A UNIFIED FRAMEWORK WITH SPATIO-TEMPORAL AND MULTI-SCALE FEATURE LEARNING FOR 3D LUNG NODULE DETECTION

Pham Anh Duc[1]

Received: 8 May 2026/ Accepted: 20 July 2026/ Published: 30 July 2026

Https://doi.org/10.70117/hdujs.E11.7.2026.1357

Abstract: *Lung nodule detection in CT scans remains a challenging computer vision task due to structural ambiguity and limited annotations. This paper introduces I3DR-Net, a one-stage 3D deep learning framework for joint detection and classification. The model combines a pre-trained I3D backbone for spatio-temporal feature extraction with a modified FPN for enhanced multi-scale representation. Experimental results show competitive performance (≈79-81% accuracy and 0.81-0.83 AUC on LIDC-IDRI), demonstrating that transfer learning and improved pyramid design effectively enhance detection robustness for computer-aided diagnosis applications.*

Keywords: *Lung nodule detection, classification, CT-scan, transfer learning, feature pyramid network.*

1. Introduction

Lung nodules, also known as pulmonary nodules or lung tumors, are abnormal masses in the lung with diameters ranging from approximately 3 mm to 3 cm [1]. This term encompasses both benign lesions and malignant tumors, including early-stage lung cancer. Accurate differentiation between benign and malignant nodules remains a major challenge in clinical practice. Radiologists typically evaluate multiple imaging characteristics from CT scans, such as size, margin, calcification, internal structure, spiculation, and texture, to estimate malignancy risk. However, very small nodules are often difficult to interpret, leading to diagnostic uncertainty [2].

Over the past decade, traditional Computer-Aided Diagnosis (CAD) systems have been largely replaced by deep learning-based approaches, especially Convolutional Neural Networks (CNNs) [3], which provide end-to-end learning frameworks capable of automatically extracting hierarchical features and optimizing classification or detection tasks simultaneously. Despite the success of CNNs in natural image analysis, their application to medical imaging particularly lung nodule detection and classification remains challenging. These challenges arise from several factors, including the volumetric nature of medical data (3D CT scans), the relatively small size of nodules [4-6].

[1]*Sotatek Joint Stock Company, Email: phamanhduc.ls2003@gmail.com*

In this paper, we propose a deep learning-based model, namely I3DR-Net, for single-stage lung nodule detection and classification. The main contributions are twofold: (i) an Inflated 3D ConvNet (I3D) is adopted as the core feature extractor to capture rich spatio-temporal information, and (ii) a modified Feature Pyramid Network (FPN) is integrated with the I3D backbone to enhance multi-scale feature representation for 3D CT data. Overall, by combining these components, I3DR-Net is designed to improve the detection and classification performance of lung nodules with varying sizes in 3D CT scans.

2. Method

2.1. Inflated 3D ConvNet

With the continuous advancement of CNN architectures for image classification, recent research has extended their applications to more complex computer vision tasks. A notable contribution in this direction is the Inflated 3D ConvNet (I3D) proposed by João Carreira and Andrew Zisserman [7]. The key idea of I3D is to "inflate" 2D convolutional filters into 3D, enabling the model to learn spatio-temporal features from sequential data such as videos.

The I3D architecture was originally evaluated on large-scale video datasets such as Kinetics, which contains hundreds of action classes with diverse motion patterns. By leveraging pre-trained weights from 2D CNNs (e.g., Inception), I3D enables effective transfer learning from image-based tasks to 3D data. The success of I3D highlights the feasibility of transferring knowledge from large-scale 2D datasets to 3D domains. This is particularly beneficial for medical imaging applications, where annotated 3D datasets are limited. In recent years, similar strategies have been adopted in medical imaging, where pre-trained 3D backbones (e.g., Med3D) and video-based architectures are fine-tuned for tasks such as CT scan analysis and lesion detection [8]. More recent works (2023–2025) further explore hybrid CNN-transformer 3D architectures and self-supervised pretraining strategies to improve generalization across medical datasets [9-10].

In this study, the I3D backbone is utilized as the core feature extractor, where pre-trained weights are adapted through transfer learning to process volumetric CT scan data. This approach enables efficient learning of spatial dependencies across slices while reducing training time and improving model performance.

