Practice 4: A Study of Segmentation and Classification of 3D Lung CT Images in LUNA16 Dataset

Hoang Khanh Dong - 22BA13072
February 2026

***Abstract*—Lung cancer remains the leading cause of cancer-related mortality worldwide, and early detection of pulmonary nodules on computed tomography (CT) scans is critical to improving patient survival. This practice presents a two-stage deep learning pipeline for automated lung nodule detection using the LUNA16 challenge dataset.**

I. Introduction

Pulmonary nodules are small, round lesions found in the lungs that may indicate early-stage lung cancer. Detecting these nodules in chest CT scans is a challenging task due to the large volumetric data, diverse nodule morphologies, and the overwhelming number of non-nodule structures that can mimic nodules.

The LUNA16 (Lung Nodule Analysis 2016) challenge provides a standardized benchmark for evaluating nodule detection algorithms. It consists of 888 CT scans (10 subsets) with expert-annotated nodule locations drawn from the LIDC-IDRI dataset. The primary evaluation metric is the FROC score, defined as the average sensitivity at seven predefined false-positive rates (1/8, 1/4, 1/2, 1, 2, 4, and 8 FP/scan).

This work adopts just **5 subsets** of the dataset for training and testing, using a two-stage approach:

- **Stage 1 — Lung Segmentation:** A lightweight 3D U-Net (4-level encoder-decoder, channels 16–128) segments the lung region from the raw CT volume.
- **Stage 2 — Nodule Classification:** A 3D ResNet-18 (channels 32–256) classifies each candidate as a true nodule or false positive.

II. Dataset

This study uses 5 out of 10 official subsets (subset 0–4) from the LUNA16 challenge, comprising a total of 445 CT scans. Each scan is stored in MetaImage (`.mhd/.raw`) format. Table I summarizes the subset statistics.

TABLE I
Summary of LUNA16 subsets used in this study.

Subset	Scans	Avg. Slices
0	89	256.97
1	89	243.28
2	89	278.17
3	89	268.61
4	89	262.99
Total	**445**	**261.80**

Nodule annotations are provided in `annotations.csv`, which contains the world coordinates (x, y, z) and diameter (in mm) for each nodule identified by at least 3 out of 4 radiologists. Figure 1 shows the distribution of nodule diameters across the 5 subsets.

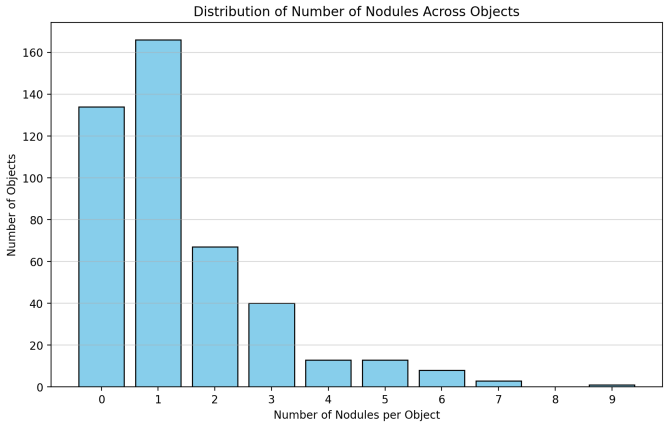

Fig. 1. Distribution of nodule diameters in the selected LUNA16 subsets.

III. Methodology

A. Data Preprocessing

All CT scans are loaded from MetaImage (`.mhd`) format using SimpleITK. The raw Hounsfield Unit (HU) intensities are clipped to the range $[-1000, 400]$ and linearly normalized to $[0, 1]$. Each volume is then resampled to 1 mm isotropic spacing via trilinear interpolation (`scipy.ndimage.zoom`), ensuring consistent physical resolution across scans with varying slice thickness.

