t \in \{2.5, 3, 3.5, 4, 4.5, 5, 5.5\}$ by maximizing the balanced accuracy on the training data for each fold. The evaluation is then performed with the optimal t on the respective test data.

For the autoPET approach, we developed a pipeline based on the results of the autoPET challenge hosted at MICCAI, which aims to segment lesions on PET/CT scans.[35] The winner algorithm of 2022 published the code for inference (https://github.com/Yejin0111/autoPET2022_Blackbean), which we used to process the PET/CT scans of our datasets.[36] The segmented lesions were matched with the LNS segmentations to generate a list of pathological LNS per patient. For each target lesion, the LNS with the highest overlap above a threshold $o \in \{0.05, 0.1, 0.2\}$

Table 4 Quantitative results of lymph node assessment on PET/CT to predict cN for all datasets. TS, training strategy. Best results are marked in bold. TS 1 of the proposed algorithm is significantly different from both baseline methods according to McNemar's test for LNS classification (p value <0.05). For N-staging, TS 1 of the proposed algorithm is significantly different from the thresholding algorithm according to a permutation test.

		LNS classification				N-staging		
TS	Algorithm	Acc.	Bal. Acc.	Sens	Spec.	Acc.	Sens.	Spec.
1	AI-based LNS classifier and rule-based N-staging	0.88 ± 0.02	0.81 ± 0.03	0.72 ± 0.08	0.90 ± 0.03	0.63 ± 0.04	0.58 ± 0.04	0.87 ± 0.01
2	AI-based LNS classifier and N-staging	**0.91** ± 0.00	0.73 ± 0.02	0.49 ± 0.05	0.96 ± 0.00	**0.64** ± 0.06	0.52 ± 0.05	0.86 ± 0.02
3	Freeze pretrained LNS classifier and refine N-staging	0.88 ± 0.02	**0.81** ± 0.03	0.71 ± 0.08	0.90 ± 0.02	0.62 ± 0.03	0.55 ± 0.04	0.86 ± 0.01
	Thresholding ($t = 4$)	0.83 ± 0.00	0.81 ± 0.01	0.77 ± 0.03	0.84 ± 0.01	0.50 ± 0.01	0.51 ± 0.02	0.84 ± 0.00
	autoPET algorithm ($o = 0.05$)[a,35]	0.89 ± 0.00	0.62 ± 0.01	0.26 ± 0.02	0.97 ± 0.00	0.59 ± 0.01	0.51 ± 0.02	0.86 ± 0.00

[a]Supervised training using lesion segmentation masks.

is considered pathological. For N-staging, the predefined rule-set was applied. A schematic overview of the pipeline can be seen in Fig. 10 in Appendix B. The overlap threshold o was considered a hyperparameter, and the best-performing configuration is reported in the following.

4.3 Quantitative Results

In this section, we present quantitative results for several training strategies and compare the results to the autoPET algorithm as well as the thresholding approach. We report five-fold cross-validated results in Table 4. Hence, the evaluation for N-staging is carried out for all available datasets, consisting of 497 patients. On the LNS-level, we report results for the UKSH-datasets only as there are no labels available on the LNS-level for the LungPETCTDx dataset.

The first training strategy, that is using a binary classifier to predict the pathological LNS and applies rules for N-staging, yields an accuracy and balanced accuracy for LNS classification of 0.88 and 0.81, respectively, whereas the accuracy for N-staging is 0.63. The third training strategy performs similarly well. The second training strategy leads to a decreasing sensitivity in LNS classification, but the accuracy for N-staging is the highest of all methods. The thresholding approach is oversensitive as can be seen in the low specificity value for LNS classification and the confusion matrices in Fig. 9. The high sensitivity on the LNS-level affects the N-staging negatively as the false positive LNS lead to high N-stages. The autoPET algorithm shows a very low sensitivity, but a high specificity on LNS-level and comparable performance for N-staging as our proposed algorithm. Although all algorithms tend to overestimate the N-stage [see Fig. 9(b)], the first training strategy of our proposed algorithm shows the best balance between the reported evaluation metrics and smaller deviations across the ordinal scale of the N-stage. For TS1, we do not observe considerable differences in the false negative or false positive rates across different LNS and see no correlations to the LNS that were incorrectly mapped more frequently through atlas-based LNS mapping. For LNS classification, our proposed algorithm significantly outperforms the baseline methods despite requiring no pixel-level annotations for training.

