PLOS ONE

RESEARCH ARTICLE

# Teacher-student approach for lung tumor segmentation from mixed-supervised datasets

**Vemund Fredriksen[1☯], Svein Ole M. Sevle[1☯], André Pedersen[2,3,4]*, Thomas Langø[2,5], Gabriel Kiss[1,5], Frank Lindseth[1,6]**

**1** Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, **2** Department of Health Research, Medical Technology, SINTEF, Trondheim, Norway, **3** Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, **4** Clinic of Surgery, St. Olavs hospital, Trondheim University Hospital, Trondheim, Norway, **5** Research Department, Future Operating Room, St. Olavs hospital, Trondheim University Hospital, Trondheim, Norway, **6** Norwegian Open AI Lab, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

☯ These authors contributed equally to this work.
* andre.pedersen@ntnu.no

## Abstract

### Purpose

Cancer is among the leading causes of death in the developed world, and lung cancer is the most lethal type. Early detection is crucial for better prognosis, but can be resource intensive to achieve. Automating tasks such as lung tumor localization and segmentation in radiological images can free valuable time for radiologists and other clinical personnel. Convolutional neural networks may be suited for such tasks, but require substantial amounts of labeled data to train. Obtaining labeled data is a challenge, especially in the medical domain.

### Methods

This paper investigates the use of a teacher-student design to utilize datasets with different types of supervision to train an automatic model performing pulmonary tumor segmentation on computed tomography images. The framework consists of two models: the student that performs end-to-end automatic tumor segmentation and the teacher that supplies the student additional pseudo-annotated data during training.

### Results

Using only a small proportion of semantically labeled data and a large number of bounding box annotated data, we achieved competitive performance using a teacher-student design. Models trained on larger amounts of semantic annotations did not perform better than those trained on teacher-annotated data. Our model trained on a small number of semantically labeled data achieved a mean dice similarity coefficient of 71.0 on the MSD Lung dataset.

OPEN ACCESS

**Citation:** Fredriksen V, Sevle SOM, Pedersen A, Langø T, Kiss G, Lindseth F (2022) Teacher-student approach for lung tumor segmentation from mixed-supervised datasets. PLoS ONE 17(4): e0266147. https://doi.org/10.1371/journal.pone.0266147

**Editor:** Robertas Damaševičius, Politechnika Slaska, POLAND

**Received:** January 6, 2022

**Accepted:** March 15, 2022

**Published:** April 5, 2022

**Data Availability Statement:** All data used in the study are fully open data sets, available to all as specified in the manuscript. We have not performed any ground truth sanitation, and the distribution between training, validation, and test sets are described in the manuscript. In addition, we have made our best model available on GitHub (https://github.com/VemundFredriksen/LungTumorMask), which enables others to benchmark our design against theirs on the same or new data.

**Funding:** This work was supported by The Norwegian National Advisory Unit for Ultrasound and Image-Guided therapy at St. Olavs hospital and The Liaison Committee for Education, Research and Innovation in Central Norway (No. 2018/42794) and The Cancer Foundation, St. Olavs hospital, Trondheim University Hospital (No. 13/2021) and SINTEF and The EEA Grant project "Improving Cancer Diagnostics in Flexible Endoscopy Using Artificial Intelligence and Medical Robotics (IDEAR)" (No. 19/2020).

**Competing interests:** The authors have declared that no competing interests exist.

## Conclusions

Our results demonstrate the potential of utilizing teacher-student designs to reduce the annotation load, as less supervised annotation schemes may be performed, without any real degradation in segmentation accuracy.

## Introduction

Cancer is becoming the leading cause of death and the most significant obstacle to increase life expectancy in many countries [1]. Lung cancer, accounting for more than 11% of all new cases, is the second most common cancer and it ranks first among the cancer-related mortality worldwide, accounting for 18% of the total cancer deaths [2]. The most common lung cancer treatments include: surgical resection, chemotherapy, radiotherapy, and immunotherapy. Many of these treatments, and also the successful diagnosis with bronchoscopy or computed tomography (CT)-guided biopsy, depend on accurately locating, and in many cases delineating (segmenting), the tumor from normal tissue in the preoperative images, typically CT.

Manual segmentation of the lesions/tumors from preoperative CT is a laborious and tedious process for oncologists, radiologists, and pulmonologists, which could result in delays of treatment and lower the survival rates, especially in clinics with inadequate resources. In addition, the quality of manual localization and segmentation relies on prior knowledge and clinical expertise. Even with adequate guidelines and standards, tumor segmentation is often prone to high inter- and intraobserver variability. On the other hand, automatic segmentation techniques has the potential to provide efficient, consistent, and more accurate results. Automatic methods can both shorten the time needed to read the images and they also allow experts to devote their limited time to optimize planning and treatment planning.

### Related work

Historically, methods like thresholding [3], region growing [4, 5], and graph cuts [6] were commonly proposed to segment lung tumors from CT images. These algorithms are suitable as semi-automatic methods, but are not suited for localization of lung tumors. Recent advancements in deep learning enables automation of tasks that until recently was only performed by trained experts [7, 8]. Advancements in hardware has enabled development of larger and increasingly complex models, but much of the improvement is caused by access to large amounts of annotated data.

Today, especially after the introduction of the U-Net [9] architecture, deep learning methods have dominated the field of medical image segmentation [10]. However, convolutional neural networks (CNNs) are memory intensive, especially for 3D volumes. It is therefore common to train networks based on 2D or 2.5D input images [11], where the model evaluates one slice at a time, chunks of slices, or 3D patches, and then applied in a sliding windows fashion across the CT volume. Patching the 3D volume comes at a cost of loss of perception, and thus more efficient multi-scale 3D CNN architectures have been proposed, which enables the use of larger input volumes [12]. An alternative approach is to perform segmentation in multiple steps, either using multiple algorithms or a cascade of CNNs [11, 13–15].

To accommodate the issue of lacking training data, unsupervised methods like supervoxel has been proposed [16]. To facilitate faster convergence and more accurate results, multimodality methods that utilize magnetic resonance imaging (MRI) or positron emission tomography (PET) scans in addition to CT have been suggested [17–19]. Neural network architectures that can utilize multiple annotation types has also been suggested [20, 21].