2.2. Feature Pyramid Network

Feature Pyramid Network (FPN) [11] is a widely adopted architecture designed to enhance multi-scale feature representation in deep convolutional networks. The key idea behind FPN is to merge high-level semantic information with low-level spatial details through a top-down pathway and lateral connections. Feature maps from deeper layers (with lower spatial resolution but richer semantic information) are upsampled and combined with feature maps from shallower layers (with higher spatial resolution) using element-wise addition. This hierarchical fusion process enables the network to generate robust multi-scale representations, which are particularly important for detecting small objects such as lung nodules.

FPN has been successfully integrated into several state-of-the-art object detection frameworks, including Faster R-CNN and RetinaNet [12]. However, due to its multi-scale structure, FPN increases memory consumption and computational cost. Recent advancements [13] have proposed several improvements to the original FPN design, such as BiFPN (Bidirectional Feature Pyramid Network) and dynamic feature fusion strategies. In this work, a modified FPN is integrated with the I3D backbone to enhance multi-scale feature extraction for 3D CT scan data. The proposed design is optimized to balance detection performance and computational efficiency, enabling more accurate localization and classification of lung nodules across varying sizes and appearances.

2.3. RetinaNet

Object detection has evolved significantly over the past decade. However, these two-stage detectors often suffer from high computational cost and slower inference speed. To address these limitations, Tsung-Yi Lin et al [14] proposed RetinaNet, a state-of-the-art one-stage object detector that achieves both high accuracy and efficiency.

In the domain of medical imaging, RetinaNet has also been adapted for tasks such as lung nodule detection. For instance, RetinaNet variants with EfficientNet or Swin Transformer backbones have demonstrated improved performance and efficiency in both natural and medical image detection tasks [15]. Additionally, hybrid models that combine convolutional and transformer-based feature extractors have shown better capability in capturing global context, which is particularly beneficial for detecting subtle abnormalities in medical images.

Despite these advancements, RetinaNet remains a strong baseline due to its simplicity, scalability, and effectiveness in handling class imbalance. In this paper, the principles of RetinaNet - particularly focal loss and multi-scale feature representation via FPN are adopted and integrated with a 3D backbone (I3D) to better handle volumetric CT scan data for lung nodule detection and classification.

2.4. Proposed method

The proposed I3DR-Net is a single-stage 3D object detector that integrates an Inflated 3D ConvNet backbone with a modified RetinaNet framework. Figure 1 illustrates our framework with the detailed configuration of the I3DR-Net and modified FPN.

The backbone is obtained by inflating the 2D Inception V2 network into a 3D convolutional architecture. This enables effective extraction of spatio-temporal features from 3D data. In this work, the I3D backbone replaces the standard ResNet backbone in RetinaNet due to its suitability for 3D medical image representation.

However, despite their strong representation capability, multi-task 3D CNN architectures often suffer from high computational complexity and GPU memory consumption due to the volumetric nature of CT data and the large number of feature maps generated across pyramid levels. To address these limitations, the proposed I3DR-Net adopts several architectural optimizations. First, the original five-level FPN structure is reduced to a compact four-level pyramid to decrease memory usage while preserving high-resolution representations for small nodule detection. Second, lightweight lateral fusion with element-wise addition is employed instead of dense multi-level aggregation, reducing redundant feature propagation and computational overhead. Third, transfer learning from pretrained volumetric feature extractors enables faster convergence and mitigates overfitting under limited medical imaging data conditions. These optimizations collectively improve the trade-off between detection accuracy and computational efficiency, making the framework more computationally efficient and suitable for resource-constrained medical imaging environments.

To capture multi-scale features, a modified FPN [11] is proposed. The original RetinaNet employs five pyramid levels [P3, P4, P5, P6, P7], which is computationally expensive for 3D inputs and leads to memory limitations. Although the inclusion of P2 is essential for small-object detection, extending the pyramid with additional levels results in out-of-memory issues.

