Full citation: Sofija Engelson, Jan Ehrhardt, Timo Kepp, Joshua Niemeijer, Stefanie Schierholz, Lennart Berkel, Yannic Elser, Malte Maria Sieren and Heinz Handels "Comparison of anatomical priors for learning-based neural network guidance for mediastinal lymph node segmentation", Proc. SPIE 12927, Medical Imaging 2024: Computer-Aided Diagnosis, 129271K (April 3, 2024).
https://doi.org/10.1117/12.3006158
Copyright 2024 Society of PhotoOptical Instrumentation Engineers (SPIE). One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this publication for a fee or for commercial purposes, and modification of the contents of the publication are prohibited.

Comparison of Anatomical Priors for Learning-based Neural Network Guidance for Mediastinal Lymph Node Segmentation

Sofija Engelson[a], Jan Ehrhardt[a,b], Timo Kepp[b], Joshua Niemeijer[c], Stefanie Schierholz[d], Lennart Berkel[e], Yannic Elser[e], Malte Maria Sieren[e], and Heinz Handels[a,b]

[a]Institute of Medical Informatics, University of Lübeck, Germany
[b]German Research Center for Artificial Intelligence, Lübeck, Germany
[c]German Aerospace Center, Braunschweig, Germany
[d]Department of Surgery, University Medical Center Schleswig-Holstein, Lübeck, Germany
[e]Department of Radiology and Nuclear Medicine & Institute of Interventional Radiology, University Medical Center Schleswig-Holstein, Lübeck, Germany

ABSTRACT

The assessment of lymph node metastases is critical for accurate cancer staging and consequently the decision for treatment options. Lymph node staging is a challenging, time-consuming task due to the fact that lymph nodes have ill-defined borders as well as varying sizes and morphological characteristics. The purpose of this study is to evaluate the effects of using different anatomical priors with the aim of guiding network attention within the application of segmentation of pathological lymph nodes in the mediastinum. The first presented prior, a distance map, displays the distance to a commonly defined point across all patients and, thus, provides an orientation of where a patch is extracted from. The second prior option, a probabilistic lymph node atlas, provides a map of areas where healthy and pathological lymph nodes are located, but also highlights lymph node stations that are more likely to become malignant. The distance map as well as the probabilistic lymph node atlas are results of an upstream atlas-to-patient registration approach. The third prior is a combination of segmentation masks of anatomical structures generated by the TotalSegmentator algorithm. A paired t-test on 5-fold cross validated results shows no significant differences in Dice score between models trained with the distance map or/and the probabilistic lymph node atlas compared to models trained with CT only. Counterintuitively, the models trained with segmentation masks of selected anatomical structures show significantly decreased segmentation accuracy. However, using the probabilistic lymph node atlas reduces the number of false negatives and consistently elevates the effect of post-processing.

Keywords: Mediastinal Lymph Node Segmentation, Anatomical Priors, Probabilistic Atlas, nnU-Net

1. INTRODUCTION

As part of the Tumor Node Metastases (TNM) classification for cancer staging, the N-staging sheds light on the infestation of metastases in regional lymph nodes. The enlargement of lymph nodes occurs as a response by the immune system, whereby a major contributing factor is the infiltration of tumor cells. In the field of medical imaging, PET/CT scans are oftentimes used to evaluate malignancy of lymph nodes. However, in cases where PET scans are unavailable, medical professionals rely on CT images and assess lymph node size based on a predefined rule set known as Response Evaluation Criteria in Solid Tumors (RECIST).[1] According to RECIST guidelines, a lymph node is considered pathological if its short-axis diameter exceeds 10 mm. Automatic segmentation of enlarged lymph nodes in CT images can serve as a basis for decisions regarding the necessity of surgical intervention and further treatment, and support automatic tumor staging based on both PET/CT or CT only. However, segmentation of lymph nodes in CT image data is difficult because, on the one hand, contrast differences to surrounding tissue are marginal and, on the other hand, lymph nodes strongly vary in size, shape, number, and location.