A more recent strategy to accommodate the lack of training data is the teacher-student design, inspired by the concept of knowledge distillation [22–24]. The teacher creates pseudo-annotations from suboptimal annotations to increase the dataset size for training the student. The teacher-student pattern can be applied to any type of network architecture, and does not dictate other hyperparameters or external configurations. A teacher-student design can be used in different ways, from utilization of unlabeled data [25–27], to exploitation of multiple modalities [18, 19, 27], and to usage of datasets with different annotation types [28, 29].

## Contributions

Our approach differs from the previously mentioned methods applied to lung tumor segmentation by utilizing annotations of different supervision on the same modality, namely CT images. Since CT scanning is less invasive to the patient than the other modalities, it is a goal to efficiently segment tumors from CT-only examinations. Our method is inspired by Sun *et al.* [28] that shows promising results using a teacher-student framework to segment liver and liver lesions given semantic and bounding box annotations. To the best of our knowledge, we are the first to implement a similar teacher-student framework for CT images to perform semantic segmentation of lung tumors. Our study suggests that even with a small dataset of semantic annotations, a student can achieve state-of-the-art performance given a large enough pseudo-annotated dataset to learn from.

# Materials and methods

Our method consists of two separate models: a semi-supervised teacher and a fully-automatic student. The method relies on two different annotation types: semantic 3D annotations and 2D bounding boxes in the axial planes. These we refer to as *strong* and *weak* annotations. Furthermore, we define our strongly and weakly annotated datasets as $D_s = \{S_x^i, S_y^i\}_{i=1}^{m_s}$ and $D_w = \{W_x^i, W_y^i\}_{i=1}^{m_w}$, respectively. An overview of our design can be seen in Fig 1.

## Data

To study the effect of our teacher-student framework we used three public datasets: Medical Segmentation Decathlon (MSD)-Lung [30], Non-Small Cell Lung Cancer (NSCLC)-Radiomics [31, 32], and Lung-PET-CT-Dx [33]. All three datasets contain manual annotations by human experts. The first two datasets consist of semantic annotations, whereas the latter dataset contains bounding box labels annotated in the axial plane.

The MSD-Lung dataset contains 64 images, whereas the NSCLC-Radiomics and Lung-PET-CT-Dx datasets contains 422 and 1295 images, respectively. Multiple images in the Lung-PET-CT-Dx dataset were discarded. The discarded images were either PET or PET/CT-fused, only contained a small portion of the thorax, or comprised of multiple scans stacked on top of each other. After removing all non-CT images and images with a real-world length (Z-axis) outside the range [16, 60] cm, 665 images from Lung-PET-CT-Dx remained in our dataset.

The three datasets varied in terms of voxel density and tumor sizes. Overall, the Lung-PET-CT-Dx and the NSCLC-Radiomics datasets contain larger tumors than the MSD-Lung dataset (see Table 1). The tumor diameter is an approximate size, measured by calculating the average of the longest and shortest diameter of the tumor in real-world coordinate space.

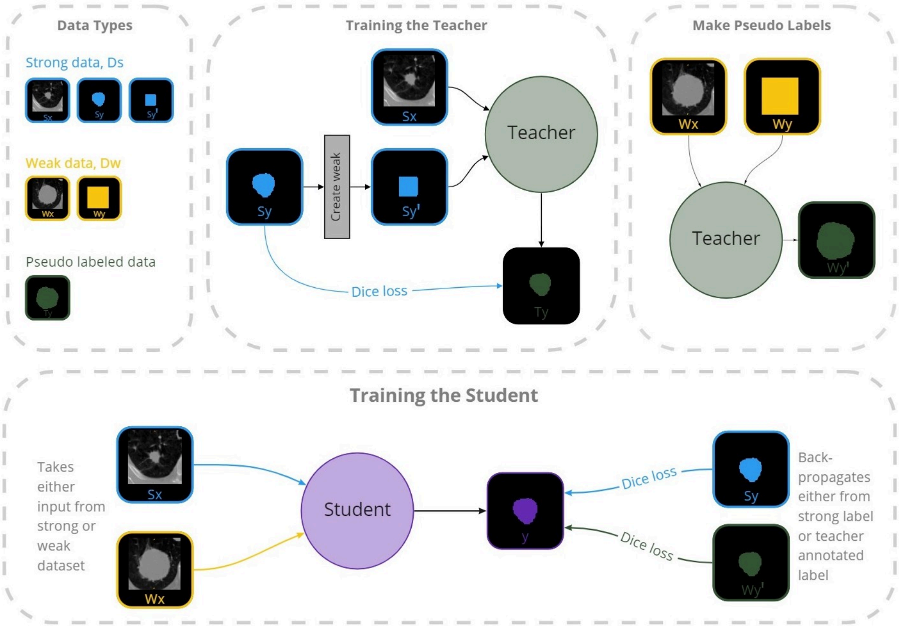

**Fig 1. The method overview during training.** Firstly, the teacher was trained using the strong dataset (semantic annotated images) represented with blue lines. The teacher was then used to make semantic annotations for the weak dataset (bounding box annotated images). The student was then trained on both the pseudo-annotated dataset $D_{w'}$, represented by the orange lines, and the strong dataset $D_s$.

https://doi.org/10.1371/journal.pone.0266147.g001

## Preprocessing

Our preprocessing pipeline consisted of multiple steps. Firstly, the voxel intensities were clipped to the range [-1024, 1000], before being standardized using the Z-score normalization method. The images' voxel spacing were then normalized to an anisotropic resolution of 1 × 1 × 1.5 mm$^3$. Lastly, a volumetric cropping was applied, which differed between the teacher and the student.

For the teacher, the images were cropped around the tumor with a fixed resolution of 128 × 128 × 128 voxels, whereas for the student, the images were split in two, each cropped around one of the lungs. The lungs were automatically segmented using the `lungmask` command line tool [34], and used when performing cropping around the lungs. The ground truth label images were voxel normalized and cropped in a similar manner as their corresponding CT image.

