# Residual Networks for Pulmonary Nodule Segmentation and Texture Characterization

Adrian Galdran(✉) and Hamid Bouchachia(✉)

Department of Computing and Informatics, Bournemouth University, Poole, UK
{agaldran,abouchachia}@bournemouth.ac.uk

**Abstract.** The automated analysis of Computed Tomography scans of the lung holds great potential to enhance current clinical workflows for the screening of lung cancer. Among the tasks of interest in such analysis this paper is concerned with the segmentation of lung nodules and their characterization in terms of texture. This paper describes our solution for these two problems in the context of the LNdB challenge, held jointly with ICIAR 2020. We propose a) the optimization of a standard 2D Residual Network, but with a regularization technique adapted for the particular problem of texture classification, and b) a 3D U-Net architecture endowed with residual connections within each block and also connecting the downsampling and the upsampling paths. Cross-validation results indicate that our approach is specially effective for the task of texture classification. In the test set withheld by the organization, the presented method ranked 4th in texture classification and 3rd in the nodule segmentation tasks. Code to reproduce our results is made available at http://www.github.com/agaldran/lndb.

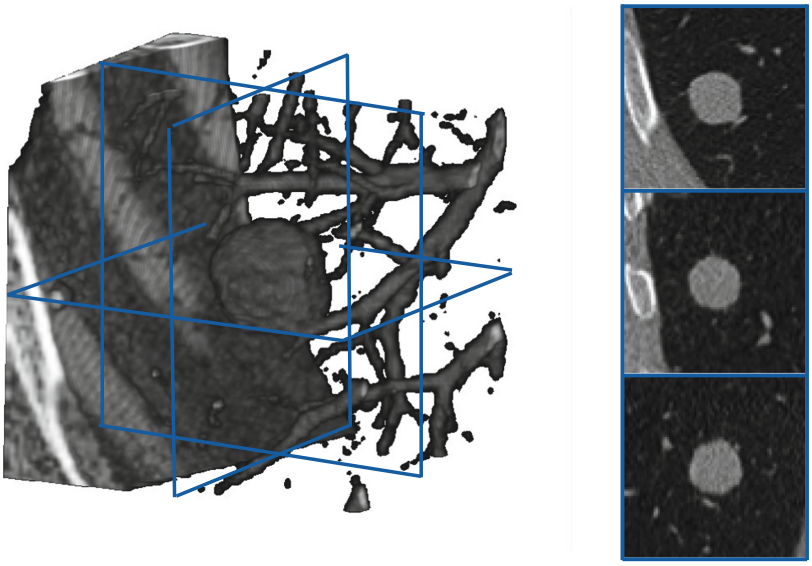
**Keywords:** Lung nodule segmentation · Texture classification · Imbalanced classification · Label smoothing

## 1 Introduction

Pulmonary cancer is known to be among the most lethal types of cancer worldwide [14]. Early detection of lung cancer may have a great impact in mortality rates, and Computed Tomography (CT) is recognized as a promising screening test for this purpose [1]. Large-scale screening programs are susceptible of becoming more effective and efficient by the deployment of Computer-Aided Diagnosis (CAD) tools. Such tools might bring standardization to a problem that suffers from great interobserver variability, and they also have the potential of reducing the workload of specialists by assisting them with complementary decisions.

The main task related with CAD in the processing of pulmonary CT scans is the automated analysis of lung nodules. This comprises several sub-tasks, namely lung nodule detection (localization of lesions within the scan), nodule segmentation (delineation of lesion borders), and lung nodule characterization (classification of each nodule into different categories, *e.g.* malignancy or texture). This


A. Campilho et al. (Eds.): ICIAR 2020, LNCS 12132, pp. 396–405, 2020.
https://doi.org/10.1007/978-3-030-50516-5_35



(a) (b)

**Fig. 1.** (a) 3-D visualization of a lung nodule from the LNdB dataset (b) Three 2-D orthogonal views of (a), used here to train a model for texture classification.

array of problems has attracted considerable attention from the medical image analysis community in the last years [16]. In particular since the advent of Deep Learning techniques, a wide range of approaches based on Convolutional Neural Networks has been proposed for lung nodule detection [3], segmentation [4], characterization [5,7], or direct end-to-end screening [2] with remarkable success.