Stage 1 (Segmentation): The resampled volumes are divided into non-overlapping chunks of $64 \times H \times W$ along the depth axis. The spatial dimensions H and W are resized to 256×256 using bilinear interpolation. Chunks shorter than 64 slices are zero-padded. Ground-truth lung masks from the `seg-lungs-LUNA16` directory are resampled with nearest-neighbor interpolation and binarized.

Stage 2 (Classification): Candidate locations from `candidates_V2.csv` are converted from world coordinates to voxel coordinates in the isotropic volume. A 32^3-voxel patch is cropped around each candidate center, with

zero-padding applied at volume boundaries. During training, random flips along all three axes (Z, Y, X) are applied as data augmentation.

B. Stage 1: Lung Segmentation

1) Model Architecture: The segmentation model is a lightweight 3D U-Net with a 4-level encoder-decoder architecture and skip connections. It takes a single-channel input of size $64 \times 256 \times 256$ and produces a binary lung mask of the same size. The total parameter count is approximately 1.40 M. Figure 2 illustrates the architecture.

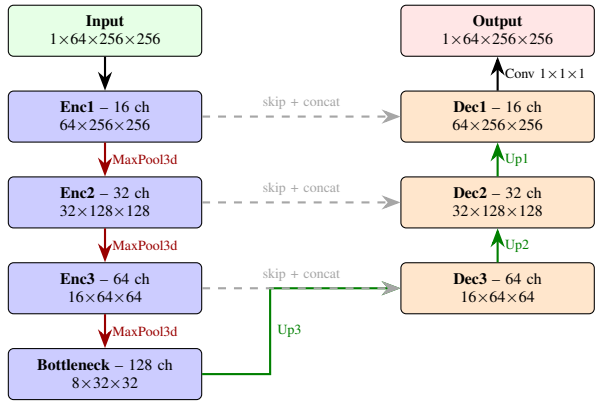

Fig. 2. 3D U-Net architecture for lung segmentation (1.40 M parameters).

Each encoder block consists of two consecutive 3D convolutions ($3 \times 3 \times 3$) followed by GroupNorm and ReLU. Max-pooling ($2 \times 2 \times 2$, stride 2) halves the spatial dimensions between levels. The decoder mirrors the encoder using transposed convolutions for upsampling, with skip connections concatenating encoder features before each decoder block. The final $1 \times 1 \times 1$ convolution produces per-voxel logits.

2) Training and Validation Strategy: I employ 5-fold cross-validation, where each fold holds out one subset for validation and trains on the remaining four. Table II summarises the data split across folds. Each CT volume is divided into non-overlapping chunks of depth 64 (after resampling to 1.0 mm isotropic spacing), with H and W resized to 256×256.

TABLE II
STAGE 1: 5-FOLD CROSS-VALIDATION DATA SPLITS.

Fold	Train Subsets	Train Chunks	Val Subset	Val Chunks
1	{1, 2, 3, 4}	1930	{0}	475
2	{0, 2, 3, 4}	1926	{1}	479
3	{0, 1, 3, 4}	1927	{2}	478
4	{0, 1, 2, 4}	1922	{3}	483
5	{0, 1, 2, 3}	1915	{4}	490

Training. The 3D U-Net is trained for 10 epochs with a batch size of 1 (due to large volume sizes) using AdamW optimizer (weight decay 10^{-4}) and OneCycleLR scheduler (max LR $= 10^{-3}$, 10% warmup, cosine annealing). The loss function combines Dice loss and BCE loss with equal weighting. Automatic mixed-precision (AMP) is used throughout, and gradient norms are clipped at 5.0.

Validation. At the end of each epoch, the model is evaluated on the held-out validation subset. I compute per-voxel Dice coefficient and Intersection-over-Union (IoU) for the lung region. The best checkpoint is selected based on the lowest validation loss. Figures 3–5 show the training curves.