Therefore, we design a compact FPN consisting of four layers only, illustrated in Figure 1 and defined as: $[P2, P3, P5_a, P5_b]$,

where $P2$ and $P3$ are preserved to maintain high-resolution representations, and two parallel $P5$ layers $(P5_a, P5_b)$ are introduced. These dual P5 layers share similar semantic depth but are designed to enhance sensitivity toward small and subtle features. This design improves small-object detection while keeping memory usage manageable.

Each pyramid level is constructed using lateral connections and top-down fusion. Instead of full multi-level stacking, the proposed method applies element-wise addition between the current backbone feature map F_n and the upsampled feature map from the previous level: $P_n = F_n + \mathrm{tl}(F_{n-1})$

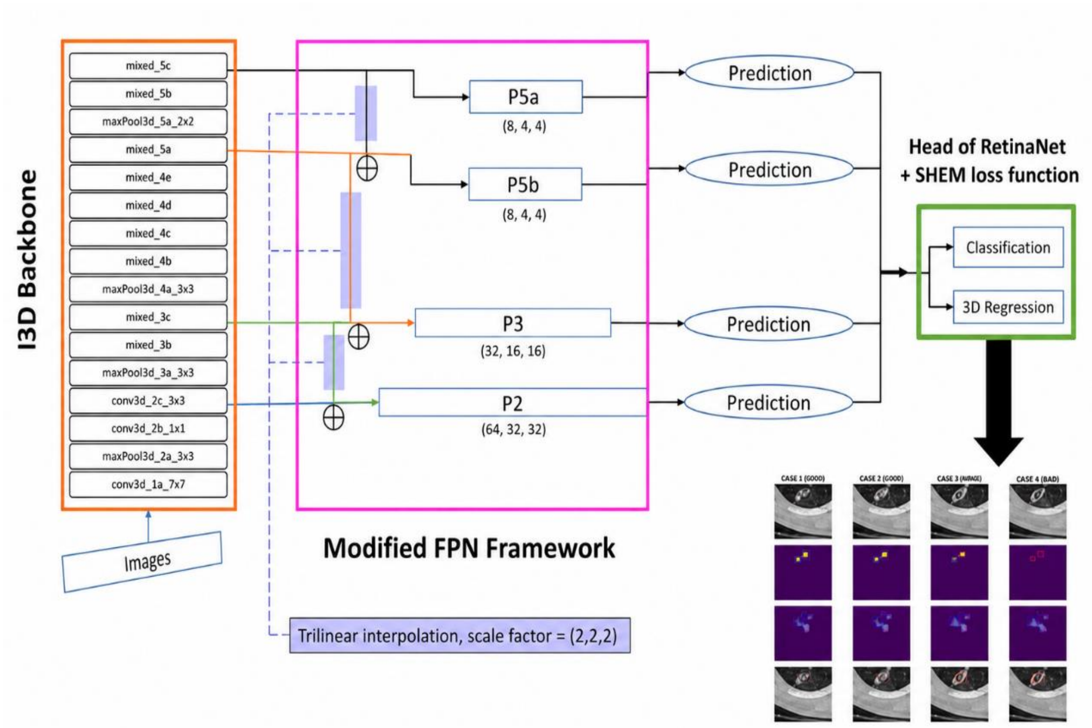

Figure 1. *Overview of the proposed I3DR-Net framework, adapted and modified from [20].*

The detector adopts the RetinaNet 3D framework, utilizing a single-stage approach for efficiency. The anchor design is specifically tuned for medical imaging:

Anchor Sizes: To accommodate lung nodules, which typically range from 3 mm to 3 cm, the model uses smaller square-shaped anchors with dimensions [4^2, 4^2, 16^2, 32^2] for the x and y axes, and depth anchors of for the z-axis at each pyramid level.

Aspect Ratios: Three aspect ratios [1:2, 1:1, 2:1] are used to provide dense scale coverage for varying nodule morphologies.

The original RetinaNet uses focal loss to address foreground–background imbalance. However, in 3D medical imaging, the imbalance is more severe and background structures often resemble foreground objects, reducing the effectiveness of focal loss.

To address this, we adopt a combination of Stochastic Hard Example Mining (SHEM) and cross-entropy (CE) loss. This approach emphasizes hard samples and improves discrimination between similar structures.