This paper describes a solution to the LNdB challenge, held jointly with ICIAR 2020. This challenge is built around the release of a new database of lung CT scans (termed itself LNdB), accompanied with manual ground-truth related to lung nodule localization, segmentation, texture categorization, and follow-up recommendation based on 2017 Fleischner society guidelines [12]. A sample of one of the nodules from the LNdB database is shown in Fig. 1.

Our solution is solely concerned with the tasks of lung nodule segmentation and texture characterization. In the remaining of this paper, we describe our approach to each of these tasks, which is based on the effective training of two residual networks. For the texture categorization scenario, we adopt a regularization scheme based on a custom manipulation of manual labels that better optimizes the $\kappa$ score in this kind of problems. For the nodule segmentation problem, we construct a modified UNet architecture by adding residual connections inside each of its blocks, and also from the downsampling path to the upsampling path. The reported cross-validation results are promising, specially in the texture classification task, where our approach seems to be able to successfully overcome the difficulties associated to a highly imbalanced dataset.

## 2 Lung Nodule Texture Characterization

In the context of the LNdB challenge, sub-challenge C corresponded to the classification of lung nodule's texture into three distinct categories, namely Solid, Sub-Solid, and Ground-Glass Opaque (GGO).

The solution proposed in this paper was based on four main components: 1) input pre-processing, 2) a standard Residual Neural Network, 2) a specialized label smoothing technique, and 3) application of oversampling on the minority classes.

### 2.1 Input Pre-processing

Initially, for each provided nodule center a cubic volume of size $64 \times 64 \times 64$ was extracted and stored separately to facilitate model training. In addition, instead of attempting to process the input data by means of three-dimensional convolutions, we simplified the input volumes by first extracting three orthogonal planes of dimension $64 \times 64$ centered around each nodule, and then stacking them into a single 3-channel image. This turned the inputs into tensors amenable to standard 2D-Convolutional Neural Networks, and resulted in a 95% dimensionality reduction. A representation of this process is displayed in Fig. 1.

### 2.2 Convolutional Neural Network

We experimented with Residual Networks of different depths (18-layers, 50-layers, and 101-layers depth networks). Several modifications were also tested, namely the size of the filters in the very first layer was reduced from $7 \times 7$ to $3 \times 3$, and an initial Batch-Normalization/Instance Normalization layer was inserted in each architecture. In addition, initialization with weights pre-trained on the ImageNet database was also tested. Eventually, our best configuration based on cross-validation analysis was a 50-layer Residual Network, trained from random weights and with no initial normalization layer.

### 2.3 Gaussian Label Smoothing

A simple but powerful approach to regularization in CNNs consists of performing Label Smoothing (LS) [15]. LS is often applied for multi-class classification tasks, where the Cross-Entropy loss function is employed, and annotations are available in the form of one-hot encoding. The idea of LS consists of replacing these original one-hot encoded labels by a smoothed version of them, where part of their value is redistributed uniformly among the rest of the categories. LS has been proven useful to prevent neural networks from becoming over-confident and avoid overfitting in a wide range of problems [11].

Recently, a modified version of LS has been introduced in the context of Diabetic Retinopathy Grading from retinal fundus images [8], termed Gaussian Label Smoothing (GLS). The main assumption of GLS is that in a scenario

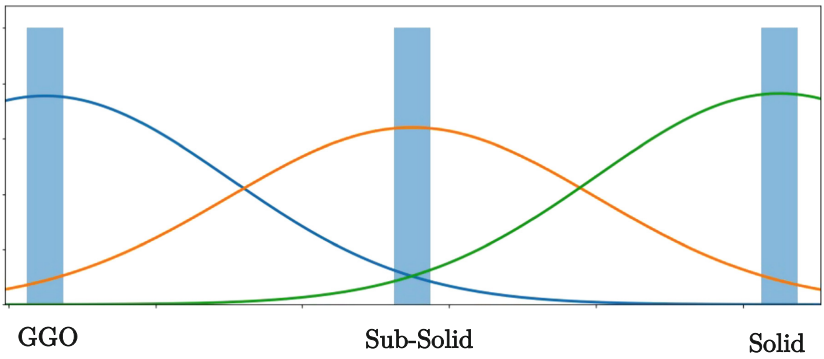



**Fig. 2.** Gaussian Label Smoothing technique (GLS) applied to the texture characterization problem. Light-blue bars represent the original one-hot encoded labels for each category, whereas the mean of the Gaussian curves represented in blue, orange, and green represent the corresponding smoothed labels. (Color figure online)