4.3.1 *Ablation study*

We report five-fold cross-validated results for using the weighted patch sampling versus using a patch of fixed size around the centroid of the segmented LNS. Similarly, we compare the simultaneous training of the LNS classifier and N-staging with and without the pseudo-labeling loss.

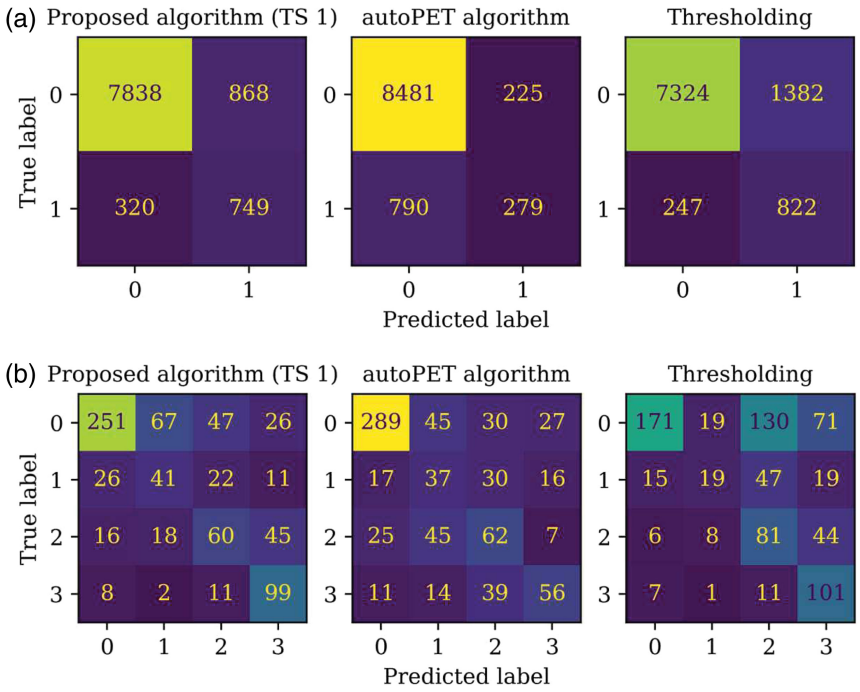


Fig. 9 Confusion matrices for LNS classification and *N*-staging for performance comparison.

Table 5 Ablation study for weighted patch sampling for training strategy 1 for the UKSH dataset. Best results are marked in bold. The differences between the models are statistically significant according to McNemar's test for LNS classification (p value <0.05).

Patch sampling	LNS classification				*N*-staging		
	Acc.	Bal. Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
No	0.90 ± 0.00	0.77 ± 0.04	0.60 ± 0.07	0.93 ± 0.01	0.60 ± 0.06	0.56 ± 0.06	0.86 ± 0.02
Yes	0.88 ± 0.02	0.81 ± 0.03	0.72 ± 0.08	0.90 ± 0.03	**0.63** ± 0.04	**0.59** ± 0.05	**0.88** ± 0.01

Table 6 Ablation study for pseudo-labeling loss for training strategy 2. Best results are marked in bold. The differences between the models are statistically significant according to a permutation test for *N*-staging (p value <0.05).

Pseudo-labeling	LNS classification				*N*-staging		
	Acc.	Bal. Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
No	0.91 ± 0.00	0.72 ± 0.05	0.49 ± 0.11	0.96 ± 0.01	0.59 ± 0.02	0.50 ± 0.04	**0.86** ± 0.01
Yes	0.91 ± 0.00	**0.73** ± 0.02	**0.49** ± 0.05	**0.96** ± 0.00	0.64 ± 0.06	0.52 ± 0.05	0.86 ± 0.02

The results in Tables 5 and 6 show that using the patch sampling improves LNS classification and pseudo-labeling improves *N*-staging significantly.

4.3.2 *Analysis of computational efficiency*

Table 7 shows the required inference run time for the individual processing steps for all algorithms. All algorithms have a similar processing time of ∼13 min per PET/CT.

Table 7 Computational efficiency analysis for the processing of one PET/CT for inference.

