

Benchmarking Deep Neural Networks for Lung Nodule Classification in LUNA25

M. Taha Gökcan

Department of Computer Engineering
Yıldız Technical University,
Istanbul, Türkiye
taha.gokcan@std.yildiz.edu.tr

Songül Varlı

Department of Computer Engineering
Yıldız Technical University,
Istanbul, Türkiye
svarli@yildiz.edu.tr

Abstract—Early and accurate classification of pulmonary nodules as benign or malignant is critical for improving lung cancer survival rates, particularly when tumors are detected at an asymptomatic stage. In this study, we present a comprehensive benchmark of 2D and 3D deep learning models for malignancy risk estimation on the LUNA25 dataset. We explore a wide range of architectures, from classical CNNs like ResNet to modern models such as ConvNeXt. Furthermore, we assess the impact of Focal Loss versus Binary Cross-Entropy and propose a custom 3D ResNet model that outperforms other models. Our findings highlight that simple 3D architectures, when carefully optimized, offer significant gains in performance.

Keywords—Lung CT, Pulmonary Nodule, LUNA25, Focal Loss

I. INTRODUCTION

Lung cancer is among the leading cancers worldwide in terms of morbidity and mortality. According to the 2020 Global Cancer Statistics, it accounts for approximately 18% of all cancer cases and is associated with 1.8 million deaths globally [1]. While the five-year survival rate for locally advanced disease remains below 5%, detecting the tumor when it is small and asymptomatic can raise the five-year survival rate above 60% [2]. Therefore, early diagnosis and timely treatment are critical for improving patient survival. The primary early indicator of lung cancer is the pulmonary nodule. The complex morphology of pulmonary nodules, their variety in size and shape (ranging from 3 mm to 30 mm in diameter), and their location surrounded by background tissue make the automatic classification of these nodules a challenging task. Early diagnosis relies largely on analyzing these nodules in computed tomography (CT) images.

Accurate classification of pulmonary nodules on CT images as benign or malignant is essential for assessing lung health. However, slice-by-slice inspection of CT data, the complexity and multivariate nature of diagnostic criteria, and the fact that some benign and malignant nodules share similar visual characteristics make this task challenging for radiologists. The constantly increasing volume of medical imaging in hospitals further elevates workload, thus increasing the risk of errors. To address these challenges and provide radiologists with an objective decision-support tool, the development of computer-aided diagnosis (CAD) systems has great importance.

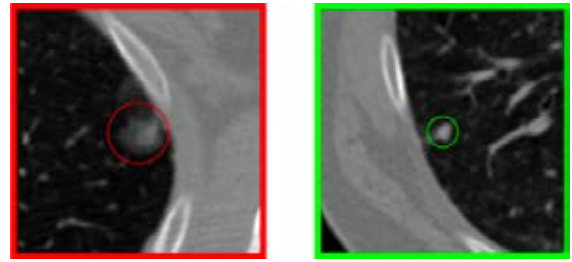


Fig. 1: A malignant and benign nodule from LUNA25 Dataset

In pulmonary nodule classification, convolutional neural networks (CNNs) remain the most widely adopted deep-learning approach because they learn hierarchical features directly from CT data. Both 2D CNNs and 3D CNNs such as ResNet, DenseNet, GoogleNet, Xception are used, while attention mechanisms and Vision Transformers (ViTs) have recently gained popularity for modeling long-range context [1].

Some approaches proposed in recent years will be explored. To counter data scarcity and class imbalance, MSCS-DeepLN [3] ensembles lightweight ResNet over multi-scale inputs. To mitigate imbalance, it incorporates an AUC-based loss function as a penalty term. 3D SAACNet [2] is designed to use volumetric spatial features for benign/malignant classification by utilizing asymmetric convolutions within the segmentation-attention network and then combining deep features through a gradient boosting machine. FocalNeXt [4] merges ConvNeXt-style feature extraction with focal-modulation attention from FocalNet, targeting efficient global-context modeling in high-resolution CT volumes. Using a widened inverted-ResNet backbone augmented with 3D Global Coordinate Attention, GC-WIR [5] aims to improve accuracy with relatively few parameters and stable training dynamics. The PARE [6] model mimics radiologist reasoning by parsing spatial context and recalling learned prototypes. It utilizes stacked Vision Transformer (ViT) blocks. A U-Net based segmentation module parses the nodule's spatial context, while a prototype recall module stores learned cases as references for comparative analysis. Combining multiple self-attention heads with a 3D RTConvBlock to capture local-global dependencies while

