UniToChest: A Lung Image Dataset for Segmentation of Cancerous Nodules on CT Scans

Hafiza Ayesha Hoor Chaudhry[1][0000−0001−5932−0103],
Riccardo Renzulli[1][0000−0003−0532−5966],
Daniele Perlo[2][0000−0001−6879−8475], Francesca Santinelli[3],
Stefano Tibaldi[3][0000−0003−0326−1399], Carmen Cristiano[3],
Marco Grosso[3][0000−0002−7083−09295], Giorgio Limerutti[3],
Attilio Fiandrotti[1][0000−0002−9991−6822],
Marco Grangetto[1][0000−0002−2709−7864], and Paolo Fonio[1][0000−0002−1106−2990]

[1] University of Turin
[2] Fondazione Ricerca Molinette Onlus
[3] Città della Salute e della Scienza di Torino

Abstract. Lung cancer has emerged as a major causes of death and early detection of lung nodules is the key towards early cancer diagnosis and treatment effectiveness assessment. Deep neural networks achieve outstanding results in tasks such as lung nodules detection, segmentation and classification, however their performance depends on the quality of the training images and on the training procedure. This paper introduces *UniToChest* , a dataset consisting Computed Tomography (CT) scans of 623 patients and 10071 lesions labelled for segmentation with nodules spanning different diameters ranges. Then, we propose a lung nodules segmentation scheme relying on a convolutional neural architecture that we also re-purpose for a nodule detection task. The experimental results show accurate segmentation of lung nodules across a wide diameter range and better detection accuracy over a traditional detection approach. The datasets and the code used in this paper are publicly made available as a baseline reference.

Keywords: Medical image segmentation · Deep Learning · U-Net · Dataset · Chest CT Scan · Lung nodules · DeepHealth

1 Introduction

Lung cancer has become the leading cause of death for men and women in 2021, surpassing breast and prostate cancer [19]. With such a low survival rate of 14-15% at late stages of lung cancer, detecting and monitoring malign cancerous nodules is the key towards better recovery rates [7]. Conventionally, first a thoracic *Computed Tomography* (CT) scan of the lungs generates high resolution images of the chest structures [16]. The same procedure is used to monitor the growth of lung nodules over time, as an indicator of the success of the treatment or as a warning in case of a sudden volume change [13].

Manual lung nodules analysis is time-consuming, so *Computer-Aided Diagnosis* (CAD) systems are commonly employed for the detection and segmentation tasks.Over the past decade, several systems based on traditional or deep learning based image processing techniques have been proposed for the detection and segmentation of lung nodules [8, 21, 10]. Differences in size and shape of the nodules, age and gender of the patients, imaging device model and brand along with the similarity between nodules and their surrounding, make this a challenging task. Most of the methods rely on supervised approaches , so an important factor towards precise segmentation and detection is the training dataset quality. Images and relative annotations in fact often lack in terms of quality or quantity or both, due to the cost of acquiring and annotating the images by a radiologist. While some methods attempt to cope with small sample sizes [22] or noisy labels [6], a good training set remains of paramount importance. Finally, releasing annotated medical images to the public requires abiding by the privacy protection laws, which includes making sure that neither the images nor the annotations leak any sensitive information.

This paper present a twofold contribution towards accurate lung nodule detection and segmentation.
First, we present UniToChest, a dataset collected and annotated by Radiology Unit in Città della Salute e della Scienze Hospital within the framework of the EU-H2020 *DeepHealth* project [4]. The dataset [15] includes 306440 lung cancer screening thoracic computed tomography (CT) scans of 623 patients. Each patient file contains diagnostic lung cancer CT scan images and associated segmentation masks for the annotated lesions. This dataset is the largest of its kind with most diversity in lesions (lung nodule) size.

Second, we propose a complete nodule detection and segmentation pipeline designed around a convolutional neural network. Namely, we first segment nodules using an autoencoder with skip connection that we train in a fully supervised way. Next, we recast the nodule detection task as a segmentation problem, showing better performance than a baseline nodule detector.

2   Background and Related Works

In this section we first provide the medical background relevant to the understanding of this work, next we review existing techniques in pulmonary nodules detection highlighting the main limitations that prompted this research.