Proposed algorithm (TS 1)		autoPET algorithm		Thresholding	
Processing step	AVG time (s)	Processing step	AVG time (s)	Processing step	AVG time (s)
TotalSegmentator for 5 atlases	105	autoPET segmentation	868	TotalSegmentator for 5 atlases	105
Multi-atlas registration	645	Atlas-based LNS mapping + rule-based *N*-staging	8	Multi-atlas registration	645
Patch sampling	53	—	—	Thresholding + rule-based *N*-staging	0.19
Inference + rule-based *N*-staging	0.09	—	—	—	—
Total	803	Total	876	Total	750

5 Discussion

There exists a lot of research on subtasks of *N*-staging, i.e., binary lymph node classification, on PET/CT given manually extracted features or image regions, but labor-intensive manual interaction is not feasible in practice. Consequently, the main focus of this study is to introduce a deep-learning–based solution for both LNS classification in the mediastinum and *N*-staging on PET/CT, which was trained using image-level labels only. Our proposed algorithm performs significantly better than the standard thresholding-based approach as well as the autoPET algorithm, which was trained with segmentation masks, for LNS classification. Our study shows that the division into subtasks and the integration of prior knowledge are paramount steps in the reduction of labeling effort on pixel-level for lymph node analysis, which is in alignment with current literature.[7,37]

LNS mapping is affected by high interrater variability and anatomical heterogeneity across patients.[22,38] Our analyses show that the variances in atlas annotation have no significant impact on the classification accuracy of our proposed algorithm, but it remains unclear to what extent label fusion through majority voting accounts for the localization uncertainty in the supervised comparison method. For instance, the LNS mapping based on the autoPET segmentations could be improved by a probabilistic approach in future work. The supervised comparison algorithm was trained on publicly available data. Although retraining with larger annotated datasets might enable supervised methods to outperform the proposed, such data are currently unavailable. Another aspect that could lead to the forward propagation of errors is the fact that one wrongly predicted LNS can change the prediction on the patient-level from *N*0 to *N*3 or the other way around. Oversensitive LNS classification can impair *N*-staging accuracy, as observed in the baseline methods, which do not account for the ordinal nature of *N*-staging. As stated in Sec. 4, AI-based *N*-staging in training strategy 2 shows a better classification accuracy than the rule-based *N*-staging in training strategy 1, whereas the performance on LNS-level deteriorates. This indicates that the AI-based *N*-staging does not solely rely on the causal rule-set. The higher amount of training data and the pseudo-labeling do not suffice to ensure that the *N*-staging rule set is considered by the algorithm. Segmentation-based methods based on the whole PET/CT such as the autoPET algorithm provide a layer of interpretability and verifiability, that is lost, when the network is trained based on patches. Furthermore, AI-based *N*-staging in TS 2 and 3 compromises transparency regarding, which LNS is responsible for changes in the *N*-stage. However, the predicted list of pathological LNS in TS 1 and the LNS segmentation masks from atlas-based registration can serve as auxiliary tools for manual verification of the predicted *N*-stage, when lymph node segmentation masks are not available.

For lymph node assessment based on images, multiple factors can have a high predictive power, e.g., lymph node size, absolute tracer uptake, relative tracer uptake to the liver, tumor size, tumor tracer uptake, and lymphatic routes. A large patch size and the weight map–based patch

sampling should enable learning features from the lymph node surrounding anatomy and primary tumor location, but this is not guaranteed. Previously published literature shows that adding information about the primary tumor resulted in a significant performance gain.[20] Methodologically, future research could focus on adding information about the primary tumor into the proposed methodology.

We also carried out experiments, in which we used the histologically proven labels as the ground-truth and predicted the pN-stage by training only on UKSH-1. However, we observe that the sensitivity for LNS classification drops below 0.5, which carries through to an *N*-staging accuracy of 0.45. We assume that the performance drop results from the severe class imbalance, as the collected dataset mostly contains patients with no, little, or mediocre metastatic spread to the lymph nodes. To obtain a dataset that shows a representative prevalence, it would be necessary to also surgically resect the lymph nodes of patients with severe metastatic spread and then do a histological examination. However, for those patients, surgery is not an appropriate treatment option; therefore, it would be unethical to carry out a lymph node resection. The acquisition of a more balanced dataset and including a diverse range of positive lymph nodes and LNS is a crucial step for future research. Multicentric data acquisition will be necessary to collect sufficient data from patients with a large pN-stage, who were still treated surgically. However, the challenge of class imbalance across the LNS will remain due to ethical constraints in patient management and the nature of metastatic spread. With this study, we make an important step toward the automation of mediastinal lymph node assessment on PET/CT images and provide insight into the influencing factors.