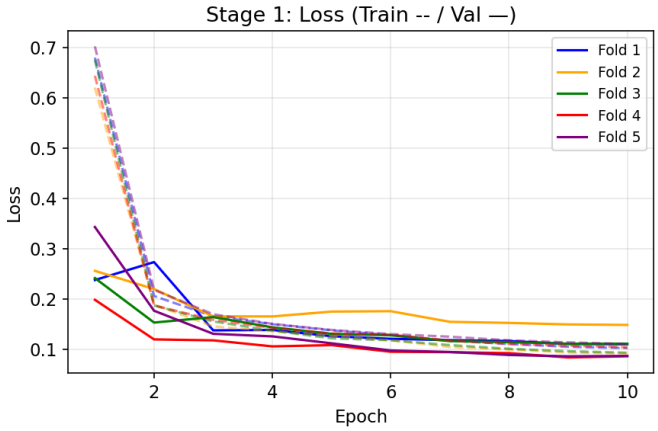

Fig. 3. Stage 1: Training and Validation Loss.

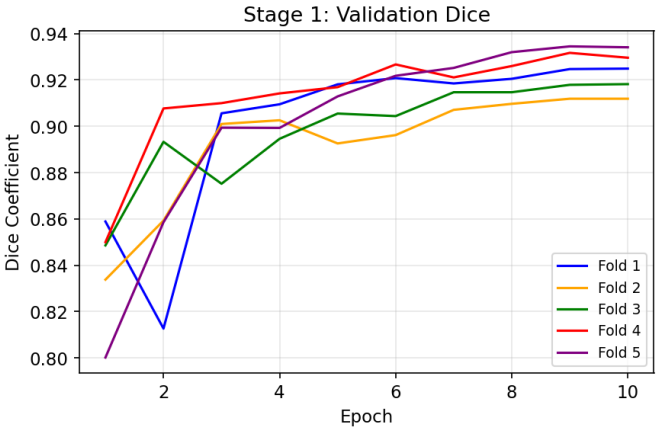

Fig. 4. Stage 1: Validation Dice Coefficient.

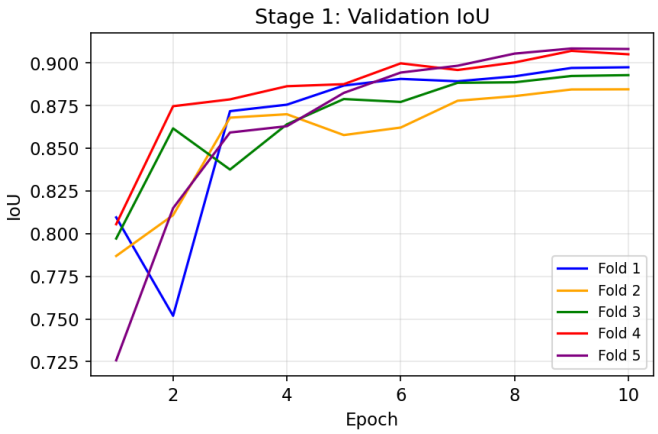

Fig. 5. Stage 1: Validation IoU.

C. Stage 2: Nodule Classification

1) Model Architecture: The classification model is a 3D ResNet-18 adapted for volumetric input. It takes a single-channel 32×32×32 isotropic patch and outputs a single nodule probability logit. The network uses narrower channels (32–256) for efficiency, totalling approximately 3.60 M parameters. Figure 6 illustrates the architecture.

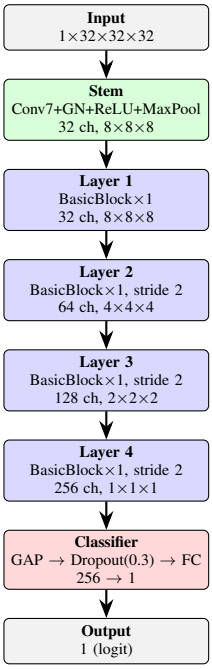

Fig. 6. 3D ResNet-18 architecture for nodule classification (3.60 M parameters).