## Teacher-student design

The teacher was trained on 3D patches surrounding the tumor, guided by the corresponding bounding box annotations. Once trained, the teacher was applied to $D_w$ to generate pseudo-strong labels, $D_{w'}$. Although expert labeled images are the gold standard, teacher pseudo-

**Table 1. Tumor sizes of the three datasets.**

| Dataset | Volume [$cm^3$] | Diameter [$mm$] |
|---|---|---|
| MSD-Lung | 21.98 ± 51.66 | 37.63 ± 20.08 |
| NSCLC-Radiomics | 75.37 ± 96.30 | 63.63 ± 29.62 |
| Lung-PET-CT-Dx | 63.67 ± 86.26 | 48.66 ± 19.85 |

https://doi.org/10.1371/journal.pone.0266147.t001

annotated images can enhance training of fully automatic models, or even be used to aid experts in clinical use.

The student, like any ordinary automatic method, takes CT images as input and produces 3D segmentations of the potential lung tumors without user interaction. During training, the student exploits the pseudo-annotated images in $D_{W'}$ produced by the teacher, using the extended dataset, $\{D_s, D_{W'}\}$. Once trained, the student can perform end-to-end segmentation without human intervention. Algorithm 1 describes the training scheme. $W_x$ and $S_x$ denote inputs of weakly labeled and strongly labeled dataset, respectively. Likewise, $W_y$ and $S_y$ denote weak and strong annotations.

**Algorithm 1** Model training scheme

```
for each: $(S_x, S_y) \in D_s$    ▷ Train teacher on strong dataset
    $S_{y'} \leftarrow transform\_to\_weak(S_y)$    ▷ Create weak annotations from strong labels
    $W'_y \leftarrow teacher.predict(S_x, W_y)$    ▷ Prediction with image and weak label as input
    $loss \leftarrow calculate\_loss(S_y, W_{y'})$    ▷ Calculate loss from output and strong label
    $teacher.adjust\_weights(loss)$    ▷ Backpropagate after every batch
for each: $(W_x, W_y) \in D_w$    ▷ Annotate weak dataset with trained teacher
    $W'_y \leftarrow teacher.predict(W_x, W_y)$    ▷ Prediction with image and weak label as input
    $D_{W'}.insert(W_x, W_y)$    ▷ Store input and teacher-annotated output in $D_{W'}$
for each: $(x, y) \in D_s \bigcup D_{W'}$    ▷ Train student on extended dataset
    $y' \leftarrow student.predict(x)$    ▷ Prediction on image
    $loss \leftarrow calculate\_loss(y, y')$    ▷ Calculate loss from output and label
    $student.adjust\_weights(loss)$    ▷ Backpropagate after every batch
```

## Implementation

All our networks are based on the U-Net architecture [9], and share common building blocks (see Fig 2). U-Net was used as it performs well as a baseline architecture, and has shown competitive performance on various datasets from different modalities, of different organs, cancer types, and data types [14, 35].

The teacher consists of three levels, one of each downsampling operation, going from an image resolution of 128 × 128 × 128 to 16 × 16 × 16. The students are comprised of four levels. In contrast to the U-Net architecture, our design performs downsampling by applying 3D convolutions with a stride of two. We also substituted the ReLU [36] activation function with PReLU [37]. We implemented two related students: one that produces semantic segmentation output only, which we call the Single Output Student (SO Student), and one that produces an additional output approximating the bounding box surrounding the tumor in the axial plane, which we call the Dual Output Student (DO Student). The architecture of the student networks can be seen in Fig 2.

Firstly, the original U-Net design was too heavy to be applied in 3D directly. The architecture was therefore tuned to be better suited for the task and dataset. Hyperparameters were chosen through a systematic search. However, a rigorous search was not feasible due to the long training runtime. The teacher architecture and training hyperparameters were then frozen, before the teacher-student design was introduced. This was done to make comparison fair between the designs.

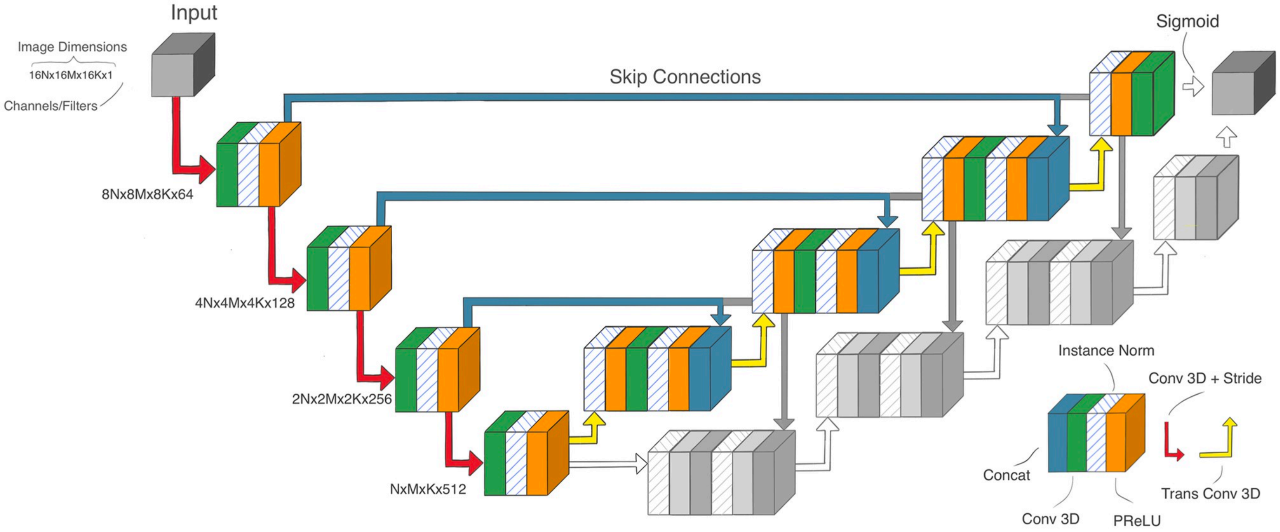

**Fig 2. The network architectures.** The single output (SO) Student is highlighted in colors whereas the decoder branch of the dual output (DO) Student is implicated in gray. For the DO Student the ouput of the two decoder branches are concatenated to form a dual-channeled output. The teachers have the same architecture as the SO Student, but with three downsamplings rather than four.

https://doi.org/10.1371/journal.pone.0266147.g002

The SO student was one level deeper than the teacher, but trained in the same manner. The DO student was identical to the SO student, but a second decoder branch was added to investigate the potential benefit of using both annotation types (semantic and bounding box labels) in training. The second decoder branch had a second loss to predict bounding boxes. The aim of the second branch was to improve localization of the tumor, as using the bounding box labels would make it learn different features.