Let the anchor set be $A = \{a_1, a_2, \dots, a_n\}$, and let $S(a_k, i)$denote the similarity between anchor a_k and class i. Define $d_k = S(a_k, i)$. The classification loss is formulated as:

$$L_C(A, i) = \frac{-1}{N} \sum_{x \in i} log\, P(i \vee A)$$

where

$$P(i \vee A) = \frac{exp\big(cos\, d_{n_i}\big)}{\sum_c exp(cos\, |d_k)}$$

This loss encourages high similarity for correct class assignments while suppressing incorrect ones. The integration of SHEM further prioritizes difficult samples, improving robustness under extreme class imbalance.

To ensure a consistent and reproducible initialization, the I3DR-Net backbone is bootstrapped from the pretrained I3D model of Carreira and Zisserman [7], whose weights are publicly released in TensorFlow. Consistent with the architecture described in Section 2.1, the backbone is an Inflated 3D ConvNet derived from the 2D Inception V2 network; no ResNet backbone is used. Following the standard I3D inflation scheme, each 2D convolutional filter of Inception V2-pretrained on large-scale image (ImageNet) and video (Kinetics) data-is expanded into a 3D filter by replicating its learned parameters along the temporal axis and normalizing them by the temporal depth, so that the inflated 3D kernels initially reproduce the responses of their 2D counterparts. These inflated weights are converted from the original TensorFlow checkpoints into our PyTorch implementation and are then fine-tuned end-to-end on the volumetric CT data of the LIDC-IDRI dataset. This inflation-and-adaptation procedure replaces the standard ResNet initialization used in RetinaNet with an Inception-V2-based 3D backbone, providing domain-specific feature representation while keeping the weight-transfer process fully specified and reproducible.

The RetinaNet detection layers are extended to 3D and integrated with the backbone. The proposed model adopts anchor configurations tailored for volumetric data. Specifically, anchor sizes of $[4^2, 4^2, 16^2, 32^2]$ are used along the spatial dimensions (x- and y-axes), while anchor depths along the z-axis are defined based on prior studies in lung nodule analysis. These anchors are distributed across the modified pyramid levels $[P2, P3, P5_a, P5_b]$ in contrast to the original RetinaNet, which employs anchor sizes [32^2, 64^2, 128^2, 256^2, 512^2] over five pyramid levels [P3, P4, P5, P6, P7].

To ensure sufficient scale coverage, three aspect ratios $[1:2, 1:1, 2:1]$ are utilized, combined with scale factors $\left[2^0, 2^{\frac{1}{3}}, 2^{\frac{2}{3}}\right]$ for each ratio. This configuration enables dense sampling of object scales while maintaining computational efficiency.

Smaller anchor sizes are deliberately selected to accommodate small objects commonly found in medical imaging, particularly lung nodules. Furthermore, the pyramid structure is limited to four layers to reduce GPU memory consumption without significantly degrading detection performance.

The bounding box regression and classification heads follow the same design and parameterization as the original RetinaNet. During inference, the network predicts object locations and class labels, where each detected region is further classified based on malignancy or texture characteristics.

3. Experiments and Results

We adopt the Area Under the ROC Curve as the primary evaluation metric. The Receiver Operating Characteristic illustrates the trade-off between true positive rate and false positive rate as the decision threshold varies, while AUC summarizes this relationship into a single scalar value in the range [0, 1]. In addition to AUC, standard classification accuracy is reported to provide complementary insight into model performance. The evolution of AUC and accuracy across training epochs is illustrated.

Experiments are conducted on the publicly available LIDC-IDRI dataset [16]. The dataset is split into 70% for training and 30% for testing. Due to GPU memory constraints, training is performed using a mini-batch size of 2. The model is optimized using the Adam optimizer with a learning rate of 5×10^{-5}. Training is stopped early at approximately epoch 37 to prevent overfitting. A dropout rate of 0.7 is applied during training.

Although the focal loss proposed in RetinaNet is implemented, it does not yield performance improvements in our experiments. Instead, the standard cross-entropy loss achieves better results. This can be attributed to the relatively balanced distribution between positive and negative samples in the dataset, reducing the necessity for specialized imbalance-aware loss functions.