Computed Tomography (CT) scan is a medical imaging procedure that uses a computer linked to an x-ray machine to grab series of pictures of areas of the inner body. The pictures are taken from different angles and are used to create 3-dimensional views of tissues and organs. Sometimes to increase the chances of seeing diseases, a drug called contrast medium is used which is injected into the venous circulation to make the blood vessels opaque and reveal neoplastic lesions. Pulmonary nodules are small, focal, radiographic opacities that may be

---

[4] https://deephealth-project.eu/

solitary or multiple. A classic solitary pulmonary nodule (SPN) [4] is a single, spherical, well-circumscribed, radiographic opacity measuring less than or equal to 30 mm in diameter and is surrounded completely by aerated lung. The SPN is a coined term that in the past described solitary nodules detected incidentally by chest radiography (CXR). Today, most nodules are detected by computed tomography (CT). The detailed CT images frequently identify more than one nodule, or enlarged lymph nodes. Indeterminate nodules are those that do not possess features clearly associated with a benign etiology, such as a benign pattern of calcification or stability on imaging for > 2 years. On CT scans, a nodule appears as a rounded or irregular opacity, well or poorly defined, measuring up to 3 cm in diameter. Advances in chest imaging and the increased use of CT as a diagnostic modality have lead to incidental identification of many small pulmonary nodules. The vast majority of nodules detected on CT are sub-centimeter based on early lung screening trials (61%-89%). The overwhelming majority of these are benign. The prevalence of pulmonary nodules changes significantly across studies. This variation stems from the inconsistency among studies in method, enrolled population, and reporting results. Most lung nodules are detected incidentally on CXR or CT scans obtained for other purposes. The actual risk for malignancy in sub-centimeter nodules is lower than the predicted risk based on clinical and radiographic criteria for pulmonary nodules [11]. The risk factor varies with the nodule diameter, staying under 35% for nodules below to 1 cm in diameter and exceeding 97% for nodules above 3 cm. For this reason, accurate nodule segmentation is of paramount importance to estimate its malignancy probability. Methods for the detection and segmentation of lung nodules can be categorized into traditional and learning-based.

Traditional methods rely on handcrafted feature extraction [2], often coupled with shallow classifiers or regressors. The main problem with such techniques is the manually designing feature extractors time consuming activity and features may be tailored to some specific dataset.

Learning-based methods usually rely on a type of artificial neural networks known as *Convolutional Neural Networks(CNN)* [9]. The underlying idea is to let the convolutional layers learn feature extractors that maximize some loss metric on an annotated dataset rather than handcrafting the feature extractor. Such architectures include millions of learnable parameters and represent the state of the art in a number of medical applications today [20]. In particular, the U-Net [17] architecture is designed around an autoencoder topology with skip connections and represents a standard for semantic segmentation tasks. Due to the amount of learnable parameters they include, their performance strongly depends on the amount of data available for training, prompting the collection of large annotated datasets. The *Lung Image Database Consortium image collection (LIDC-IDRI)* dataset [1] is the largest publicly available dataset for the detection and segmentation of lung nodules. *LUNA16* (Lung Nodule Analysis 2016) [18] is a segmentation challenge that uses a subset of *LIDC-IDRI*. The *LIDC-IDRI* dataset contains 7371 nodules annotated by atleast 3 out of 4 radiologists performing the study. For nodules greater than $3mm$ the com-

plete volumetric nodule boundary is given as Region of Interests (ROIs) [12]. Whereas for nodules less than $3mm$ only the centroid point(x,y,z) is provided as ROI instead of whole nodule boundary, which makes detecting the smaller nodules harder.