Send correspondence to S. Engelson: sofija.engelson@uni-luebeck.de

Automatic lymph node classification, detection, and segmentation has a long history in medical research. Feuerstein et al.[2] transfer the use of atlases for segmentation purposes from the field of brain imaging to lymphatic tissue in the chest. Similarly, Feulner et al.[3] create a probabilistic atlas from lymph node segmentation masks and use this as a spatial prior for a multistep approach based on conventional methods. With Roth et al.,[4] who provide a data set of 3D CT volumes of 90 patients and 388 segmented lymph nodes, learning-based methods become popular in this field of research. The authors train a CNN using three reformatted, orthogonal 2D slices through the centroid coordinates of a volume of interest. Multiple authors build upon this work and experiment with variations of network architectures. Iuga et al.[5] introduce a 3D fully convolutional foveal neural network, which extracts features at different resolutions. Nayan et al.[6] test a modified upsampling strategy for U-Net++. A combination of both worlds – the introduction of spatial prior information and learning-based approaches, is proposed by Bouget et al.[7] They use segmentation masks of the esophagus and other anatomical structures as additional channel input to a 3D U-Net with the aim to prevent the network from generating false positives in these areas.

This work focuses on the comparison of different spatial anatomical priors consolidated as additional input to deep learning methods for the segmentation of enlarged, mediastinal lymph nodes. We propose upstream atlas-to-patient registration to generate strong anatomical priors, in turn, to assist the neural network in overcoming the challenges of robustly detecting lymph nodes. Here, the priors used were a distance maps normalized to the atlas' coordinate system, probability maps for the occurrence of lymph nodes, and segmentation masks of various anatomical structures. We used a modified version of nnU-Net as the network architecture. In our experiments, we investigated the influence of different prior information on segmentation accuracy and selected distance metrics.

2. METHODS

As a basis for our segmentation network, we used the nnU-Net[8] with additional residual connections in the encoder.[9,10] Standard augmentations such as random cropping, rotating, and flipping were used.

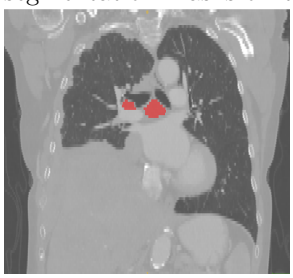
2.1 Generation of Anatomical Priors and Pre-processing

Three anatomical priors were investigated for comparison, which were defined and generated as follows: A *distance map* (DM) was calculated with respect to a manually defined point (underneath the bifurcation of the trachea) on a reference patient, our atlas.[11] This distance map was then warped to all training patients, resulting in an individual distance map for each patient normalized to a standard coordinate system. This prior provides coordinates relating to a reference anatomy, so that identical anatomical localizations have identical coordinates in all images. The motivation behind using this prior is that it provides an orientation of where a patch is extracted from. The *probabilistic lymph node atlas* (PA) originated from the registration of annotated, publicly available CT images of 119 patients[4,7] to the atlas. The registered segmentation masks were averaged to create a probability map and warped to each patient image to indicate potential regions with lymph nodes. In addition, this prior option highlights lymph node stations (i.e. 4, 7, and 10) that are more likely to become malignant due to their proximity to the lungs. The *segmentation masks* (SM) for the third prior option resulted from a selection of anatomical structures, e.g. bones, heart, esophagus, trachea and aorta, segmented by the TotalSegmentator algorithm.[12] Segmentation masks of anatomical structures can also serve better orientation, as the same structures are marked with the same class value throughout all patients. Additionally, all segmented regions besides the background indicate where lymph nodes are not to be found. The suggested collection of priors can be reviewed in Fig. 1 and was provided in various combinations as network input in addition to the CT data.