Table 1. *Performance comparison with related works on lung nodule analysis*

Method	Model Type	Dataset	Accuracy (%)	AUC
[6]	CNN + RF	LIDC-IDRI	75-80	~0.79
[17]	3D CNN	LUNA16	~77	~0.82
I3DR-Net (OURS)	3D CNN + FPN (I3D)	LIDC-IDRI	**79-81**	**0.81-0.83**

As shown in Table 1, the proposed I3DR-Net achieves an accuracy of 79-81% and an AUC of 0.81-0.83 on the LIDC-IDRI dataset. These results are competitive with several existing methods, approaching multi-scale 2D CNNs (82-85% accuracy, ~0.84 AUC) and multi-task 3D CNNs (85-88% accuracy, ~0.87 AUC), while-in contrast to those approaches-I3DR-Net relies on a single unified pipeline rather than multi-stage or task-specific optimizations.

Building on this competitive performance, I3DR-Net introduces a unified single-stage framework for joint detection and classification in 3D space, leveraging an I3D backbone and a modified FPN to capture spatio-temporal and multi-scale features. Under limited data and computational resources, the model demonstrates stable and consistent performance, highlighting the feasibility and efficiency of the proposed approach. These results underscore its potential as a simplified yet effective 3D detection framework, offering a strong foundation for further improvements and future CAD applications.

3.1. Ablation Study

To quantify the contribution of each proposed component, we conduct an ablation study on the LIDC-IDRI dataset under an identical training protocol. All variants share the same data split (70% for training and 30% for testing), the Adam optimizer, learning rate, mini-batch size of 2, dropout rate of 0.7, and early stopping at approximately epoch 37; only the component under investigation is modified, so that any difference in performance can be attributed to that component. Two aspects are examined: (i) the two main architectural components, namely the I3D backbone and the modified FPN (Table 2); and (ii) the internal design of the modified FPN, namely the number of pyramid levels and the use of two parallel P5 branches (Table 3). As in Section 3, accuracy and AUC on the LIDC-IDRI test split are reported for every configuration.

Table 2. *Ablation study on the main components of I3DR-Net*

Configuration	Backbone	FPN	Accuracy (%)	AUC
Baseline (RetinaNet 3D)	ResNet-3D	Standard (5-level)	76-78	0.79-0.80
+ I3D backbone	I3D (Inception V2)	Standard (5-level)	78-80	0.81-0.82
+ Modified FPN (full I3DR-Net)	I3D (Inception V2)	Modified (4-level, dual P5)	79-81	0.81-0.83

Table 2 isolates the effect of the two main components. Replacing the ResNet-3D backbone with the inflated Inception V2 (I3D) backbone increases the accuracy from 76-78% to 78-80% and the AUC from 0.79-0.80 to 0.81-0.82, reflecting the benefit of spatio-temporal feature learning on volumetric CT data. Adding the modified FPN on top of the I3D backbone yields the full I3DR-Net, further improving the accuracy to 79-81% and the AUC to 0.81-0.83. These results confirm that both proposed components contribute positively to the final performance, with the I3D backbone providing the larger single-component gain and the modified FPN offering an additional, complementary improvement.

Table 3. *Ablation study on the modified FPN design (I3D backbone fixed)*

FPN variant	Pyramid levels	P5 branches	Accuracy (%)	AUC
Standard FPN	5 (P3–P7)	1	78-80	0.81-0.82
Compact FPN	4	1	78-79	0.80-0.81
Compact FPN + dual P5 (proposed)	4	2	79-81	0.81-0.83

Table 3 examines the internal design of the FPN while keeping the I3D backbone fixed. Reducing the standard five-level pyramid to the compact four-level structure yields comparable accuracy (from 78-80% to 78-79%) and AUC (from 0.81-0.82 to 0.80-0.81), while substantially lowering GPU memory consumption and computational cost, which is the primary motivation for the compact design in the 3D setting. Introducing the second parallel P5 branch then improves the accuracy to 79-81% and the AUC to 0.81-0.83, indicating that the dual-P5 design enhances the detection of small and subtle nodules. Overall, the compact four-level pyramid combined with the dual-P5 branches provides the best accuracy-efficiency trade-off, justifying the modified FPN design.