where labels are not independent, but reflect some underlying "ordering", it can be better to replace the uniform smoothing process in LS by a weighted smoothing, where neighboring categories receive more weight than further away ones. It was shown in [8] that, for ordinal classification of Diabetic Retinopathy grades, GLS outperformed standard LS.

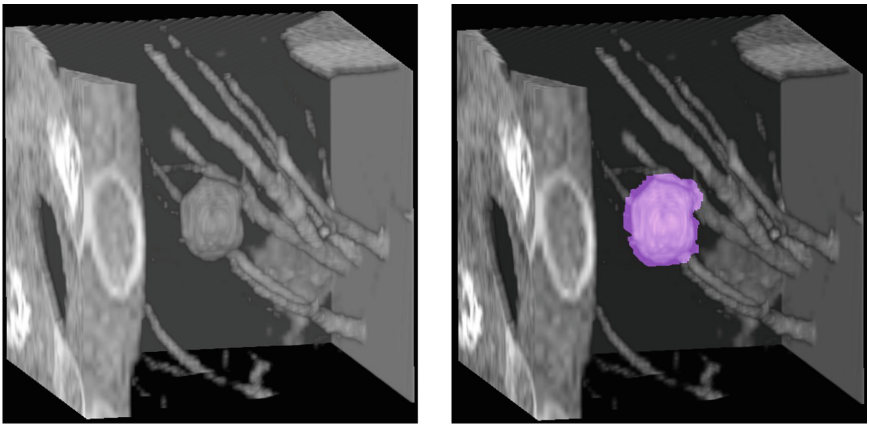
For the problem of texture classification, we use the standard cross-entropy loss and we adapt the GLS technique to three classes:

$$\mathcal{L}(y, \text{gls}(\hat{y})) = \sum_{k=1}^{3} y^{(k)} \log(\text{gls}(\hat{y}^{(k)})), \tag{1}$$

where $y$ is the output of the CNN followed by a soft-max mapping, $\hat{y}$ is the original one-hot encoded label, and $\text{gls}(\hat{y}) = G \circ \hat{y}$ is the transformation of $\hat{y}$ by a GLS mapping. As an example, a nodule belonging to the solid category is no longer represented by a one-hot encoded vector of the form $(1, 0, 0)$ but rather by a vector close to $(0.80, 0.18, 0.02)$, as shown in Fig. 2. As opposed to this strategy, the LS technique would encode the label as a vector of the form $(0.80, 0.10, 0.10)$. The effect of GLS is to induce a larger penalization when the prediction of the network is far away from the true class, promoting decisions closer to the actual label. This is particularly useful for the LNdB challenge, where the evaluation metric is the Quadratic-Weighted Kappa score.

### 2.4 Minority Class Oversampling

Given the relatively low amount of examples and high ratios of class imbalance (a proportion of approximately 5%/7%/88% for each class respectively), special care was taken when considering the sampling of the training set during model training. Our experiments revealed that the optimal strategy in this setting was



(a) (b)

**Fig. 3.** (a) Three-dimensional visualization of a lung nodule from the training set, and (b) same nodule as in (a) with an overlayed manual segmentation.

to perform oversampling on the two minority classes, which is consistent with previous works [6]. Solid nodules were oversampled by a factor of 9 and sub-solid by a factor of 6, which resulted in a class ratio of 25%/25%/50% during training. It must be noticed that this approach lends itself to easily overfitting minority examples, which are shown to the model much more frequently. However, we observed that the application of Test-Augmentation Techniques mitigated this effect considerably, as explained in Sect. 4.