TABLE I: Comparison of Focal Loss and Binary Cross-Entropy in Custom 3D ResNet Model

Fold	AUC	Sens@Spec	Spec@Sens	Malignant F1	Benign F1
1	0.837	0.336	0.343	0.428	0.862
2	0.871	0.446	0.472	0.386	0.894
3	0.861	0.471	0.358	0.380	0.870
4	0.874	0.367	0.468	0.447	0.879
5	0.868	0.414	0.548	0.445	0.907
BCE Loss Avg.	0.862	0.407	0.438	0.417	0.882
1	0.846	0.367	0.476	0.452	0.888
2	0.886	0.500	0.637	0.381	0.893
3	0.865	0.471	0.416	0.353	0.836
4	0.881	0.458	0.589	0.444	0.879
5	0.881	0.360	0.619	0.447	0.906
Focal Loss Avg.	0.872	0.431	0.547	0.416	0.880

keeping computation practical, MSA-Net [7] is a recent transformer-based alternative for 3D nodule volumes.

In this context, the LUNA25 Challenge [8] focuses on classifying pulmonary nodules from low-dose chest CT scans as benign or malignant. In addition, a reader study is planned on an isolated data set to compare the relative performance of human experts and artificial intelligence models. The aim of the competition is to strengthen the potential for concrete contributions to clinical decision support workflows by improving the performance gains achieved in various architectures.

The LUNA25 challenge establishes standardized 2D and 3D baselines for the task of malignancy-risk estimation. The 3D component of the challenge utilizes an inflated-3D (I3D) architecture based on I3DR-Net [9], a model designed for nodule detection. While, a 2D ResNet18 [10] model is employed for the 2D stream.

II. MATERIALS AND METHODS

A. Dataset

The primary dataset was LUNA25 from the National Lung Screening Trial (NLST), containing 2,120 patients and 4,069 low-dose chest CT studies, with 555 malignant and 5,608 benign nodules. We employed 5-fold cross-validation with patient-level partitioning to prevent leakage, meaning that all nodules from a given patient were confined to a single fold. Except for the test fold, other folds allocated 80% of cases to training and 20% to validation. Two nodules, one benign and one malignant, are shown in the Figure 1.

Nodules were cropped into $64 \times 64 \times 64$ voxel cubes, clipped to $[-1000, 400]$ Hounsfield Units, rescaled to the 0–1 range, and augmented with random rotations and translations. To overcome data imbalance, malignant nodules were sampled more frequently during training.

B. Experimental Setup

Training was conducted for up to 50 epochs with a batch size of 32. We used the Adam optimizer with a weight decay of 0.0005. Early stopping terminated training if the validation AUC did not improve for 10 consecutive epochs. As loss function, Binary Cross-Entropy (BCE) and Focal loss [11] were examined. Focal Loss improves classification performance in imbalanced datasets by down-weighting easy negatives and forcing the model to focus on harder, informative