The stem consists of a 7×7×7 convolution (stride 2) followed by GroupNorm, ReLU, and 3×3×3 max-pooling (stride 2), reducing the input from 32^3 to 8^3. Each of the four residual layers contains one BasicBlock with two 3×3×3 convolutions and a skip connection; layers 2–4 use stride-2 downsampling. The classifier head applies global average pooling, dropout (0.3), and a fully-connected layer to produce the final logit.

2) Training and Validation Strategy: The same 5-fold cross-validation scheme is used, with each fold holding out one subset for validation. Table III summarises the candidate data statistics across folds, highlighting the extreme class imbalance (∼1:440 positive-to-negative ratio in the raw data).

TABLE III
STAGE 2: 5-FOLD CROSS-VALIDATION CANDIDATE STATISTICS.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train subsets	{1,2,3,4}	{0,2,3,4}	{0,1,3,4}	{0,1,2,4}	{0,1,2,3}
Val subset	{0}	{1}	{2}	{3}	{4}
Train scans	356	356	356	356	356
Val scans	89	89	89	89	89
Train total candidates	298,003	306,126	302,680	301,188	300,555
Train positives	679	647	636	659	647
Train negatives	297,324	305,479	302,044	300,529	299,908
Per-epoch samples	1,358	1,294	1,272	1,318	1,294
Val total candidates	79,135	71,012	74,458	75,950	76,583
Val positives	138	170	181	158	170
Val negatives	78,997	70,842	74,277	75,792	76,413
Val samples (10:1)	1,518	1,870	1,991	1,738	1,870

Training. The 3D ResNet-18 is trained for 30 epochs with a batch size of 64 using AdamW optimizer (weight decay 10^{-4}) and OneCycleLR scheduler (max LR $= 10^{-3}$, 10% warmup, cosine annealing). The loss function is binary cross-entropy (BCE). To handle the severe class imbalance, I adopt the following sampling strategy for each epoch:

1) All positive candidates (n_{pos}) are included.
2) An equal number of negatives (n_{pos}) are randomly sampled *without replacement* from the full negative pool, using a different random seed per epoch.
3) The balanced set ($2 \times n_{\text{pos}} \approx 1{,}300$ samples) is ordered by scan ID (scan-level shuffling) to improve cache locality during I/O, then fed through the DataLoader.

Since each epoch sees only ∼0.2% of all negatives, approximately 440 epochs would be needed to cycle through the entire negative pool; however, 30 epochs already provide sufficient diversity. Data augmentation consists of random 3D flips along Z, Y, and X axes (each with 50% probability).

Validation. After each epoch, the model is evaluated on a *fixed* validation subset (deterministic seed per fold). To approximate a realistic class distribution while keeping evaluation tractable, I subsample negatives at a 10:1 ratio relative to positives (e.g., Fold 1: 138 pos + 1,380 neg = 1,518 total). The following metrics are computed using scikit-learn:

- **Accuracy**: overall correct predictions.
- **Precision** and **Recall**: binary metrics with threshold 0.5.
- **F1 Score**: harmonic mean of precision and recall.
- **AUC-ROC**: area under the receiver operating characteristic curve, computed from continuous probabilities (sigmoid output).

The best checkpoint is selected based on the highest validation AUC-ROC, as it is threshold-independent and more robust for imbalanced datasets than loss-based selection. Figures 7–11 show the training curves.

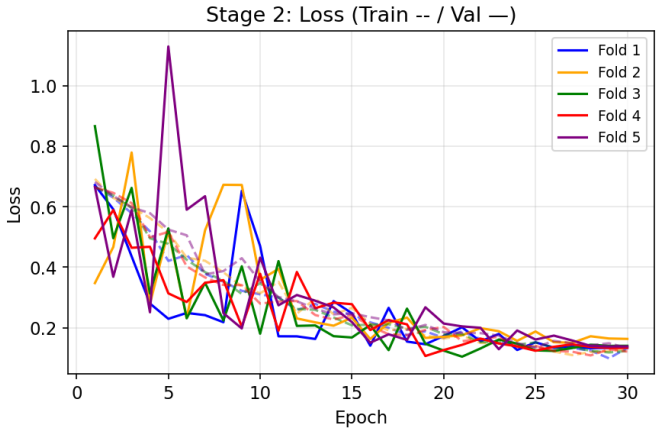

Fig. 7. Stage 2: Training and validation loss.