The Adam [38] optimizer with a learning rate of $10^{-4}$ was used for training until DSC validation convergence. The batch size was set to one and virtually increased to eight using accumulated gradients. The gradients were computed using the Dice Loss function [39], based on the Dice Similarity Coefficient (DSC). The models were trained for a maximum of 350 epochs, or until overfitting occured. The best model was selected based on the lowest validation loss.

## Empirical evaluation

To evaluate our framework we considered two primary scenarios, each with two sub-experiments. We considered one scenario where the size of the strongly annotated dataset ($\sim$ 500 images) is similar to the size of the weakly annotated dataset ($\sim$ 750 images), and another scenario where the strongly annotated dataset was considerably smaller ($\sim$ 50 images) than the weakly annotated dataset($\sim$ 1000 images). Within each scenario, we evaluated two semi-supervised models and three fully-automatic models. Among the three fully-automatic models, one model was trained solely on strongly annotated data, whereas the two other were student networks trained both on strongly annotated data and the teacher-annotated pseudo labels.

We used different metrics for evaluating and comparing the models. The DSC was used to measure the semantic segmentation performance, whereas F1-score was used to determine object-wise localization performance. We also used DSC-TP to evaluate the segmentation accuracy considering only true positives (TPs). We considered objects to be true positives if there were $\geq$ 25% overlap between the predicted mask and the GT mask, motivated by a prior study [40].

**Table 2. Teacher results.**

| Dataset | Model | DSC |
|---|---|---|
| MSD-Lung | Point Guided | 74.78 ± 11.83 |
| | Box Guided | **84.91 ± 06.09** |
| NSCLC-Radiomics | Point Guided | 59.57 ± 23.90 |
| | Box Guided | **86.65 ± 08.77** |
| Both | Point Guided | 61.48 ± 23.29 |
| | Box Guided | **86.44 ± 08.50** |

The best performing model with respect to mean dice similarity coefficient (DSC) is highlighted in bold.

https://doi.org/10.1371/journal.pone.0266147.t002

The test set was sampled at random and accounted for 15% of the total dataset. The same split was used for all experiments to preserve fairness in evaluation. Patients with multiple scans were stratified into the three subsets: train, validation, and test. To counter the tumor size imbalance, we balanced the train and validation sets with regard to tumor sizes. Images containing tumors of more rare sizes were upsampled.

Models were trained using a workstation with a 14-core Intel Core i9 10940X @3.30 GHz CPU, 128 GB RAM, and two NVIDIA RTX 8000 (48 GB) GPUs. The most memory intensive student used, at its peak, ∼22.54GB VRAM during training, but inference can be performed with 3GB VRAM. Implementation was done in Python 3.7, built upon the MONAI [41] framework (v0.4.0), using PyTorch v1.6, and CUDA 11.0. The best performing model and corresponding inference code are made openly available as a command line tool at https://github.com/VemundFredriksen/LungTumorMask.

# Results

## Vast strongly annotated dataset

As seen in Table 2, the teacher guided by the bounding boxes, outperformed the point guided (without bounding boxes as input) teacher on both datasets in terms of DSC. The difference between the two models was less prominent measured on the MSD-Lung dataset than for the NSCLC-Radiomics dataset.

For the final inference models, the DSC was highest on the MSD-Lung dataset, across all three models (see Table 3). The best performing student network overall was the SO Student, with highest DSC on the MSD-Lung dataset. There was negligible difference between the three

**Table 3. Student results.**

| Dataset | Model | DSC | DSC-TP | F1-score | Recall | Precision |
|---|---|---|---|---|---|---|
| MSD-Lung | Baseline | **67.31 ± 21.17** | **73.12 ± 15.16** | **81.48 ± 31.86** | **88.89 ± 31.43** | **77.78 ± 34.25** |
| | SO Student | 64.27±16.05 | 71.32±8.06 | **81.48 ± 31.86** | **88.89 ± 31.43** | **77.78 ± 34.25** |
| | DO Student | 55.37 ± 29.03 | 70.49 ± 8.82 | 74.07 ± 40.91 | 77.78 ± 41.57 | 72.22 ± 41.57 |
| NSCLC-Radiomics | Baseline | 51.06 ± 28.22 | 68.81 ± 18.27 | 63.56 ± 36.36 | **83.82 ± 36.82** | 56.68 ± 38.65 |
| | SO Student | **52.92 ± 31.13** | **69.39 ± 19.21** | 64.18 ± 37.37 | 79.90 ± 39.66 | 58.76 ± 39.28 |
| | DO Student | 52.25 ± 30.18 | 68.69 ± 19.47 | **68.43 ± 38.71** | 79.17 ± 39.75 | **64.95 ± 40.81** |
| Both | Baseline | 52.96 ± 27.98 | 69.34 ± 17.97 | 65.66 ± 36.32 | **84.41 ± 36.27** | 59.14 ± 38.76 |
| | SO Student | **54.25 ± 29.99** | **69.63 ± 18.19** | 66.20 ± 37.19 | 80.95 ± 38.90 | 60.98 ± 39.21 |
| | DO Student | 52.61 ± 30.06 | 68.90 ± 18.59 | **69.09 ± 39.02** | 79.00 ± 39.97 | **65.80 ± 40.97** |

For each respective metric, the best performing models are highlighted in bold.

https://doi.org/10.1371/journal.pone.0266147.t003

**Table 4. Scarcely trained teacher results.**

| Dataset | Model | DSC |
|---|---|---|
| MSD-Lung | Scarce Point Guided | 48.52 ± 31.18 |
| | Scarce Box Guided | **81.65 ± 07.40** |
| NSCLC-Radiomics | Scarce Point Guided | 43.83 ± 25.65 |
| | Scarce Box Guided | **84.69 ± 06.59** |
| Both | Scarce Point Guided | 44.42 ± 26.45 |
| | Scarce Box Guided | **84.31 ± 06.77** |

The best performing model with respect to mean dice similarity coefficient (DSC) is highlighted in bold.

https://doi.org/10.1371/journal.pone.0266147.t004

models on the NSCLC-Radiomics dataset. The Baseline model performed best on the MSD-Lung dataset, both in terms of DSC and F1-score.