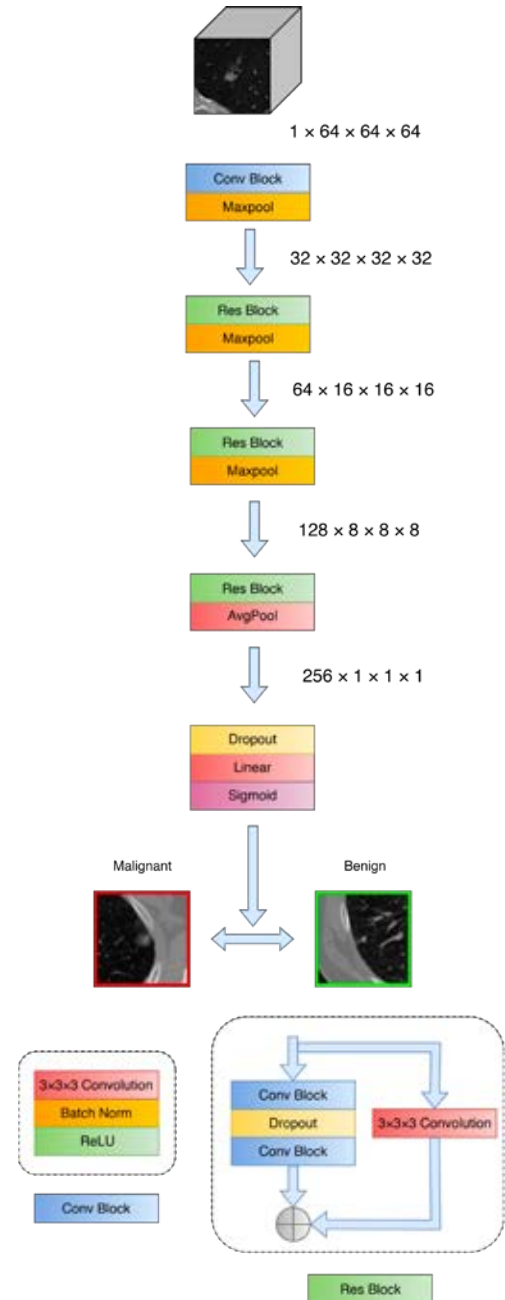


Fig. 2: Custom 3D ResNet Model

TABLE II: Results of 2D Models

Model	# Params	AUC	Sens@Spec	Spec@Sens	Malignant F1	Benign F1
ResNet18	11.2	0.807	0.276	0.379	0.356	0.877
ResNet34	21.3	0.810	0.328	0.366	0.332	0.846
ResNet50	23.5	0.808	0.261	0.375	0.326	0.822
EfficientNet-B0	4.0	0.826	0.295	0.417	0.365	0.858
EfficientNet-B1	6.5	0.818	0.308	0.353	0.348	0.838
EfficientNet-B2	7.7	0.810	0.299	0.382	0.327	0.826
DenseNet121	7.5	0.790	0.305	0.290	0.347	0.885
EfficientNet-V2-S	20.2	0.826	0.330	0.380	0.360	0.857
MobileNet-V3-Small	1.5	0.809	0.288	0.362	0.358	0.848
ConvNeXt-Tiny	27.8	0.858	0.406	0.505	0.411	0.901
ConvNeXt-Small	49.5	0.759	0.192	0.280	0.303	0.836
ConvNeXt-V2-Tiny	27.8	0.828	0.287	0.467	0.339	0.835

examples. Since Focal loss was the most effective, it was utilized in all setup. The initial learning rate was 0.0001 and was annealed to 0.00001 using a cosine scheduler.

Our ResNet-based model begins with a convolutional block comprising a 3×3×3 convolution, Batch Normalization, and ReLU. This is followed by three stacked ResNet blocks. Each block contains two convolutional Conv Blocks separated by Dropout (0.2), wrapped by a residual skip connection. When channel dimensions differ, the skip path uses a 3×3×3 convolution, and outputs are combined by element-wise addition. Spatial resolution is progressively reduced via 2×2×2 max-pooling. The final feature map is compressed using adaptive average pooling, after which a Dropout–Linear–Sigmoid head produces the benign/malignant probability. The proposed model can be seen in the Figure 2.

Beyond our model, we benchmarked a range of off-the-shelf encoders in both 2D and 3D. For 2D, ResNet-18 [10] served as the official LUNA25 baseline. We also evaluated other ResNet variants, EfficientNet [12], EfficientNet-V2 [13], DenseNet [14], MobileNet-V3 [15], ConvNeXt [16], ConvNeXt-V2 [17], and Xception [18]. For 3D, beside the I3D baseline model, we assessed ResNet, EfficientNet, ConvNeXt, ConvNeXt-V2, and Xception [18]. The 3D versions were obtained by replacing 2D convolution and normalization layers with their 3D versions while preserving the original configurations as closely as possible.