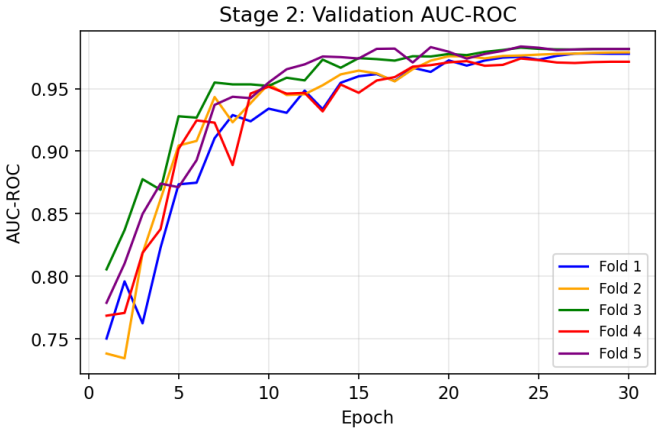

Fig. 8. Stage 2: Validation AUC-ROC.

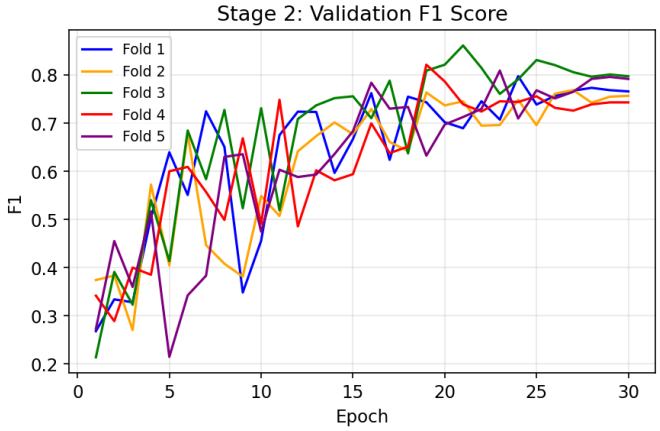

Fig. 9. Stage 2: Validation F1 Score.

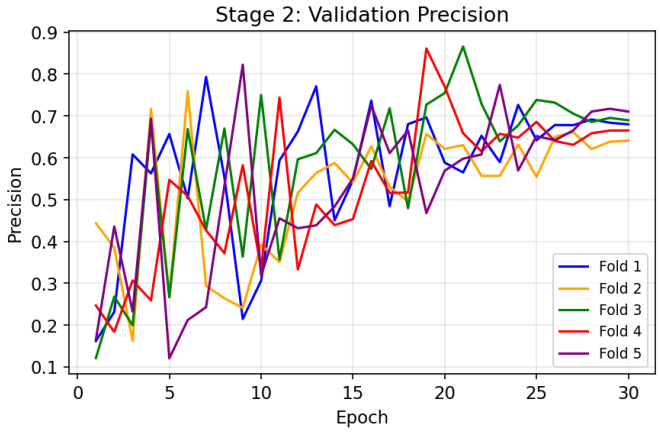

Fig. 10. Stage 2: Validation Precision.