## Scarce strongly annotated dataset

When reducing the strongly labeled dataset, the performance of the point guided teacher was degraded, whereas the box guided teacher still performed well (see Table 4).

A similar trend applies to the final inference models (see Table 5). The baseline model performed poorer, whereas the student networks still performed well. The same can be seen from the object-wise metrics, although the difference was more prominent. Contrary to the results shown in Table 3, the SO Student had the highest DSC measured in this scenario. Fig 3 shows a sample of the outputs produced by the models in the scarce scenario.

# Discussion

In this paper, a teacher-student design to segment lung tumors from CTs has been proposed. Three datasets of two different annotation types were used for this purpose. The teacher model was first trained on the datasets that had strong annotations. It was then used to generate pseudo-strong annotations for the student. Both the teacher and the student used U-Net-like architectures, and were evaluated on segmentation performance. In addition, the student networks were evaluated on sensitivity to annotation type and sample size.

We observed that the box guided teacher outperformed the point guided teacher in both scenarios. This was expected as the bounding box annotations assist the teacher by serving as a segmentation and localization constraint. The effect of the box guidance is especially visible in

**Table 5. Scarcely trained student results.**

| Dataset | Model | DSC | DSC-TP | F1-score | Recall | Precision |
|---|---|---|---|---|---|---|
| MSD-Lung | Baseline | 26.45 ± 26.56 | 75.24 ± 15.90 | 09.10 ± 13.26 | 33.33 ± 47.14 | 05.31 ± 07.80 |
| | SO Student | 64.74 ± 11.82 | 71.56 ± 10.40 | 61.85 ± 16.49 | **100.0 ± 0.00** | 47.22 ± 20.79 |
| | DO Student | **71.00 ± 16.01** | **76.67 ± 07.08** | **85.18 ± 31.86** | 88.89 ± 31.43 | **83.33 ± 33.33** |
| NSCLC-Radiomics | Baseline | 28.23 ± 28.05 | 55.39 ± 22.80 | 32.13 ± 36.99 | 51.47 ± 49.24 | 26.99 ± 35.68 |
| | SO Student | 51.06 ± 30.75 | **68.56 ± 20.69** | 62.65 ± 36.73 | 79.41 ± 39.51 | 56.67 ± 38.79 |
| | DO Student | **53.89 ± 29.75** | 67.44 ± 21.70 | **66.96 ± 35.57** | **84.56 ± 35.62** | **60.44 ± 38.24** |
| Both | Baseline | 28.02 ± 27.89 | 56.91 ± 22.96 | 29.44 ± 35.82 | 49.35 ± 49.34 | 24.45 ± 34.35 |
| | SO Student | 52.66 ± 29.51 | **68.98 ± 19.60** | 62.55 ± 34.98 | 81.82 ± 37.72 | 55.56 ± 37.26 |
| | DO Student | **55.89 ± 29.01** | 68.56 ± 20.71 | **69.09 ± 35.64** | **85.06 ± 35.18** | **63.12 ± 38.41** |

For each respective metric, the best performing model is highlighted in bold.