3   The UniToChest Dataset

The *UniToChest* dataset has been collected within the EU-H2020 *DeepHealth* [3, 14] project and consists of about than 300k lung CT scans of pulmonary lungs from 623 different patients. The scans are in DICOM format and each scan comes with a manually annotated segmentation mask in black and white PNG format, both being 512×512 in size. The slice thickness of CT scans ranges from 1.25 to 6.5 mm and the pixel spacing from 0.41 to 0.97 mm. A comparison with similar datasets in Tab. 3 shows that our dataset has more nodules and from a wider diameter range especially at the top end. In fact, for most patients are available images collected over multiple exams over the years including late stages as shows in Fig. 1a. The dataset contains data collected from a gender-balanced population (377 males and 246 females) and spanning across a wide rage of ages (from 7 to 9, most of the population being between 60 and 80), as in Fig. 1b For many comparable datasets, the images come from a single acquisition device that may hide some specific bias; conversely, our dataset includes images acquired using 10 different devices as in Figure 1c. For each and every image, the radiologist inspected the image for nodules and, where found, each nodule was manually segmented across multiple slices. Finally, compliance with the UE regulation on privacy is guaranteed since any sensitive information (name, birth date, identity) was carefully removed from images and annotations.

Splits	Number of Patients	Male	Female	Average Age
Training	498	303	195	66
Validation	62	39	23	68
Test	63	35	28	65
Total	623	377	246	66

Table 1: Dataset population for the three splits we provide

The total number of nodules in the malignant CT scans of our dataset surpasses any publicly available dataset. The distribution of overall nodule diameter in our dataset is represented in Figure 1a, and a detail description of nodule diameter with respect to splits made in our experiment can be seen in Table 2.

For the purpose of training a neural network, we split the dataset into training, validation and test set randomly as 80-10-10 of patients. We maintain data consistency across multiple splits by assigning a single split to each patient. The data population with respect to the splits is summarized in Table 1 All the three sets (train, validate, and test) have a 60 to 40 ratio between the number of male and female patients.