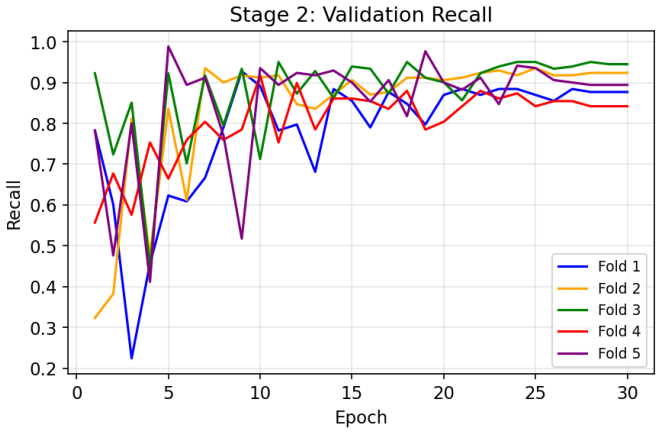

Fig. 11. Stage 2: Validation Recall.

IV. Results

A. Stage 1: Lung Segmentation

Table IV reports the best validation metrics for each fold, selected by the lowest validation loss. The 3D U-Net achieves a mean Dice coefficient of 0.9242 ± 0.0084 and a mean IoU of 0.8980 ± 0.0089 across all five folds, demonstrating consistent and reliable lung segmentation performance.

TABLE IV
Stage 1 best validation results per fold (selected by lowest val loss).

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean ± Std**
Best Epoch	10	10	10	9	9	–
Val Loss	0.1097	0.1481	0.1107	0.0833	0.0857	0.108 ± 0.023
Dice	0.9249	0.9119	0.9182	0.9317	0.9345	0.924 ± 0.008
IoU	0.8974	0.8845	0.8928	0.9070	0.9084	0.898 ± 0.009

B. Stage 2: Nodule Classification

Table V reports the best validation metrics for each fold, selected by the highest AUC-ROC. The 3D ResNet-18 achieves a mean AUC-ROC of 0.9795 ± 0.0035, indicating excellent discriminative ability. The mean recall (sensitivity) of 0.9130 ± 0.0322 shows that the model successfully detects most true nodules, while the precision of 0.6453 ± 0.0423 reflects the remaining challenge of false positives in the heavily imbalanced candidate pool.

TABLE V
Stage 2 best validation results per fold (selected by highest AUC-ROC).

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean ± Std**
Best Epoch	28	30	24	24	24	–
Val Loss	0.1323	0.1630	0.1474	0.1384	0.1908	0.154 ± 0.021
Accuracy	0.9532	0.9460	0.9543	0.9453	0.9299	0.946 ± 0.009
Precision	0.6914	0.6408	0.6772	0.6479	0.5694	0.645 ± 0.042
Recall	0.8768	0.9235	0.9503	0.8734	0.9412	0.913 ± 0.032
F1	0.7732	0.7566	0.7908	0.7439	0.7095	0.755 ± 0.028
AUC-ROC	0.9781	0.9790	0.9828	0.9741	0.9837	0.980 ± 0.004

C. FROC Evaluation

The FROC (Free-Response Receiver Operating Characteristic) score is the official LUNA16 evaluation metric, defined as

the average sensitivity at seven predefined false-positive rates: $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$ FP/scan. Ground-truth nodule locations are taken from the LUNA16 `annotations.csv` file, which contains expert-annotated nodule centres and diameters. A candidate detection is counted as a true positive if its world-coordinate distance to a ground-truth nodule centre is less than the annotated nodule radius (diameter/2). All remaining detections are counted as false positives.

For each fold, the best Stage 2 classifier (selected by AUC-ROC) is applied to *all* candidates in the held-out validation subset, and the FROC curve is computed by sweeping the probability threshold. Table VI reports the sensitivity at each operating point.