https://doi.org/10.1371/journal.pone.0266147.t005

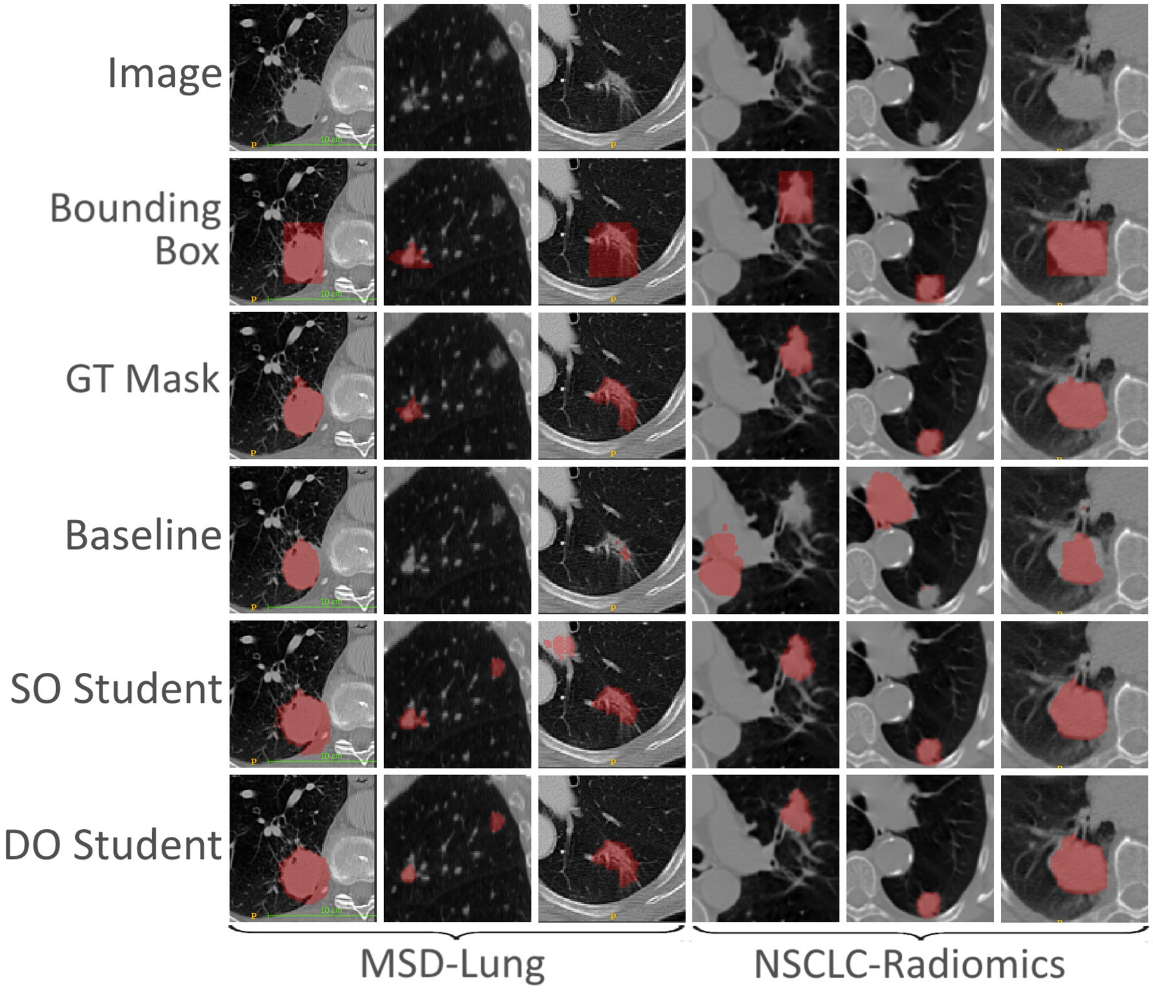

**Fig 3. A sample of the results produced by the scarce students on the test set.** The figure shows the input image, bounding box, and ground truth (GT) mask in the three top rows, respectively. The baseline model, single output (SO) Student, and dual output (DO) Students corresponding outputs are shown in the three bottom rows, respectively.

https://doi.org/10.1371/journal.pone.0266147.g003

the scarce scenario, where the box guided teacher achieved almost double the DSC as the point guided one. The scarce box guided teacher also outperformed the vast point guided teacher. This suggests that training a teacher on a smaller set of bounding box annotated images can be advantageous compared to training a teacher on a large set of point guided images.

Surprisingly, the students did not perform better than the baseline in the scenario with vast strongly annotated data (see Table 3). Measured on the MSD-Lung dataset, the baseline model outperformed the two students, whereas the opposite was observed for the NSCLC-Radiomics dataset. A potential explanation might be that the Lung-PET-CT-Dx dataset contains tumors with sizes more similar to the NSCLC-Radiomics dataset than to the tumors in MSD-Lung. The introduction of the Lung-PET-CT-Dx dataset may have led the students to perform better on larger tumors, but may have degraded the results on smaller tumors typically found in MSD-Lung. Another explanation might be that the ratio between strong and weak labels were not large enough to make a noticeable difference. This was further demonstrated when the models were evaluated in the scarce scenario (see Table 5). In this scenario, the students significantly outperformed the baseline supervised model. This demonstrates that the introduction of suboptimal annotations into the teacher-student design can improve performance of an end-to-end segmentation model.

## State-of-the-art comparison

We observed a DSC comparable with state-of-the-art performance measured on the MSD-Lung dataset, with a F1-score of 85.18, and a DSC of 71.00, for one of our students. Isensee *et al*. [14] reported a DSC of 69.2 on the MSD-Lung dataset, whereas Carvalho *et al*. [11] reported a DSC of 70.9. Our model trained on only 40 human annotated images scored marginally better, although on a different test set. Other state-of-the-art results demonstrated better performance on the radiomics dataset. Pang *et al*. [42] reported a DSC of 77.67 whereas Kamal *et al*. reported a DSC of 72.28, and Hossain *et al*. [13] a DSC of 65.77, respectively.

However, all of these related work performed considerable data sanitation, which we did not, making the comparison unfair. Furthermore, our results suggest that 40 images is not enough to train a supervised model, but enough to train a semi-supervised model that can enhance a supervised model by increasing the available data in a cheaper way than manual delineation. This finding highlights the advantage of using a teacher-student design, such as ours, that can utilize datasets with poorer annotations. It is considerably faster to annotate tumors with bounding boxes than with semantic segmentation, but with negligible loss in performance. This finding suggests that it is advantageous to spend the time annotating more images with poorer supervision than to spend the same amount of time annotating fewer images with higher quality. The other highlighted papers performing lung tumor segmentation cannot take advantage of this effect as they rely solely on fully supervised training on high quality data.

## Data noise argument

The datasets were of varying quality. The MSD-Lung dataset was of high standard, whereas the NSCLC-Radiomics dataset was less so. Other publications that used the NSCLC-Radiomics dataset reported heavy data sanitation, effectively removing large parts of the dataset [13, 42, 43]. We did not override the expert's annotations, as we also seek to handle suboptimal annotations, if these should be present in a data set. The flawed dataset explains why the difference between the box guided and point guided teacher is larger on the NSCLC-Radiomics dataset than MSD-Lung. Images where the tumor is poorly, or even completely wrongly annotated, the box guided teacher can rely on the bounding boxes to achieve a good DSC, but since the annotation itself is wrong, the point guided teacher struggles.

## Limitations

One of the major limitations in this experiment was the scarce amount of data. The test set was sampled randomly from each dataset. It is plausible that a different sample of the test set would have given different result. Although K-fold cross validation could be used to eliminate this concern, it was dropped due to time limitations. K-fold cross validation is a time consuming strategy. It depends on training K different models, which would take a considerable amount of time, even with a small K in our situation. Since our method is a two step method that relies on two training steps, the K-fold cross validation would take nearly double the time of a similar single-step method as well.

Another limitations of this experiments was that the students were sensitive to voxel spacing. By reducing the voxel spacing during normalization/preprocessing, thus increasing the resolution of the image, the DSC did not improve, but actually degraded. Therefore, it is possible that the proposed architecture is sensitive to small adjustments in the preprocessing pipeline.

### Future work

An alternative to using 2D bounding boxes in the axial plane, is to use a 3D bounding box. As one 3D bounding box contains much fewer corners than multiple 2D bounding boxes, this could further reduce annotation load. It is reason to believe that a teacher trained on 3D boxes will perform worse than one trained on 2D axial boxes. However, if the reduction in annotation load is significant, the amount of data that can be annotated for the teacher might weigh up the loss in precision of the annotation. After all, this is the very fundamental idea behind the teacher-student design. However, we feel that a much larger dataset should be used to explore this properly.