(a) Nodule Diameter Distribution

(b) Age distribution wrt. Gender

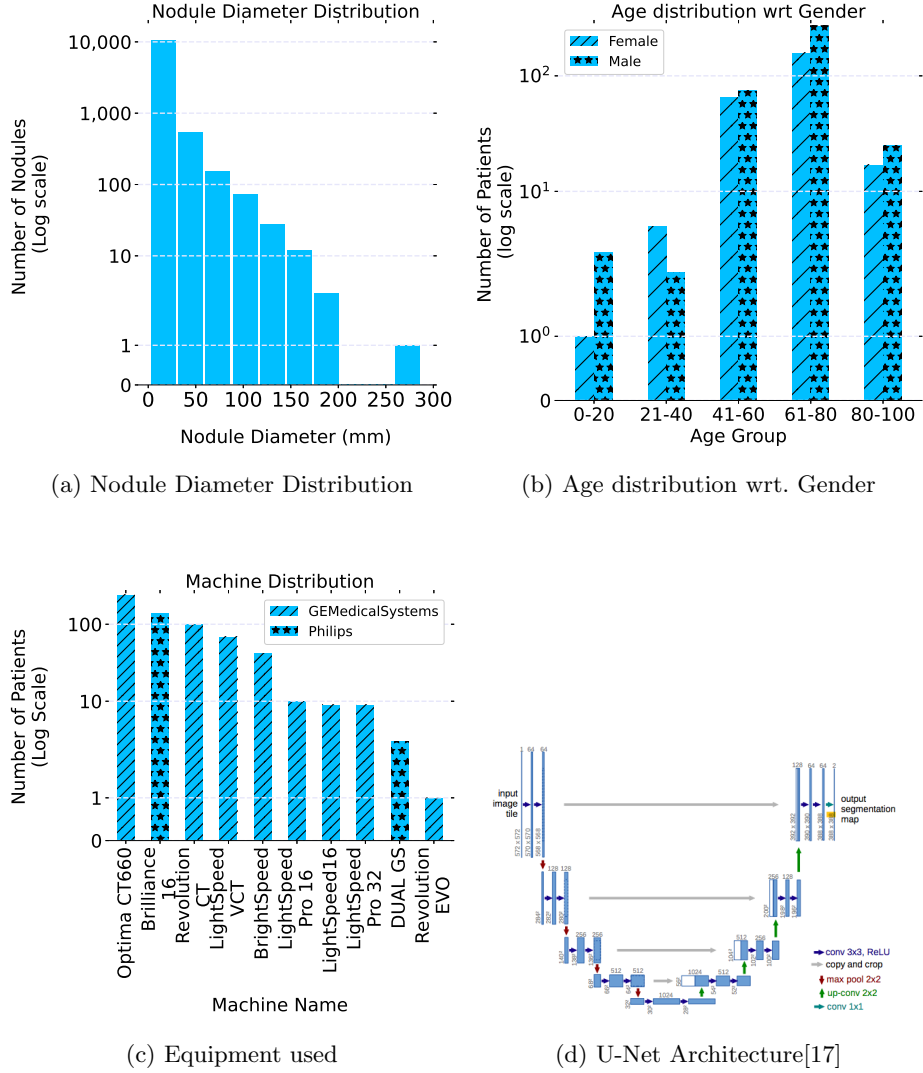(c) Equipment used

(d) U-Net Architecture[17]

Fig. 1: Dataset distributions show no sign of bias with respect to nodule diameter range, patient gender and machines used. The U-Net architecture is also displayed for reference

	$< 3mm$	$< 10mm$	$< 30mm$	$> 30mm$	Total
Train	149	6527	1861	249	8786
Validation	7	315	116	23	461
Test	21	575	195	33	824
Total	177	7417	2172	305	10071

Table 2: Nodule diameter distribution across three splits

Dataset	Number of Patients	Number of Scans	Total Nodules count	Nodule Diameter range(mm)
$LIDC - IDRI$	1010	244527	7371	$2 - 69$
$LUNA16$	1010	888	1836	$3 - 33$
$UniToChest$	623	306440	10071	$1 - 136$

Table 3: Comparison with similar public dataset shows that our dataset has more clinical lung cancer CT scan slices and annotated lung nodule count with a diverse diameter range

4   Methodology

This section describes the proposed method for pulmonary nodules segmentation, including the preprocessing stage, the architecture of the deep neural convolutional architecture we rely upon and the relative training procedures.

4.1   Data Preprocessing

DICOM files produced by CT machines tipically contain pixel intensity values in Hounsfield Units (HU) , i.e. they indicate radiometric density per pixel (low values indicating air, higher values indicating bones). Following a standard medical practice, a clipped windowing transformation function is applied to such desity values. The window width and center indicate the range of the Hounsfield Units covered inside the converted pixel values, everything outside this range will be equivalent to either zero or one. According to standard practice, we have used a window width of 1600 and a window center of -500 to account for the radiometric density of body structures actually useful for nodule detection.

4.2   Network Architecture

Our approach relies on the U-Net implementation [17] in Figure 1d. The encoder consists of 5 convolutional layers with max-pooling for featuremap downsampling. As in other convolutional architectures, as the size of the featuremaps

shrinks the number of featuremaps increases by a two factor. The decoder includes 5 convolutional layers followed by an upconvolutions, where the size of the featuremaps increases while their number decreases at each layer. A number of encoder and decoder layers are matched with skip connections, where the feature maps generated by the respective encoder layer is concatenated with the output of decoder layer, enabling the precise learning and localization of image object by allowing different tradeoffs between semantic level and spatial accuracy of the featuremaps.

4.3   Training Procedure

The training method is fully supervised and consists in randomly initializing the network weights (*from scratch*) and then training the network for nodule segmentation minimizing the loss between the network output and the segmentation mask relative to the input image. As for similar segmentation tasks, we minimize the *DICE* loss since it has a derivative allowing for error gradient backpropagation and minimizing the dice loss amount to maximizing the IoU between predicted and ground truth mask. As a preliminary stage, we found beneficial pretraining the network over the LIDC dataset prior to the training on UniToChest. The rationale behind this pretraining is to have the network learning additional features from the LIDC dataset that may be possibly useful when trained for segmentation on UniToChest. Next, the network is trained over UniToChest train set until the *Intersection over Union* (IoU) score as measured over UniToChest validation set did not improved for 50 epochs. For this training, only scans with one or more nodules have been considered, since other scans we experimentally verified do not bring any useful information for segmentation. The CT slices are provided in input to the network in batches of 5, as that enabled a reasonable tradeoff between memory footprint and performance. We found beneficial resorting to on the fly data augmentation during the training to avoid overfitting to the training data. The augmentation technique we used consist in cropping random patches from the slices and performing random flips and rotations (the very same transformations are also applied to the corresponding segmentation mask). The optimizer used in our experiment is Adam with an initial learning rate of 0.001 and weight decay of 0.0001. The whole architecture has been implemented in PyTorch and is available on github.[5]

5   Results and Discussion

In this section, we first experiment over the UniToChest dataset with the neural network based method described in the previous section for nodule segmentation. Next, we repurpose and the same method for nodules detection, with particular attention to the tradeoff between sensitivity and specificity. All results are relative to UniToChest test set, i.e. images that the network has never seen at training time.