C. Evaluation Metrics

To assess the performance of each model in predicting lung nodule malignancy, we employ the following evaluation metrics:

- **Area Under the ROC Curve (AUC):** Measures the likelihood of correctly ranking positive samples higher than negatives.
- **Sensitivity at 95% Specificity (Sens@Spec):** Represents the true positive rate (sensitivity) when the model is constrained to a fixed false positive rate of 5%.
- **Specificity at 95% Sensitivity (Spec@Sens):** Indicates the true negative rate (specificity) when sensitivity is maintained at 95%.
- **F1 Score:** The harmonic mean of precision and recall, offering a single measure that balances both false positives and false negatives.

For all employed 2D and 3D architectures, the final performance metrics are reported as the average results obtained from 5-fold cross-validation.

III. RESULTS AND DISCUSSION

To evaluate models performance, 5-Fold cross-validation was employed. The final metrics were obtained by averaging the results across all folds based on the performance metrics introduced in the previous section.

Table I presents a detailed comparison between Focal Loss and Binary Cross-Entropy (BCE) loss, along with the corresponding results for each fold. As can be seen from the table, Focal Loss consistently achieved higher AUC scores across all folds and yielded nearly a 1-point improvement in the overall AUC compared to BCE. Importantly, it maintained comparable F1 scores, indicating that its enhancement in discriminative ability has not resulted in any loss. Moreover, the notable improvement in the Spec@Sens metric demonstrates the model's strengthened ability to make high-confidence predictions.

Performance results for 2D models are summarized in Table II. To identify optimal model complexity, various configurations of ResNet and EfficientNet architectures were evaluated. The baseline model ResNet-18 achieved an AUC of 0.807, establishing an initial benchmark. While more complex ResNet variants did not show substantial gains in AUC, their F1 scores tended to decline, suggesting a degree of overfitting. EfficientNet architectures exhibited a trend in which increasing parameter counts led to diminished performance. EfficientNet-B0, the smallest variant, achieved the best AUC score among its group (0.826). Notably, EfficientNet-V2-S, despite having approximately five times more parameters than EfficientNet-B0, failed to outperform it. MobileNet-V3-Small delivered competitive performance with very low complexity, achieving results comparable to ResNet-18, highlighting its suitability for lightweight deployment. In contrast, DenseNet-121 yielded unsatisfactory outcomes in this setting. Among ConvNeXt models, the lower-complexity variant outperformed its deeper counterparts, with ConvNeXt-Tiny reaching an the best score of 0.858 AUC. However, ConvNeXt-V2-Tiny did not show further improvement.

Motivated by the effectiveness of lightweight 2D models, their 3D counterparts were subsequently explored. Results are presented in Table III. The I3D baseline model achieved

TABLE III: Results of 3D Models

Model	# Params	AUC	Sens@Spec	Spec@Sens	Malignant F1	Benign F1
I3D	12.2	0.811	0.334	0.330	0.393	0.911
CNN	1.2	0.811	0.324	0.348	0.357	0.853
ResNet	33.2	0.821	0.317	0.375	0.376	0.916
EfficientNet-B0	4.7	0.833	0.342	0.354	0.383	0.882
Xception	22.4	0.828	0.338	0.364	0.386	0.889
ConvNeXt-Tiny	31.3	0.772	0.236	0.248	0.312	0.812
ConvNeXtV2-Atto	4.2	0.835	0.302	0.451	0.377	0.869
Custom ResNet	4.6	0.872	0.431	0.547	0.416	0.880

an AUC of 0.811. In the CNN architecture, where standard convolutional blocks replaced the ResNet blocks, performance remained comparable in terms of AUC, though F1 scores were quite lower. The 3D ResNet variant improved the AUC by 1.4% compared to 2D ResNet18, while EfficientNet-B0's 3D version showed a 0.6% gain, both supporting the advantage of volumetric feature extraction.