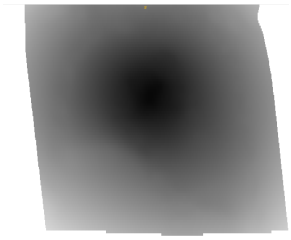
The registration pipeline of the atlas to the input data consisted of a rigid registration (first step), followed by an affine registration (second step), and finally a non-linear registration using ITK's VariationalRegistration module[13] (third step). As mentioned in Sec. 1, the CT images are strongly heterogeneous, vary in field of view, and contain pathologies, compromising the robustness of pure intensity-based registration. Therefore, rigid and affine registration was based on the anatomical segmentation masks described above.

Further pre-processing, included cropping all input data to the size of the segmentation mask of the lung by using the lung masks from TotalSegmentator[12] to reduce computational costs. Default CT normalization of the

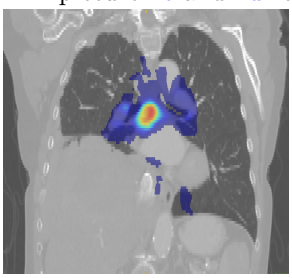
Figure 1: Image 1a shows the CT image and the according lymph node segmentation masks of an exemplary patient. Image 1b, the distance map, has its center approximately underneath the bifurcation of the trachea. The deformation of the map is visible, which comes from the fact that distances are measured according to the atlas' coordinate system. The probabilistic lymph node atlas registered on the exemplary patient as well as the segmentation masks of its anatomical structures are shown in picture 1c and 1d respectively.



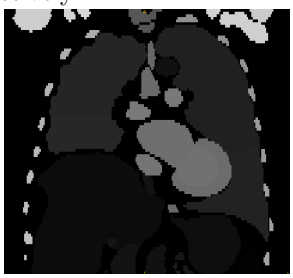
(a) CT and lymph node segmentation mask



(b) Distance map



(c) Warped probabilistic lymph node atlas



(d) Segmentation masks from Totalsegmentator

nnU-Net, that is, taking the 0.5^{th} and 99.5^{th} percentile of all intensity values of the foreground class, was applied to the input data. The resulting values are similar to the soft tissue intensity window. The distance maps were divided by the largest value of all distance maps, the probabilities of lymph node occurrence were smoothed with a Gaussian filter ($\sigma = 5$) and scaled to a range between zero and one.

2.2 Post-processing

The post-processing consisted of padding the resulting masks to the original input size, accounting for the strong class imbalance between fore- and background by reducing the threshold for binarization, and removing some falsely segmented pixels. The threshold for class binarization for pixel at position $(i, j, k) \in \mathbb{N}^{m \times n \times s}$ was calculated according to:

$$m_{ijk} = t \times (1 - 0.5 \times p_{ijk}), \tag{1}$$

where m_{ijk} is the threshold depending on the probabilistic lymph node atlas at pixel with indices i, j, k, the constant threshold is denominated with t, and p_{ijk} is a pixel of the probabilistic lymph node atlas P. Similar to Bouget et al.,[7] the threshold t was set to 0.5, 0.3, or 0.2. By predefining areas of where lymph nodes are more likely to occur, we can allow for more uncertainty in a controlled manner. This resulted in enlarged segmentation masks. To fulfill the RECIST criterion, we restrict the ellipsoid diameter of a connected component to be larger than 7 mm, 5 mm or 3 mm.

3. RESULTS

The self-configuration process of the nnU-Net configuration set the patch size to 128×112×160 and a batch size of two. The learning rate scheduling was modified to be as follows: From epoch 1 to 1,040 the learning rate linearly decreases from 0.01 to 2×10^{-5}, from epoch 1,040 to the end of training (epoch 2,000) the learning rate linearly reduces to 4×10^{-8}. In this way, the learning rate decreases at a larger rate in the first training half than in the second half to encourage fast weight adaptation.