## 3 Lung Nodule Segmentation

Sub-challenge C corresponded to the task of segmenting lung nodules from a CT scan, given the location of their centroids. An example of a nodule and the associated manual segmentation is shown in Fig. 3. For this task, we implemented a standard U-Net taking as input three-dimensional volumes of $80 \times 80 \times 80$ resolution. We introduced several modifications to the architecture presented in [13]: 1) All 2-dimensional learnable filters were replaced by 3 dimensional filters 2) Batch-normalization layers were inserted in between every convolutional block, and also prior to the first layer in the architecture, 3) Skip connections were added to every convolutional blocks, and 4) Convolutional layers connecting the downsampling path in the architecture with the upsampling path were also added. Note that some of these modifications have been explored in previous works developing enhancements of the standard U-Net architecture [18]. A representation of the resulting architecture is provided in Fig. 4.

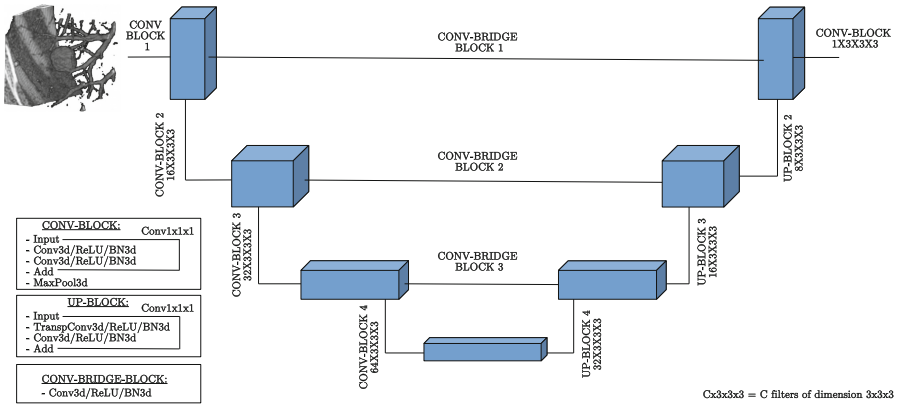



**Fig. 4.** A description of the 3d-unet used for this project.

The loss function we minimized in this case was the channel-wise Dice soft function, as suggested in [10]:

$$\mathcal{L}(y, \hat{y}) = \frac{2 \sum_{i=1}^{n} y_i \cdot \hat{y}_i}{\sum_{i=1}^{n} y_i^2 + \hat{y}_i^2 + \epsilon,} \tag{2}$$

where $y_i$ is the output of the CNN at each location $i$, $\hat{y}_i$ is the binary label associated to each voxel, and $\epsilon$ is a small constant to prevent division by zero.

## 4 Training Details

The training of the CNN, both for texture classification and nodule segmentation, followed similar stages. In both cases, an initial learning rate of 0.01 was set, and the weights were updated by means of the Adam optimizer so that the corresponding loss was minimized. In the nodule segmentation problem, the optimizer was wrapped by the look-ahead algorithm [17], since severe instabilities were observed when using the original Adam optimizer. Regarding batch sizes, for the texture classification task a batch size of 8 samples was applied, whereas for the segmentation problem a reduced batch size of 4 had to be employed due to computational constraints Both models were trained for 500 epochs, but training was stopped after no performance improvement was observed in the validation set during 25 epochs. In addition, after no improvement in 15 epochs, the learning rate was decreased by a factor of 10.

As for data augmentation techniques, we performed random reflection along each of the three axis of a given volume, as well as random small offsets, scalings and rotation. It is important to note that, even if the texture classification model was trained on two-dimensional images with three intensity channels (the three orthogonal views depicted in Fig. 1), in this case we also performed data augmentation on three dimensional volumes before sampling the three planes

**Table 1.** Cross-validation performance analysis on nodule texture classification.

| | Vl. Fold 1 | Vl. Fold 2 | Vl. Fold 3 | Vl. Fold 4 | Avg $\pm$ Std |
|---|---|---|---|---|---|
| Quad. kappa score | 0.542 | 0.607 | 0.475 | 0.617 | $0.560 \pm 0.057$ |
| Balanced accuracy | 0.518 | 0.621 | 0.449 | 0.543 | $0.533 \pm 0.061$ |
| Mean AUC | 0.849 | 0.885 | 0.852 | 0.803 | $0.847 \pm 0.029$ |

of the input image. It is also worth noting that for the texture classification scenario, the metric that dictated if there was an improvement in the validation set was initially set to be the quadratic kappa score between predictions and actual labels. However, we observed the kappa score to be too noisy during the training process. For this reason, we replaced it by a metric aggregating the average Area Under the Curve for three classes and the kappa score itself. For the segmentation case, the validation metric was the dice score computed over each scan after thresholding the network output by means of the Otsu algorithm (this was done due to the high variability of results depending on the threshold selection), computed per volume and averaged afterwards.