The main motivation of using a teacher-student design is to improve models by learning from additional suboptimal annotated or unannotated data. We observed a benefit of using such a design for lung tumor segmentation in CTs. However, a single-step teacher might not be sufficient. It has been proposed to train both the teacher and student end-to-end in an iterative fashion [25]. This makes sense as the teacher could improve from the student's feedback. Especially from multiple students, which iteratively could improve the students as well, as the teacher become more experienced. However, for 3D applications this is likely infeasible. Alternatively, one could potentially use multiple teachers, trained on different types of images that focus on different lung tumor types and sizes. Having specialized teachers to train the student in an ensemble manner makes sense as it more closely represents the natual teacher-student relation from academia.

## Conclusion

We present the first known implementation of a mixed-supervised teacher-student framework for lung tumor segmentation from CT images. Our method utilized both semantic and axial bounding box annotations to maximize lung tumor segmentation performance. We demonstrated that with sufficient bounding box annotated data, our teacher-student framework achieved state-of-the-art performance, even with scarce semantic annotated data. In a scenario with only 40 semantic labeled images and ∼1000 bounding box labeled images, one of our models reached a mean DSC of 71.0 measured on nine images from the MSD dataset.

## Acknowledgments

This work was conducted at the machines Idun [44], Malvik (a cluster node owned and maintained by NTNU AI Lab), and Bohaga (a machine owned by NTNU, maintained by Gabriel Kiss).

## Author Contributions

**Conceptualization:** Frank Lindseth.

**Formal analysis:** Vemund Fredriksen, Svein Ole M. Sevle.

**Funding acquisition:** Thomas Langø, Frank Lindseth.

**Investigation:** Vemund Fredriksen, Svein Ole M. Sevle, André Pedersen.

**Methodology:** Vemund Fredriksen, Svein Ole M. Sevle, André Pedersen, Thomas Langø, Gabriel Kiss, Frank Lindseth.

**Project administration:** Thomas Langø, Frank Lindseth.

**Resources:** Gabriel Kiss, Frank Lindseth.

**Software:** Vemund Fredriksen, Svein Ole M. Sevle, André Pedersen.

**Supervision:** André Pedersen, Thomas Langø, Gabriel Kiss, Frank Lindseth.

**Validation:** Vemund Fredriksen, Svein Ole M. Sevle, André Pedersen, Thomas Langø, Gabriel Kiss, Frank Lindseth.

**Visualization:** Vemund Fredriksen, Svein Ole M. Sevle.

**Writing – original draft:** Vemund Fredriksen, Svein Ole M. Sevle, André Pedersen, Thomas Langø, Gabriel Kiss, Frank Lindseth.

**Writing – review & editing:** Vemund Fredriksen, Svein Ole M. Sevle, André Pedersen, Thomas Langø, Gabriel Kiss, Frank Lindseth.

## References

1. World Health Organization. WHO report on cancer: setting priorities, investing wisely and providing care for all. 2020;.
2. Sung H, Ferlay J, Siegel R, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians. 2021; 71. https://doi.org/10.3322/caac.21660
3. Uzelaltinbulat S, Ugur B. Lung tumor segmentation algorithm. Procedia Computer Science. 2017; 120:140–147. https://doi.org/10.1016/j.procs.2017.11.221
4. Adams R, Bischof L. Seeded Region Growing. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1994; 16(6):641–647. https://doi.org/10.1109/34.295913
5. Dehmeshki J, Amin H, Valdivieso M, Ye X. Segmentation of Pulmonary Nodules in Thoracic CT Scans: A Region Growing Approach. Medical Imaging, IEEE Transactions on. 2008; 27:467–480. https://doi.org/10.1109/TMI.2007.907555 PMID: 18390344
6. Gu Y, Kumar V, Hall LO, Goldgof DB, Li CY, Korn R, et al. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. Pattern Recognition. 2013; 46 (3):692–702. https://doi.org/10.1016/j.patcog.2012.10.005 PMID: 23459617
7. Ramanto KN, Parikesit AA. The usage of deep learning algorithm in medical diagnostic of breast cancer. Malaysian Journal Fundam Appl Sci. 2019; 15(2):274–281. https://doi.org/10.11113/mjfas.v15n2.1231
8. Kim M, Yun J, Cho Y, Shin K, Jang R, Bae Hj, et al. Deep learning in medical imaging. Neurospine. 2019; 16(4):657. https://doi.org/10.14245/ns.1938396.198 PMID: 31905454
9. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–241.
10. Du G, Cao X, Liang J, Chen X, Zhan Y. Medical Image Segmentation based on U-Net: A Review. Journal of Imaging Science and Technology. 2020; 64. https://doi.org/10.2352/J.ImagingSci.Technol.2020.64.2.020508
11. Carvalho J, Moreira J, Figueiredo M, Papanikolaou N. Automatic Detection and Segmentation of Lung Lesions using Deep Residual CNNs; 2019. p. 977–983.
12. Jiang J, Hu YC, Liu CJ, Halpenny D, Hellmann MD, Deasy JO, et al. Multiple Resolution Residually Connected Feature Streams for Automatic Lung Tumor Segmentation From CT Images. IEEE Transactions on Medical Imaging. 2019; 38(1):134–144. https://doi.org/10.1109/TMI.2018.2857800 PMID: 30040632
13. Hossain S, Najeeb S, Shahriyar A, Abdullah ZR, Ariful Haque M. A Pipeline for Lung Tumor Detection and Segmentation from CT Scans Using Dilated Convolutional Neural Networks. In: ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019. p. 1348–1352.
14. Isensee F, Jaeger P, Kohl S, Petersen J, Maier-Hein K. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 2021; 18:1–9. https://doi.org/10.1038/s41592-020-01008-z PMID: 33288961

15. Gan W, Wang H, Gu H, Duan Y, Shao Y, Chen H, et al. Automatic segmentation of lung tumors on CT images based on a 2D & 3D hybrid convolutional neural network. The British Journal of Radiology. 2021; 94:20210038. https://doi.org/10.1259/bjr.20210038 PMID: 34347535