6 Appendix A: Literature Review for Lymph Node Assessment on PET/CT

Table 8 shows the summary of a literature review for lymph node classification and *N*-staging on PET/CT images. Balanced accuracy ranges from 0.70 to 0.86 for lymph node classification and reaches 0.99 and 0.77 for *N*-staging.

Table 8 Literature review for lymph node classification (LN class) or prediction of the *N*-stage based on PET/CT.

Author	Year	# images	# LNs	Features	Task	Sensitivity	Specificity	Bal. Acc.	AUC	ACC
Dong et al.[19]	2024	320	509	Radiomics, image features	LN class	0.85	0.75	0.80	0.85	0.80
Laros et al.[20]	2022	148	504	Radiomics	LN class	0.80	0.90	0.85	—	0.88
Wallis et al.[21]	2021	184	241	Image features	LN class	0.88	—	—	—	
Taralli et al.[13]	2021	540	4156	Demographic information, diagnostic features	LN class	0.58	0.81	0.70	0.77	0.77
Yoo et al.[14]	2020	980	1329	Demographic information, diagnostic features, and texture features	LN class	0.86	0.75	0.80	0.85	0.80
Cheng et al.[15]	2019	45	222	Diagnostic features	LN class	0.82	0.85	0.84	—	0.85
Wang et al.[16]	2017	168	1397	Image features, diagnostic features, texture features	LN class	0.84	0.88	0.86	0.91	0.86
Toney et al.[17]	2013	133	—	Diagnostic features	*N*-staging	0.996	0.997	0.997	—	0.99
Vesselle et al.[18]	2003	133	—	Diagnostic features	*N*-staging	0.65	0.88	0.77	—	0.87

7 Appendix B: Model Overview of the Inference of the autoPET Algorithm

An overview of the autoPET inference process is shown in Fig. 10. PET-lesion segmentation using an established segmentation model is followed by overlap assessment for each LNS and rule-based N-staging.

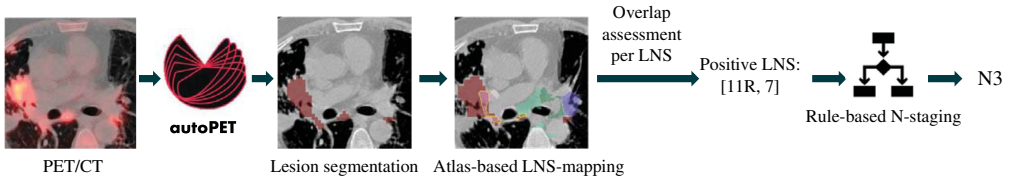
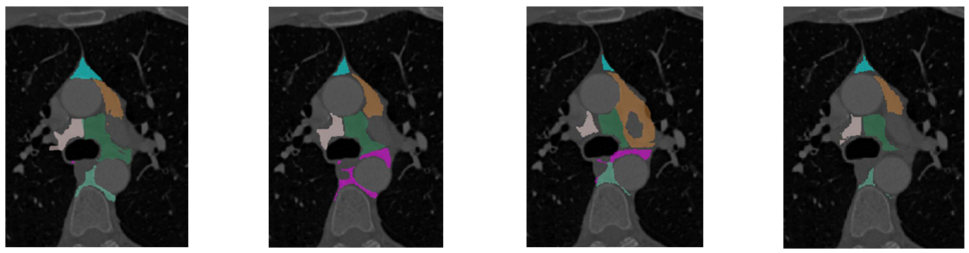


Fig. 10 Overview of the inference of the autoPET algorithm. First, the PET/CT image is processed via the autoPET challenge winners algorithm of 2022 to generate segmentation masks for PET lesions. Second, the LNS segmentation masks from atlas registration are matched with the PET lesion segmentation masks. Based on the overlap of the PET lesion and the LNS segmentation mask, the LNS is classified as positive. The resulting list of positive LNS determines the *N*-stage according to the ruleset in Table 1.

8 Appendix C: Results of Atlas Registration for Different Atlas Patients as Moving Image

Figures 11(a)–11(c) show LNS segmentation masks from atlas-to-patient registration for three different atlases used as the moving image. The resulting segmentation masks displayed in axial view do not show clear consensus for most LNS, which can be led back to the high interrater variability in annotation of the atlases.[22,38] Figure 11(d) shows the fusion of the LNS segmentation masks via majority voting.