---

[5] The git repository address has been removed not to disclose the authors' identity, and will be revealed after acceptance

5.1   Nodules Segmentation

(a) Nodule diameter = 7.8mm IoU = 0.90

(b) Nodule diam. = 35-77mm IoU = 0.91

(c) Nodule diameter = 37mm IoU = 0.93
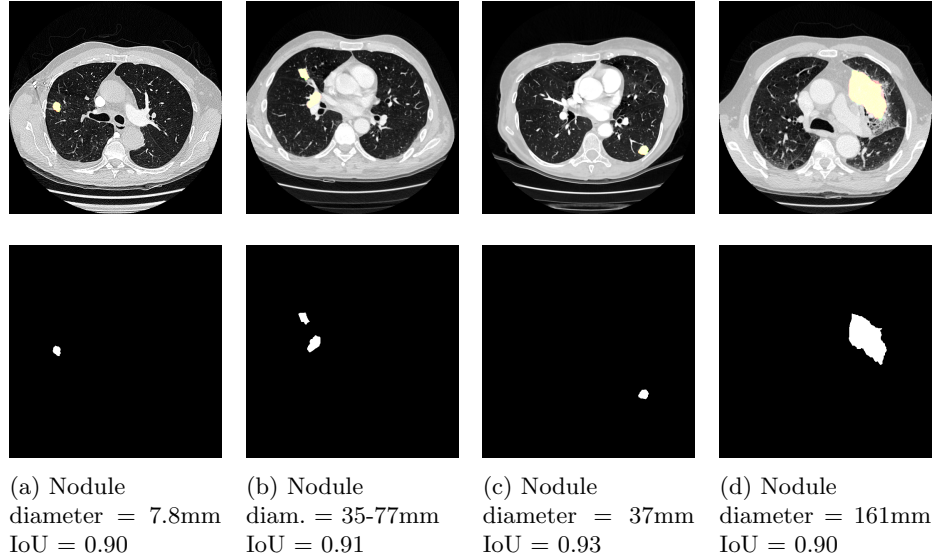
(d) Nodule diameter = 161mm IoU = 0.90

Fig. 2: Segmentation results over different nodule diameter range; regions in yellow represent the overlap between predicted and annotated segmentation masks.

As a first experiment, we evaluate the performance of the U-Net trained as in the previous section at nodule segmentation. For this task, we consider only positive slices from the test set, i.e. slices with at least one nodule. Figure 3 shows the *Intersection over Union* (IoU) and DICE scores for both the network pretrained as in the previous section and a reference network that was trained from scratch. The pretraining improves both segmentation accuracy (about +10% IoU) and convergence speed. As further experiment, we tested the network pretrained on LIDC only over UniToChest dataset, the top IoU settling at about 43% as a proof of the benefit yield by pretraining.

Table 4 correlates the average IoU and DICE scores with the nodule diameter size for both the cases without and with pretraining. The number of nodules in the sub 10mm bin is approximately equal to the number of nodules in the above 10mm bin. The average IoU is as large as 61% and even on the nodules having a diameter of less than 3 mm, we achieve an average IoU of 59%. We hypothesize that the above 10mm bin benefits the most from the pretraining because LIDC contains nodules ranging mainly in the 10mm to 50mm range.