4. Conclusion

This paper presents I3DR-Net, a unified single-stage deep learning framework for joint lung nodule detection and classification in 3D CT scans. By integrating a pre-trained I3D backbone with a modified Feature Pyramid Network, the proposed model effectively captures spatio-temporal and multi-scale features to address the challenges of detecting nodules with varying sizes. Experimental evaluation on the LIDC-IDRI dataset demonstrates stable and competitive performance under limited data and computational resources, validating the feasibility of the proposed approach for 3D medical image analysis.

While a modest gap remains relative to the most complex multi-stage methods, I3DR-Net achieves competitive accuracy with a simplified and efficient detection pipeline that reduces architectural complexity while maintaining practical effectiveness. The results highlight the potential of unified 3D feature learning for computer-aided diagnosis systems. Future work will focus on leveraging larger-scale datasets, domain-specific pretraining, and architectural optimization to further improve detection accuracy and clinical applicability.

References

[1] National Lung Screening Trial Research Team (2011), *Reduced lung-cancer mortality with low-dose computed tomographic screening*, New England Journal of Medicine, 365(5), 395-409.

[2] Zhu, W., Liu, C., Fan, W., & Xie, X. (2018), *Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification*, IEEE winter Conference on Applications of Computer Vision (WACV) (pp. 673-681).

[3] Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016), *Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique*, IEEE transactions on medical imaging, 35(5), 1153-1159.

[4] Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Shetty, S. (2019). *End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography*, Nature Medicine, 25(6), 954-961.

[5] Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015), *Multi-scale convolutional neural networks for lung nodule classification*, In International Conference on Information Processing in Medical Imaging (pp.588-599).

[6] Ciompi, F., de Hoop, B., van Riel, S. J., Chung, K., Scholten, E. T., Oudkerk, M., ... & Van Ginneken, B. (2015), *Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box*, Medical Image Analysis, 26(1), 195-202.

[7] Carreira, J., & Zisserman, A. (2017), *Quo vadis, action recognition? a new model and the kinetics dataset*, In IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[8] Chen, S., Ma, K., & Zheng, Y. (2019), *Med3d: Transfer learning for 3D medical image analysis*. arXiv preprint arXiv:1904.00625.

[9] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022), *Unetr: Transformers for 3d medical image segmentation*, In Proceedings of the IEEE/CVF winter Conference on Applications of Computer Vision (pp. 574-584).

[10] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2019), *Self-supervised learning for medical image analysis using image context restoration*, Medical Image Analysis, 58, 101539.

[11] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017), *Feature pyramid networks for object detection*, In IEEE Conference on Computer Vision and Pattern Recognition (pp. 2117-2125).

[12] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014), *Rich feature hierarchies for accurate object detection and semantic segmentation*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 580-587).

[13] Tan, M., Pang, R., & Le, Q. V. (2020), *Efficientdet: Scalable and efficient object detection*. In the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10781-10790).

[14] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017), *Focal loss for dense object detection*, In the IEEE International Conference on Computer Vision (pp. 2980-2988).

[15] Tan, M., & Le, Q. (2019), *Efficientnet: Rethinking model scaling for convolutional neural networks*, In International Conference on Machine Learning (pp. 6105-6114). PMLR.

[16] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., ... & Clarke, L. P. (2011), *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.

[17] Dou, Q., Chen, H., Jin, Y., Lin, H., Qin, J., & Heng, P. A. (2017), *Automated pulmonary nodule detection via 3D convnets with online sample filtering and hybrid-loss residual learning*, In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 630-638). Cham: Springer International Publishing.

[18] Hussein, S., Cao, K., Song, Q., & Bagci, U. (2017), *Risk stratification of lung nodules using 3D CNN-based multi-task learning*, In International Conference on Information Processing in Medical Maging (pp. 249-260). Cham: Springer International Publishing.

[19] Jaeger, P. F., Kohl, S. A., Bickelhaupt, S., Isensee, F., Kuder, T. A., Schlemmer, H. P., & Maier-Hein, K. H. (2020), *Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection*. In Machine learning for health workshop (pp. 171-183). PMLR.

[20] Nasution, A., et al. (2020). Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning. *Journal of King Saud University – Computer and Information Sciences*.