(a) Atlas 1 as moving image. (b) Atlas 2 as moving image. (c) Atlas 5 as moving image. (d) Majority voting.

Fig. 11 Result of atlas registration using different atlas patients and their LNS segmentation masks as the moving image in axial view.

Disclosures

The authors declare that there are no financial interests, commercial affiliations, or other potential conflicts of interest that could have influenced the objectivity of this research or the writing of this paper.

Code and Data Availability

The dataset used in this study is derived from internal sources and contains sensitive information that is subject to privacy and legal restrictions. Due to these concerns, the data cannot be made publicly available. Code is available at https://github.com/MICAI-IMI-UzL/kimedlung.

Ethics Statement

This study was conducted in full compliance with ethical guidelines and principles for research involving human participants. Ethical approval was obtained from the Ethics Committee of the University of Lübeck under file number 20-227.

Acknowledgments

This work was supported by grants of the collaborative research project "AI ecosystem in health care" within the subproject titled "Health system-based analysis of disease patterns using lung diseases as an example" financed by the federate state Schleswig-Holstein, Germany.

References

1. F. C. Detterbeck et al., "Invasive mediastinal staging of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition)," *Chest* **132**, 202S–220S (2007).
2. C. F. Mountain and C. M. Dresler, "Regional lymph node classification for lung cancer staging," *Chest* **111**, 1718–1723 (1997).
3. M. B. Amin et al., "The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging," *CA Cancer J. Clin.* **67**(2), 93–99 (2017).
4. R. Dorent et al., "LNQ 2023 challenge: benchmark of weakly-supervised techniques for mediastinal lymph node quantification," *Mach. Learn. Biomed. Imaging* **3**, 1–15 (2025).
5. S. Engelson et al., "LNQ challenge 2023: learning mediastinal lymph node segmentation with a probabilistic lymph node atlas," *Mach. Learn. Biomed. Imaging* **2**, 817–833 (2024).
6. S. Engelson et al., "Comparison of anatomical priors for learning-based neural network guidance for mediastinal lymph node segmentation," *Proc. SPIE* **12927**, 1292719 (2024).
7. Y. Cao et al., "LNAS: a clinically applicable deep-learning system for mediastinal enlarged lymph nodes segmentation and station mapping without regard to the pathogenesis using unenhanced CT images," *Radiol. Med.* **129**, 229–238 (2024).
8. J. Liu et al., "Mediastinal lymph node detection on thoracic CT scans using spatial prior from multi-atlas label fusion," *Proc. SPIE* **9035**, 90350M (2014).
9. A.-I. Iuga et al., "Automated mapping and N-staging of thoracic lymph nodes in contrast-enhanced CT scans of the chest using a fully convolutional neural network," *Eur. J. Radiol.* **139**, 109718 (2021).
10. T. Arita et al., "Is it possible to differentiate malignant mediastinal nodes from benign nodes by size? Reevaluation by CT, transesophageal echocardiography, and nodal specimen," *Chest* **110**, 1004–1008 (1996).
11. K. L. Prenzel et al., "Lymph node size and metastatic infiltration in non-small cell lung cancer," *Chest* **123**, 463–467 (2003).
12. Y. Wu et al., "Diagnostic value of fluorine 18 fluorodeoxyglucose positron emission tomography/computed tomography for the detection of metastases in non-small-cell lung cancer patients," *Int. J. Cancer* **132**(2), E37–E47 (2013).
13. S. Taralli et al., "Application of artificial neural network to preoperative 18F-FDG PET/CT for predicting pathological nodal involvement in non-small-cell lung cancer patients," *Front. Med.* **8**, 664529 (2021).
14. J. Yoo et al., "Machine learning-based diagnostic method of pre-therapeutic 18F-FDG PET/CT for evaluating mediastinal lymph nodes in non-small cell lung cancer," *Eur. Radiol.* **31**, 4184–4194 (2021).
15. T. Cheng, N. Chiu, and Y. Fang, "Automatic classification of lymph node metastasis in non-small-cell lung cancer (NSCLC) patient on F-18-FDG PET/CT," in *Future Trends Biomed. Health Inform. Cybersecurity Med. Devices – Proc. Int. Conf. Biomed. Health Inform. (ICBHI 2019), IFMBE Proc.*, K.-P. Lin, R. Magjarevic, and P. de Carvalho, Eds., Springer Science and Business Media Deutschland GmbH, Germany, pp. 138–142 (2020).
16. H. Wang et al., "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images," *EJNMMI Res.* **7**, 11 (2017).
17. L. K. Toney and H. J. Vesselle, "Neural networks for nodal staging of non-small cell lung cancer with FDG PET and CT: importance of combining uptake values and sizes of nodes and primary tumor," *Radiology* **270**, 91–98 (2014).
18. H. Vesselle et al., "Application of a neural network to improve nodal staging accuracy with 18F-FDG PET in non-small cell lung cancer," *J. Nucl. Med.* **44**(12), 1918–1926 (2003).
19. S. Dong, A. Fu, and J. Liu, "Prediction of metastases in confusing mediastinal lymph nodes based on fluorine-18 fluorodeoxyglucose (18F-FDG) positron emission tomography/computed tomography (PET/CT) imaging using machine learning," *Quant. Imaging Med. Surg.* **14**, 4723–4734 (2024).
20. S. S. A. Laros et al., "Machine learning classification of mediastinal lymph node metastasis in NSCLC: a multicentre study in a Western European patient population," *EJNMMI Phys.* **9**, 66 (2022).