To reduce overfitting and improve the performance of both networks at inference time, we also implemented a straightforward Test-Time Augmentation strategy. Besides considering the prediction on a given volume, such volume also went through a reflection over each axis from the set $\{x, y, z, xy, xz, yz, xyz\}$, predictions were computed on the modified volume and the same reflection was applied again on the predictions in the segmentation case. We observed a considerable benefit when applying this strategy, specially in mitigating the overfitting that arised from the heavy oversampling of minority classes in the texture classification scenario, as described in Sect. 2.4.

## 5 Results

In this section we report numerical results for cross-validation performance as well as performance in the final test set.

### 5.1 Lung Nodule Texture Classification

The organization of this challenge provided an official split of all the training scans. This represented a set of 768 nodules, that were split in four subsets of $200, 194, 186$ and $192$ respectively. Each of this subset was used for validation purposes once, while a model would be trained in the union of the remaining three subsets, which resulted in four different models being trained. Table 1 reports the quadratic weighted kappa score, and other metrics of interest, for each of this folds. We also display confusion matrices for each fold in Fig. 5.

For testing purposes, each nodule in the test set was run through each of the above four models, and the resulting probabilities were averaged to build our final submission. This produced a quadratic-weighted kappa score of $\kappa = 0.6134$,

| T/P | GGO | S-S | Solid |
|---|---|---|---|
| GGO | 4 | 3 | 4 |
| S-S | 0 | 3 | 10 |
| Solid | 0 | 7 | 168 |

(a)

| T/P | GGO | S-S | Solid |
|---|---|---|---|
| GGO | 5 | 1 | 2 |
| S-S | 2 | 4 | 8 |
| Solid | 2 | 6 | 163 |

(b)

| T/P | Solid | S-S | GGO |
|---|---|---|---|
| GGO | 2 | 3 | 5 |
| S-S | 1 | 2 | 10 |
| Solid | 0 | 1 | 161 |

(c)

| T/P | GGO | S-S | Solid |
|---|---|---|---|
| GGO | 7 | 1 | 4 |
| S-S | 1 | 1 | 8 |
| Solid | 0 | 9 | 160 |

(d)

**Fig. 5.** (a)–(d): Confusion matrices corresponding to each of the validation folds.

**Table 2.** Validation (top) and test (bottom) results as provided by the organization

| | $J^*$ | MAD | HD | Inv. Pearson CC | Bias | Std. Dev. |
|---|---|---|---|---|---|---|
| Validation | 0.4321 | 0.4576 | 2.2364 | 0.1216 | 125.4638 | 706.6704 |
| Test | 0.4447 | 0.4115 | 2.0618 | 0.1452 | 41.4341 | 129.47 |

which ranked fourth in this competition. The greater performance in the test set can likely be attributed to differences in class proportions between this and the several validation sets used in Table 1.

### 5.2 Lung Nodule Segmentation

In sub-challenge B, the nodule segmentation task was evaluated under a number of different metrics, including Modified Jaccard index (J*), Mean average distance (MAD), Mean Hausdorff distance (HD), Inverted Pearson correlation coefficient, Bias, and Standard deviation[1]. In addition, the organization considered the largest interconnected object as the final segmentation in each case.

Since the functionalities to compute the above metrics were not provided by the organizers, we were unable to perform an analysis similar to the one in the previous section for our cross-validation analysis. For this reason, we report in the first row of Table 2 results obtained by predicting each nodule with the corresponding model trained on the appropriate split of the training set, aggregating those predictions, and computing an overall score over the entire training set this way.

Our final submission was again built by averaging the predictions of each of our four models trained in different folds of the training set. The numerical analysis corresponding to our segmentations in the test set is shown in the bottom row of Table 2. Our approach ranked third in the official challenge leaderboard.