Interestingly, despite its strong 2D performance, the 3D version of ConvNeXt-Tiny performed poorly. In contrast, ConvNeXt-V2-Atto, a simpler variant, delivered significantly better results in 3D, suggesting that reduced complexity may benefit learning. Xception also outperformed the ResNet model.

The proposed Custom ResNet model, achieved the highest AUC score of 0.872, outperforming the baseline by nearly 6% and the second-best model (ConvNeXt-V2-Atto) by approximately 4%. Furthermore, improvements in both Sens@Spec and Spec@Sens by around 10% indicate that the model is particularly reliable when operating under high specificity or sensitivity constraints. Although some models demonstrated higher F1 scores for benign cases, likely due to overfitting, the malignant F1 score of the Custom model was notably superior. It is also worth noting that the best 2D ConvNeXt-Tiny outperformed most of the 3D models.

IV. CONCLUSION

This study provides a systematic evaluation of 2D and 3D deep learning models for lung nodule malignancy classification using the LUNA25 dataset. Our experiments demonstrate that ConvNeXt-Tiny 2D architecture achieve strong baseline performance, while deeper models are generally prone to overfitting. When extended to 3D, compact models continue to outperform their 2D versions. We show that using Focal Loss in place of standard Binary Cross-Entropy leads to consistent improvements in AUC and more confident predictions.

Our proposed Custom 3D ResNet model achieves the best overall performance, surpassing other architectures. These findings validate the effectiveness of architecture optimization and loss function design in building reliable, clinically useful AI tools for early lung cancer detection.

REFERENCES

- [1] A. Naik and D. R. Edla, "Lung nodule classification on computed tomography images using deep learning," *Wireless personal communications*, vol. 116, no. 1, pp. 655–690, 2021.
- [2] Z. Guo, J. Yang, L. Zhao, J. Yuan, and H. Yu, "3d saacnet with gbm for the classification of benign and malignant lung nodules," *Computers in Biology and Medicine*, vol. 153, p. 106532, 2023.
- [3] X. Xu, C. Wang, J. Guo, Y. Gan, J. Wang, H. Bai, L. Zhang, W. Li, and Z. Yi, "Mscs-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks," *Medical Image Analysis*, vol. 65, p. 101772, 2020.
- [4] T. Gulsoy and E. B. Kablan, "Focalnext: A convnext augmented focalnet architecture for lung cancer classification from ct-scan images," *Expert Systems With Applications*, vol. 261, p. 125553, 2025.
- [5] W. Wang, S. Yin, F. Ye, Y. Chen, L. Zhu, and H. Yu, "Gc-wir: 3d global coordinate attention wide inverted resnet network for pulmonary nodules classification," *BMC Pulmonary Medicine*, vol. 24, no. 1, p. 465, 2024.
- [6] J. Zhang, X. Ye, J. Zhang, Y. Tang, M. Xu, J. Guo, X. Chen, Z. Liu, J. Zhou, L. Lu *et al.*, "Parse and recall: Towards accurate lung nodule malignancy prediction like radiologists," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 199–209.
- [7] J. Pan, L. Liang, P. Sun, Y. Liang, J. Zhu, and Z. Chen, "Msa-net: multiple self-attention mechanism for 3d lung nodule classification in ct images," *BMC Medical Imaging*, vol. 25, no. 1, p. 193, 2025.
- [8] D. Peeters, B. Obreja, N. Antonissen, C. Jacobs *et al.*, "The luna25 challenge: Public dataset and reader-study for lung nodule malignancy risk estimation," <https://luna25.grand-challenge.org/>, LUNA25 Grand Challenge, 2025, [Online]. Accessed: Jun. 2025.
- [9] I. W. Harsono, S. Liawatiemena, and T. W. Cenggoro, "Lung nodule detection and classification from thorax ct-scan using retinanet with transfer learning," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 3, pp. 567–577, 2022.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [12] M. Tan, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, pp. 6105–6114, 2019.
- [13] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.
- [14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [15] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
- [16] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.
- [17] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16 133–16 142.
- [18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.