21. D. Wallis et al., "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* **49**, 881–888 (2022).
22. R. Lynch et al., "Computed tomographic atlas for the new international lymph node map for lung cancer: a radiation oncologist perspective," *Pract. Radiat. Oncol.* **3**, 54–66 (2013).
23. P. Li et al., "A large-scale CT and PET/CT dataset for lung cancer diagnosis (Lung-PET-CT-Dx)," Cancer Imaging Archive (2020).
24. S. Gatidis et al., "A whole-body FDG-PET/CT dataset with manually annotated tumor lesions," *Sci. Data* **9**(1), 601 (2022).
25. R. Werner et al., "Estimation of lung motion fields in 4D CT data by variational non-linear intensity-based registration: a comparison and evaluation study," *Phys. Med. Biol.* **59**, 4247 (2014).
26. J. Wasserthal et al., "TotalSegmentator: robust segmentation of 104 anatomical structures in CT images," *Radiol. Artif. Intell.* **5**(5), e230024 (2023).
27. V. W. Rusch et al., "The IASLC lung cancer staging project: a proposal for a new international lymph node map in the forthcoming seventh edition of the TNM classification for lung cancer," *J. Thorac. Oncol.* **4**(5), 568–577 (2009).
28. M. Nielsen et al., "Time matters: handling spatio-temporal perfusion information for automated TICI scoring," in *Med. Image Comput. Comput. Assist. Interv.* Vol. 12266, pp. 86–96, Springer, Cham, Switzerland (2020).
29. H. R. Roth et al., "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Med. Image Comput. Comput. Assist. Interv.* Vol. 8673, 520–527, Springer, Cham, Switzerland (2014).
30. D. Bouget et al., "Mediastinal lymph nodes segmentation using 3D convolutional neural network ensembles and anatomical priors guiding," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **11**, 44–58 (2022).
31. J. Hoffman et al., "Automatic identification of IASLC-defined mediastinal lymph node stations on CT scans using multi-atlas organ segmentation," *Proc. SPIE* **9414**, 94141R (2015).
32. J. Vansteenkiste et al., "Lymph node staging in non-small-cell lung cancer with FDG-PET scan: a prospective study on 690 lymph node stations from 68 patients," *J. Clin. Oncol.* **16**, 2142–2149 (1998).
33. A. S. Bryant et al., "Maximum standard uptake value of mediastinal lymph nodes on integrated FDG-PET-CT predicts pathology in patients with non-small cell lung cancer," *Ann. Thorac. Surg.* **82**(2), 417–423 (2006).
34. D. Hellwig et al., "18F-FDG PET for mediastinal staging of lung cancer: which SUV threshold makes sense?," *J. Nucl. Med.* **48**(11), 1761–1766 (2007).
35. S. Gatidis et al., "Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging," *Nat. Mach. Intell.* **6**, 1396–1405 (2024).
36. J. Ye et al., "Exploring vanilla U-Net for lesion segmentation from whole-body FDG-PET/CT scans," (2022).
37. Y. Wang et al., "Deep learning for automatic prediction of lymph node station metastasis in esophageal cancer patients from contrast-enhanced CT," *Int. J. Radiat. Oncol. Biol. Phys.* **117**, S55 (2023).
38. L. Kepka et al., "Delineation variation of lymph node stations for treatment planning in lung cancer radiotherapy," *Radiother. Oncol.* **85**, 450–455 (2007).