For our experiments, two different data sets provided by Roth et al.[4] and Bouget et al.[7] were used with 119 thoracic and abdominal CT images in total. Image resolution varies between 0.58 and 0.97 mm^3 in-plane and 0.5 to 5.0 mm^3 between slices. The manual lymph node segmentations provided for the training data show a large heterogeneity. That is, lymph nodes smaller than 10 mm in short-axis diameter are contained and, in some cases, rather regions or lymph node stations are delineated instead of single lymph nodes with identifiable borders.

For our analysis, in addition to the Dice score, we report two distance metrics – the 95^{th} percentile of the Hausdorff Distance (HD95) and the Average Symmetric Surface Distance (ASSD). While the Hausdorff distance shows the maximum, the ASSD shows the average of all shortest distances for all points of the segmented lymph node in the prediction compared to the ground truth. As it is often the case in medical applications, it is

favorable to accept a higher false positive rate at the cost of a lower false negative rate. To assess the methods regarding this, we report the Recall and the number of false positives (FP) and false negatives (FN). The metric *LN found* indicates the percentage of lymph nodes segmented in both the prediction and the ground truth annotation from all ground truth annotated lymph nodes, in more mathematical terms: $\frac{\text{true positives}}{\text{true positives+false negatives}}$. The computation of this metric follows the implementation introduced by the organizers of the BRATS 2023 challenge. Originally designed for assessing lesion segmentation in brain images, the evaluation code can be found at: https://github.com/rachitsaluja/BraTS-2023-Metrics.

Table 1: Results using 5-fold cross-validation and different input combinations for model training without post-processing. Dice scores and distance metrics significantly different to training on CT only ($p < 0.05$) are marked in *italic*.

Dataset	Dice	HD95 [mm]	ASSD [mm]	Recall	FP	FN	LN found
CT	0.6417 + 0.0297	41.7915 ± 8.5516	6.1019 ± 0.7651	0.6171 ± 0.0345	507,596	769,793	0.6891
CT & DM	0.6414 ± 0.0386	38.8456 ± 7.1209	6.0841 ± 0.8221	0.6135 ± 0.0359	516,147	738,325	0.6736
CT & PA	0.6323 ± 0.0514	39.9907 ± 4.4516	5.9800 ± 0.9250	0.6142 ± 0.0451	525,049	736,361	0.6698
CT & DM & PA	0.6321 ± 0.0381	39.2634 ± 3.9633	5.9974 ± 0.5087	0.6178 ± 0.0393	567,977	730,980	0.6767
CT & SM	*0.6144 ± 0.0481*	40.9410 ± 3.3736	6.3296 ± 0.7831	0.6046 ± 0.0428	560,350	747,953	0.6566

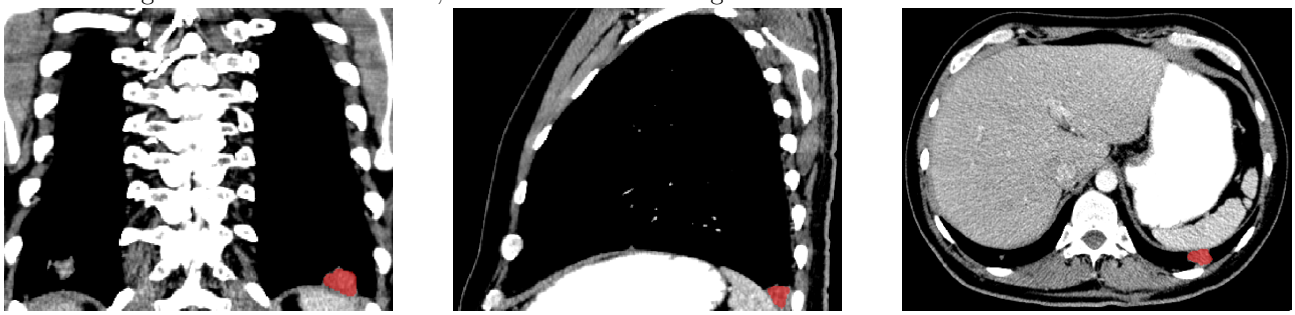
Table 2: Results using 5-fold cross-validation and different input combinations for model training with post-processing. Dice scores and distance metrics significantly different to training on CT only ($p < 0.05$) are marked in *italic*.