16. Hansen S, Kuttner S, Kampffmeyer M, Markussen TV, Sundset R, Øen SK, et al. Unsupervised supervoxel-based lung tumor segmentation across patient scans in hybrid PET/MRI. Expert Systems with Applications. 2021; 167:114244. https://doi.org/10.1016/j.eswa.2020.114244

17. Fu X, Bi L, Kumar A, Fulham M, Kim J. Multimodal Spatial Attention Module for Targeting Multimodal PET-CT Lung Tumor Segmentation. IEEE Journal of Biomedical and Health Informatics. 2021; 25 (9):3507–3516. https://doi.org/10.1109/JBHI.2021.3059453 PMID: 33591922

18. Jiang J, Hu Y, Tyagi N, Zhang P, Rimner A, Deasy JO, et al. Cross-modality (CT-MRI) prior augmented deep learning for robust lung tumor segmentation from small MR datasets. Medical Physics. 2019; 46 (10):4392–4404. https://doi.org/10.1002/mp.13695 PMID: 31274206

19. Jue J, Jason H, Neelam T, Andreas R, Sean BL, Joseph DO, et al. Integrating Cross-modality Hallucinated MRI with CT to Aid Mediastinal Lung Tumor Segmentation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. Medical Image Computing and Computer Assisted Intervention—MICCAI 2019. Cham: Springer International Publishing; 2019. p. 221–229.

20. Wang D, Li M, Ben-Shlomo N, Corrales CE, Cheng Y, Zhang T, et al. Mixed-Supervised Dual-Network for Medical Image Segmentation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, et al., editors. Medical Image Computing and Computer Assisted Intervention—MICCAI 2019. Cham: Springer International Publishing; 2019. p. 192–200.

21. Mlynarski P, Delingette H, Criminisi A, Ayache N. Deep learning with mixed supervision for brain tumor segmentation. Journal of Medical Imaging. 2019; 6(3):1–13. https://doi.org/10.1117/1.JMI.6.3.034002 PMID: 31423456

22. Phuong M, Lampert C. Towards Understanding Knowledge Distillation. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research. PMLR; 2019. p. 5142–5151. Available from: https://proceedings.mlr.press/v97/phuong19a.html.

23. Furlanello T, Lipton Z, Tschannen M, Itti L, Anandkumar A. Born Again Neural Networks. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning. vol. 80 of Proceedings of Machine Learning Research. PMLR; 2018. p. 1607–1616. Available from: https://proceedings.mlr.press/v80/furlanello18a.html.

24. Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network; 2015.

25. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. Emerging Properties in Self-Supervised Vision Transformers. In: Proceedings of the International Conference on Computer Vision (ICCV); 2021.

26. Xie Q, Luong MT, Hovy E, Le QV. Self-training with Noisy Student improves ImageNet classification; 2020.

27. Li K, Wang S, Yu L, Heng PA. Dual-Teacher: Integrating Intra-domain and Inter-domain Teachers for Annotation-Efficient Cardiac Segmentation. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, et al., editors. Medical Image Computing and Computer Assisted Intervention—MICCAI 2020. Cham: Springer International Publishing; 2020. p. 418–427.

28. Sun L, Wu J, Ding X, Huang Y, Wang G, Yu Y. A Teacher-Student Framework for Semi-supervised Medical Image Segmentation From Mixed Supervision; 2020.

29. Zhang D, Chen B, Chong J, Li S. Weakly-Supervised teacher-Student network for liver tumor segmentation from non-enhanced images. Medical Image Analysis. 2021; 70:102005. https://doi.org/10.1016/j.media.2021.102005

30. Simpson AL, Antonelli M, Bakas S, Bilello M, Farahani K, van Ginneken B, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms; 2019.

31. Aerts H, Rios Velazquez E, Leijenaar R, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications. 2014; 5:4006. https://doi.org/10.1038/ncomms5006 PMID: 24892406

32. Aerts HJWL, Wee L, Rios Velazquez E, Leijenaar R, Parmar C, Grossmann P, et al. Data From NSCLC-Radiomics [Data set]. The Cancer Imaging Archive; 2019. https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics.

33. Li P, Wang S, Li T, Lu J, HuangFu Y, Wang D. A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis [Data set].; 2020. https://doi.org/10.7937/TCIA.2020.NNC2-0461

34. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. European Radiology Experimental. 2020; 4:50. https://doi.org/10.1186/s41747-020-00173-2 PMID: 32814998

**35.** Pedersen A, Smistad E, Rise TV, Dale VG, Pettersen HS, Nordmo TAS, et al. Hybrid guiding: A multiresolution refinement approach for semantic segmentation of gigapixel histopathological images; 2021.

**36.** Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. In: ICML; 2010. p. 807–814. Available from: https://icml.cc/Conferences/2010/papers/432.pdf.

**37.** He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015. p. 1026–1034.

**38.** Kingma DP, Ba J. Adam: A Method for Stochastic Optimization; 2017.

**39.** Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. Lecture Notes in Computer Science. 2017; p. 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

**40.** Bouget D, Pedersen A, Vanel J, Leira HO, Langø T. Mediastinal lymph nodes segmentation using 3D convolutional neural network ensembles and anatomical priors guiding; 2021.

**41.** MONAI Consortium. MONAI: Medical Open Network for AI; 2020. Available from: https://doi.org/10.5281/zenodo.5525502.

**42.** Pang S, Du A, He X, Díez J, Orgun MA. Fast and Accurate Lung Tumor Spotting and Segmentation for Boundary Delineation on CT Slices in a Coarse-to-Fine Framework. In: Gedeon T, Wong KW, Lee M, editors. Neural Information Processing. Cham: Springer International Publishing; 2019. p. 589–597.

**43.** Kamal U, Rafi AM, Hoque R, Wu J, Hasan MK. Lung Cancer Tumor Region Segmentation Using Recurrent 3D-DenseUNet. In: Petersen J, San José Estépar R, Schmidt-Richberg A, Gerard S, Lassen-Schmidt B, Jacobs C, et al., editors. Thoracic Image Analysis. Cham: Springer International Publishing; 2020. p. 36–47.

**44.** Själander M, Jahre M, Tufte G, Reissmann N. EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure; 2019.