Finally, Figure 2 shows some samples of the segmentation mask predicted by the network (bottom row) for some sample test images (top row). Red pixels represent false negatives, green pixel false positives and yellow pixels correctly

segmented pixels: most of the pixels are correctly segmented, a few errors only remaining at the borders of the nodule.
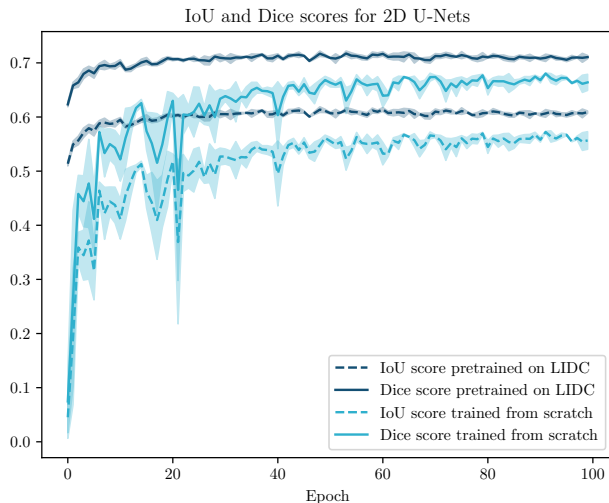

Fig. 3: Segmentation scores for the first 100 epochs: pretraining improves asymptotic IoU, plus the network converges faster.

Nodule Diameter	From scratch (not pretrained)		Pretrained on LIDC	
	IoU	DICE	IoU	DICE
$< 10mm$	0.575	0.69	0.59	0.69
$> 10mm$	0.58	0.69	0.615	0.72

Table 4: Segmentation accuracy: pretraining improves accuracy, especially for nodule sizes most represented in the dataset used for pretraining

5.2 Detection

Next, we evaluate the performance of the same U-Net network trained for segmentation on a nodule detection task, this time considering both positive and negative test scans. We provide each scans in input to the network and we count the number of white pixels in the predicted segmentation mask. If such number is greater than zero, the slice is labeled as positive, negative otherwise. Figure 4a (left) shows that the network achieves a sensitivity of 0.95 and specificity of 0.80.

We investigate whether the specificity value could be increased further, finding a balance between sensitivity and specificity, since in many medical trials the aim is also to reduce the number of false positives. For this reason, we finetune the network adding to the train set 10% of the negative training samples drawn at random. The confusion matrix on the right shows that the specificity improves from 0.8 to 0.95,reducing the number of false positives.

As a baseline reference, we also trained a binary classifier based on a ResNet18 [5] pretrained over ImageNet to discriminate each slice as positive or negative. We achieved a sensitivity of 0.74 and specificity of 0.82, so the number of false positives is higher than our segmentation-based detection method.

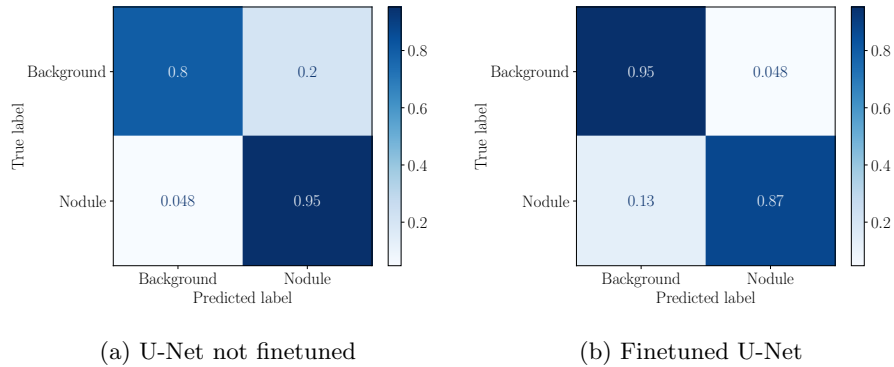(a) U-Net not finetuned   (b) Finetuned U-Net

Fig. 4: Normalized confusion matrices at nodule detection for a U-Net network trained at segmentation. The network finetuned on negative slices has better tradeoff between sensitivity and specificity

6   Conclusion and Future Works

This paper presented UniToChest, a CT scan lung nodules dataset, that is among the largest of its kind and boasts a diversity of patient ages, acquisition machines and nodules diameter. We proposed a U-Net based architecture that yield promising results at both detection and segmentation of lung nodules. Future research directions of this work include exploiting the thee-dimensional information of nodules across neighboring slices.