Dataset	Dice	HD95 [mm]	ASSD [mm]	Recall	FP	FN	LN found
CT	0.6477 ± 0.0289	43.6615 ± 9.6834	6.2353 ± 0.9130	0.6547 ± 0.0346	619789	700841	0.7115
CT & DM	0.6483 ± 0.0378	*39.1601 ± 7.0599*	6.0755 ± 0.9324	0.6536 ± 0.0364	616,594	670,527	0.6968
CT & PA	0.6374 ± 0.0490	41.9950 ± 5.3156	6.1922 ± 0.9477	0.6538 ± 0.0429	646,646	655,010	0.6953
CT & DM & PA	0.6369 ± 0.0372	*37.8621 ± 3.3415*	5.9690 ± 0.5932	0.6558 ± 0.0382	685,689	656,769	0.6976
CT & SM	*0.6184 ± 0.0475*	40.7263 ± 2.1792	6.3622 ± 0.6243	0.6362 ± 0.0463	665,955	683,977	0.6790

Models with varying combinations of priors were trained and averaged over five folds. Fold splits, network architecture and training strategy were the same for each model run. Each model was tested on a left-out test dataset per fold. The results with and without post-processing are shown in Tab. 1 and Tab. 2 respectively. The optimal post-processing hyperparameters were set via grid search. For significance testing, we carried out a paired t-test for the Dice scores and the distance metrics for the models with and without post-processing. The results of the paired t-test showed that the Dice coefficient of CT & DM, CT & PA, CT & DM & PA, as well as the model, trained on CT only varies marginally (p-value > 0.1), showing that differences between these models originated probably from random variations in training. The segmentation accuracy was within the range of $0.6369 - 0.6483$, which is in alignment with related literature.[7] Solely, the model trained with the segmentation masks consistently performed worse than the other models throughout the folds (p-value $= 0.0004$). Using priors reduced the HD95, in the case of CT & DM and CT & DM & PA significantly, if post-processing was applied. This is also reflected by the fact that the standard deviation for the baseline model was larger than for the models trained with priors. Fig. 2 shows the prediction results of the model trained on CT only for an exemplary patient. The model segmented a possibly malignant, pulmonary nodule in the bottom of the lung, which was not annotated in the ground truth as it is not a lymph node. Consequently, the HD95 became large. The ASSD was lower when the proposed priors were used in comparison to the baseline model, but the differences were not significant. It is noteworthy that the predictions of models trained with the probabilistic lymph node atlas have a smaller number of false negatives. However, this did not have an effect on the number of lymph nodes found. For all models, the optimal post-processing strategy was, to use a threshold that depends on the probabilistic atlas with a reduced constant threshold for binarization. For most models, constant threshold t was set to 0.2. This post-processing step reduced the number of false negatives, and, in this way, segmentation accuracy can be improved. Removing small segmentation masks did not improve results, thus, this post-processing step was not applied.

Even though it is surprising that additional information did not improve segmentation accuracy, these results align with Bouget et al.[7] The authors show that the model trained with anatomical priors produces less false

Figure 2: Example prediction of model trained on CT only overlaid over CT in coronal, sagittal and axial view. Here, a pulmonary nodule outside the mediastinum is segmented. As this is not a lymph node and, thus, not included in ground truth annotations, the HD95 becomes large.



positives in the no-go regions defined in the prior, but overall performance was not, or not significantly, superior to models trained using CT images alone.