## 6 Discussion and Conclusion

From the results presented above, it can be concluded that both the texture classification and the nodule segmentation tasks were solved to a reasonable level.

[1] The challenge website at https://lndb.grand-challenge.org/Evaluation/ contains rigorous definitions of each of these quantities.

In particular, results in Table 1 are well-aligned with inter-observer variability in this dataset among radiologists, as reported in [12].

Another interesting conclusion to be extracted from Table 1 is the observation that the quadratic-weighted $\kappa$ score captures different properties of a solution when compared with the averaged AUC metric (this was computed adjusting for the support of each class). For instance, the worst results in terms of $\kappa$ score were obtained in fold 3, but this fold also had the second best average AUC. In our opinion, the inclusion of the average AUC together with the $\kappa$ score as the monitoring metric based on which we early-stopped the training of the network was greatly beneficial to avoid falling in a sharp local minima during the optimization process.

Despite an overall better ranking, results for nodule segmentation were slightly poorer when compared with texture classification. A reason for this may have been our approach based on directly segmenting the 3D volume, instead of sampling 2D planes and learning from these. While a 3D model had far less learnable parameters in this case, it was much more computationally intensive in terms of number of operations performed by the network, which led to a slow hyperparameter tuning process.

In addition, we observed that the selection of the binarizing approach once the network had been trained had a great impact in the resulting segmentation, as confirmed by the large standard deviation in Table 2. In our experiments, we observed that if an optimal threshold was selected for each prediction (as opposed to a single threshold for all predictions, or even the adaptive threshold selection algorithm based on Otsu's technique we ended up using), results were much better. In other words, a reasonable binarizing threshold for a particular probabilistic prediction turned out to be very poor when applied to another prediction. We believe future work may focus on a better strategy to select a thresholding value in a per-volume basis, as has been suggested in other medical image segmentation problems [9].

## References

1. National Lung Screening Trial Research Team: Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. **365**(5), 395–409 (2011)
2. Aresta, G., et al.: Towards an automatic lung cancer screening system in low dose computed tomography. In: Stoyanov, D., et al. (eds.) RAMBO/BIA/TIA 2018. LNCS, vol. 11040, pp. 310–318. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00946-5_31
3. Aresta, G., Cunha, A., Campilho, A.: Detection of juxta-pleural lung nodules in computed tomography images. In: Medical Imaging 2017: Computer-Aided Diagnosis, vol. 10134, p. 101343N. International Society for Optics and Photonics, March 2017
4. Aresta, G., et al.: iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. Sci. Rep. **9**(1), 1–9 (2019)

5. Bonavita, I., Rafael-Palou, X., Ceresa, M., Piella, G., Ribas, V., González Ballester, M.A.: Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. Comput. Methods Programs Biomed. **185**, 105172 (2020)
6. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. **106**, 249–259 (2018)
7. Ferreira, C.A., Cunha, A., Mendonça, A.M., Campilho, A.: Convolutional neural network architectures for texture classification of pulmonary nodules. In: Vera-Rodriguez, R., Fierrez, J., Morales, A. (eds.) CIARP 2018. LNCS, vol. 11401, pp. 783–791. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13469-3_91
8. Galdran, A., et al.: Non-uniform label smoothing for diabetic retinopathy grading from retinal fundus images with deep neural networks. Translational Vision Science and Technology, June 2020
9. Galdran, A., Costa, P., Bria, A., Araújo, T., Mendonça, A.M., Campilho, A.: A no-reference quality metric for retinal vessel tree segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 82–90. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_10
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 4th International Conference on 3D Vision (3DV), pp. 565–571, October 2016
11. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 4696–4705. Curran Associates, Inc. (2019)
12. Pedrosa, J., et al.: LNDb: a lung nodule database on computed tomography. arXiv:1911.08434 [cs, eess], December 2019. http://arxiv.org/abs/1911.08434
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2019. CA Cancer J. Cin. **69**(1), 7–34 (2019)
15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826, June 2016
16. 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** (2019)
17. Zhang, M., Lucas, J., Ba, J., Hinton, G.E.: Lookahead optimizer: k steps forward, 1 step back. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 9593–9604. Curran Associates, Inc. (2019)
18. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**, 1856–1867 (2020)