Sofija Engelson received her MScience degree in data science from the University of Lüneburg in 2022. She has been working as a researcher at the University Medical-Center Hamburg-Eppendorf, University of Lübeck, and she is currently employed at the German Research Center for Artificial Intelligence as part of the working group for AI in medical image and signal processing. Her research focuses on the integration of prior knowledge into learning-based medical image and text analysis.

Yannic Elser is a junior radiologist at the Institute of Radiology and Nuclear Medicine, University Hospital Schleswig-Holstein (UKSH), Lübeck. He completed his medical degree at the University of Lübeck in 2021 and is currently pursuing his doctoral thesis under the supervision of PD Dr med. Kim Honselmann, head of the Molecular Pancreas Research Group at UKSH Lübeck. As a member of the AI research group led by PD Dr Sieren, he has coauthored several publications in the field of cardiac and thoracic artificial intelligence research.

Malte Maria Sieren received his MD from the University of Lübeck in 2016 and became a board-certified radiologist at UKSH in 2022, where he leads the artificial intelligence in radiology group. His research focuses on AI-assisted image analysis, privacy-preserving data generation, and quantitative cardiovascular and thoracic imaging. He was habilitated in 2024 on AI-based aortic aneurysm diagnosis and has authored over 60 publications and holds two international patents.

Jan Ehrhardt earned his PhD in computer science from the University of Lübeck, Germany, in 2004. He was a postdoctoral researcher at the University Medical Center Hamburg-Eppendorf (2004–2010) and has since worked at the University of Lübeck. His research focuses on medical image analysis, particularly on optimization-based and learning-based methods for image segmentation and registration. He also works on generative models for creating realistic anatomical and pathological images.

Julia Andresen obtained her BSc and MSc degrees in mathematics in medicine and life sciences from the University of Lübeck, Germany, in 2017 and 2020, respectively. In 2025, she submitted her PhD thesis at the University of Lübeck, focusing on deep learning–based image registration for unsupervised medical image segmentation. Her recent research has centered on the analysis of image time series to study disease-related changes. She currently works as an AI Test Engineer at FPI Food Processing Innovation GmbH.

Stefanie Schierholz: Biography is not available.

Tobias Keck began his medical career in surgery at the University of Heidelberg and conducted pancreatic research at Massachusetts General Hospital, Harvard Medical School. He later became vice chair in Freiburg and, since 2012, chair of surgery at the University of Lübeck. His clinical focus is minimally invasive oncologic surgery. He serves as vice chair of the CCC Lübeck, leads Sector Health at Fraunhofer IMTE, is president-elect 2026 of the Association of General and Visceral Surgeons in Germany, and holds multiple honorary titles.

Daniel Drömann received his MD in 2001 from the University of Lübeck, where he became a specialist in internal medicine in 2007 and a specialist in pneumology in 2008. He was appointed professor in 2022. Since 2006, he has worked in pneumology and internal medicine, serving as a senior physician from 2008 and as director of the Department of Medicine III at UKSH, Lübeck, since 2017. His research focuses on minimally invasive diagnostics, imaging, clinical trials, and translational oncology projects.

Jörg Barkhausen started his medical studies at the University of Essen, Germany, in 1986, where he graduated in 1993. He interned in the Department of Cardiology in Mülheim and returned to the University of Essen in 1995 for his residency in diagnostic radiology. He completed his training with his board certification in diagnostic radiology in 1999. Since 2008, he has been chairman of the Department for Radiology and Nuclear Medicine at the University Hospital in Lübeck, Germany.

Heinz Handels received his PhD in computer science from RWTH Aachen University in 1991. He became a full professor of medical informatics at the University of Hamburg in 2003 and, since 2010, has held the same position at the University of Lübeck, where he directs the Institute for Medical Informatics. Since 2020, he has also led the DFKI Research Department "AI in Medical Image and Signal Processing." His research focuses on trustworthy machine learning for medical image analysis.