4. DISCUSSION AND CONCLUSION

In this work, we present a fully automatic pipeline for mediastinal lymph node segmentation by using results from atlas registration to serve as a road map and present a comparison of performance differences using different anatomical priors as additional network input. Our comparison of anatomical priors for mediastinal lymph node segmentation shows that including anatomical priors in network training does not improve performance significantly. Variations between the priors, except for CT & SM, are also minimal. On the one hand, the information in the CT seems to be sufficient for neural networks to learning patterns. On the other hand, other reasons such as insufficient model-dependent hyperparameter tuning could play a role in anatomical priors not unfolding their full potential. However, using the probabilistic lymph node atlas as network input reduces the number of false negatives and consistently elevates the effect of post-processing.

If human learning processes are transferrable to neural networks, use cases of high complexity and little training data can benefit from additional information that is congruent for all data points. Based on this, we extended Bouget et al.'s [7] work by formulating different, potentially more intuitive, anatomical priors. But, providing the suggested priors as additional input does not seem to be sufficient to guide network attention. Instead of focussing on the formulation of priors, further research could explore other techniques to incorporate priors into network training. Possibly, integrating additional information in the training process in a supervised manner would ensure their use.

REFERENCES

[1] Eisenhauer, E., Therasse, P., Bogaerts, J., Schwartz, L., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., and Verweij, J., "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer* **45**(2), 228–247 (2009). Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers.

[2] Feuerstein, M., Glocker, B., Kitasaka, T., Nakamura, Y., Iwano, S., and Mori, K., "Mediastinal atlas creation from 3-D chest computed tomography images: Application to automated detection and station mapping of lymph nodes," *Medical Image Analysis* **16**(1), 63–74 (2012).

[3] Feulner, J., Kevin Zhou, S., Hammon, M., Hornegger, J., and Comaniciu, D., "Lymph node detection and segmentation in chest CT data using discriminative learning and a spatial prior," *Medical Image Analysis* **17**(2), 254–270 (2013).

[4] Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., and Summers, R. M., "A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations," in [*MICCAI 2014*], 520–527, Springer (2014).

[5] Iuga, A., Carolus, H., Höink, A., Brosch, T., Klinder, T., Maintz, D., Persigehl, T., Baeßler, B., and Püsken, M., "Automated detection and segmentation of thoracic lymph nodes from CT using 3D foveal fully convolutional neural networks," *BMC Medical Imaging* **21** (04 2021).

[6] Nayan, A.-A., Kijsirikul, B., and Iwahori, Y., "Mediastinal Lymph Node Detection and Segmentation Using Deep Learning," *IEEE Access* **10**, 89289–89307 (2022).

[7] Bouget, D., Pedersen, A., Vanel, J., Leira, H. O., and Langø, T., "Mediastinal lymph nodes segmentation using 3D convolutional neural network ensembles and anatomical priors guiding," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **11**, 44 – 58 (2021).

[8] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods* **18**(2), 203—211 (2021).

[9] McConnell, N., Miron, A., Wang, Z., and Li, Y., "Integrating Residual, Dense, and Inception Blocks into the nnUNet," in [*2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*], 217–222 (2022).

[10] Isensee, F., Ulrich, C., Wald, T., and Maier-Hein, K. H., "Extending nnU-Net Is All You Need," in [*Bildverarbeitung für die Medizin 2023*], Deserno, T. M., Handels, H., Maier, A., Maier-Hein, K., Palm, C., and Tolxdorff, T., eds., 12–17, Springer Fachmedien Wiesbaden, Wiesbaden (2023).

[11] Lynch, R., Pitson, G., Ball, D., Claude, L., and Sarrut, D., "Computed tomographic atlas for the new international lymph node map for lung cancer: A radiation oncologist perspective," *Practical Radiation Oncology* **3**, 54–66 (Jan. 2013).

[12] Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., and Segeroth, M., "TotalSegmentator: robust segmentation of 104 anatomical structures in CT images," *Radiology: Artificial Intelligence* (2023).

[13] Werner, R., Schmidt-Richberg, A., Handels, H., and Ehrhardt, J., "Estimation of lung motion fields in 4D CT data by variational non-linear intensity-based registration: A comparison and evaluation study," *Physics in Medicine & Biology* **59**, 4247 (July 2014).