Acknowledgement

References

1. Armato, S.G., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. Medical Physics **38**(2), 915–931 (Jan 2011). https://doi.org/10.1118/1.3528204, `https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI`

2. Chaudhary, A.H., et al.: Blood cell counting and segmentation using image processing techniques. In: Applications of Intelligent Technologies in Healthcare, pp. 87–98. Springer (2019). https://doi.org/10.1007/978-3-319-96139-2_9

3. DeepHealth: Deep-learning and hpc to boost biomedical applications for health (2019), `https://deephealth-project.eu/`

4. Gaga, M., et al.: Lung nodules: A comprehensive review on current approach and management. Annals of Thoracic Medicine **14**(4), 226 (2019). https://doi.org/10.4103/atm.atm_110_19

5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

6. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis **65**, 101759 (2020). https://doi.org/10.1016/j.media.2020.101759

7. Knight, S.B., Crosbie, P.A., Balata, H., Chudziak, J., Hussell, T., Dive, C.: Progress and prospects of early detection in lung cancer. Open Biology **7**(9), 170070 (Sep 2017). https://doi.org/10.1098/rsob.170070

8. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

10. Liu, H., et al.: A cascaded dual-pathway residual network for lung nodule segmentation in ct images. Physica Medica **63**, 112–121 (2019). https://doi.org/10.1016/j.ejmp.2019.06.003

11. MacMahon, H., et al.: Guidelines for management of incidental pulmonary nodules detected on CT images: From the fleischner society 2017. Radiology **284**(1), 228–243 (Jul 2017). https://doi.org/10.1148/radiol.2017161659

12. McNitt-Gray, M.F., Armato III, S.G., Meyer, C.R., Reeves, A.P., McLennan, G., Pais, R.C., Freymann, J., Brown, M.S., Engelmann, R.M., Bland, P.H., et al.: The lung image database consortium (lidc) data collection process for nodule detection and annotation. Academic radiology **14**(12), 1464–1474 (2007)

13. Mozley, P.D., Bendtsen, C., Zhao, B., Schwartz, L.H., Thorn, M., Rong, Y., Zhang, L., Perrone, A., Korn, R., Buckler, A.J.: Measurement of tumor volumes improves RECIST-based response assessments in advanced lung cancer. Translational Oncology **5**(1), 19–25 (Feb 2012). https://doi.org/10.1593/tlo.11232

14. Oniga, D., et al.: Applications of AI and HPC in health domain. In: HPC, Big Data, AI Convergence Toward Exascale: Challenge and Vision, chap. 11. CRC Press, Taylor & Francis Group (2021), ISBN: 9781032009841

15. Perlo, D., Renzulli, R., Santinelli, F., Tibaldi, S., Cristiano, C., Grosso, M., Limerutti, G., Grangetto, M., Fonio, P.: UniToChest (2022). https://doi.org/10.5281/zenodo.5797912

16. Puderbach, M., Kauczor, H.U.: Can lung MR replace lung CT? Pediatric Radiology **38**(S3), 439–451 (May 2008). https://doi.org/10.1007/s00247-008-0844-7

17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28

18. Setio, A.A.A., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. Medical Image Analysis **42**, 1–13 (2017). https://doi.org/10.1016/j.media.2017.06.015, `https://luna16.grand-challenge.org/`

19. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2021. CA: A Cancer Journal for Clinicians **71**(1), 7–33 (2021). https://doi.org/10.3322/caac.21654

20. Ulku, I., Akagunduz, E.: A survey on deep learning-based architectures for semantic segmentation on 2d images. arXiv preprint arXiv:1912.10230 (2019)

21. Wu, J., Qian, T.: A survey of pulmonary nodule detection, segmentation and classification in computed tomography with deep learning techniques. J. Med. Artif. Intell **2**(8), 1–12 (2019). https://doi.org/10.21037/jmai.2019.04.01

22. Zhang, P., Zhong, Y., Deng, Y., Tang, X., Li, X.: A survey on deep learning of small sample in biomedical image analysis. arXiv preprint arXiv:1908.00473 (2019)