TABLE VI
FROC EVALUATION: SENSITIVITY AT EACH FP/SCAN RATE ACROSS 5 FOLDS.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean ± Std**
GT nodules	112	128	128	119	128	–
0.125 FP/scan	0.3571	0.2422	0.2500	0.3613	0.2188	0.286 ± 0.063
0.250 FP/scan	0.5268	0.5234	0.3672	0.5462	0.3984	0.472 ± 0.078
0.500 FP/scan	0.6161	0.7578	0.6250	0.6471	0.5234	0.634 ± 0.078
1.000 FP/scan	0.7232	0.8906	0.7891	0.7479	0.7266	0.775 ± 0.066
2.000 FP/scan	0.8571	0.9766	0.9219	0.8571	0.8750	0.897 ± 0.048
4.000 FP/scan	0.9018	1.0469	1.0859	0.9412	0.9688	0.989 ± 0.071
8.000 FP/scan	0.9464	1.0938	1.1719	1.0504	1.0703	1.066 ± 0.075
FROC Score	**0.7041**	**0.7902**	**0.7444**	**0.7359**	**0.6830**	$\mathbf{0.731 \pm 0.036}$

The pipeline achieves a mean FROC score of $\mathbf{0.731 \pm 0.036}$ across the five validation folds. At low FP rates ($\leq$0.5 FP/scan), sensitivity ranges from 28.6% to 63.4%, indicating room for improvement in reducing false positives. At higher FP rates ($\geq$2 FP/scan), sensitivity exceeds 89.7%, demonstrating strong nodule detection capability when a moderate number of false positives is tolerable. Note that sensitivities exceeding 1.0 at high FP rates arise because a single ground-truth nodule can be matched by multiple nearby candidates.

V. DISCUSSION

Table VII compares the FROC score of my pipeline against published methods on the LUNA16 benchmark. Note that the published methods use all 10 subsets with full 10-fold cross-validation, whereas my pipeline uses only 5 subsets.

TABLE VII
COMPARISON OF FROC SCORES ON THE LUNA16 DATASET.

Method	**Params**	**FROC**	**Subsets**
3DFPN-HS^2 (2019)	∼28M	0.906	10
S4ND (2018)	∼25M	0.897	10
DeepLung (2018)	∼141M	0.842	10
3DFCN (2017)	∼15M	0.839	10
Mine	**5.0M**	**0.731**	**5**

My pipeline achieves a FROC score of 0.731, which is lower than state-of-the-art methods. Several factors contribute to this gap: (1) I train on only 5 of the 10 available subsets, reducing training data by half; (2) the candidate generation relies on the external `candidates_V2.csv` file rather than a learned region proposal network; and (3) the 3D ResNet-18 classifier (3.60 M) is relatively lightweight compared to deeper architectures used by top methods. Nevertheless, the pipeline demonstrates a sound two-stage approach and achieves high sensitivity (>89%) at $\geq$2 FP/scan, suggesting that the classifier is effective but would benefit from better false positive reduction at lower operating points.

VI. CONCLUSION

This practice implemented a two-stage deep learning pipeline for pulmonary nodule detection on the LUNA16 dataset using only 5 of the 10 available subsets. Stage 1 employed a 3D U-Net (1.40 M parameters) for lung segmentation, achieving a mean Dice coefficient of 0.924. Stage 2 used a 3D ResNet-18 (3.60 M parameters) to classify nodule candidates, achieving a mean AUC-ROC of 0.980. The overall pipeline attained a mean FROC score of 0.731 across 5-fold cross-validation. The gap compared to state-of-the-art methods is primarily attributed to hardware limitations: training was restricted to only 5 of the 10 subsets due to limited storage, the number of training epochs was kept small (10 for Stage 1, 30 for Stage 2) to fit within available GPU time, and data augmentation was limited to random 3D flips without more comprehensive strategies such as elastic deformation, intensity jittering, or mixup. With access to more computational resources, training on the full dataset with extended epochs and richer augmentation would likely close this performance gap.

A. Qualitative Results

Figures 12–14 show sample lung segmentation predictions from Stage 1, comparing ground-truth masks (green) with predicted masks (blue) across representative axial slices. Figures 15–21 show sample nodule detection results from Stage 2, with ground-truth nodule boundaries (green circles), true positive predictions (red dots), and false positive predictions (orange dots).

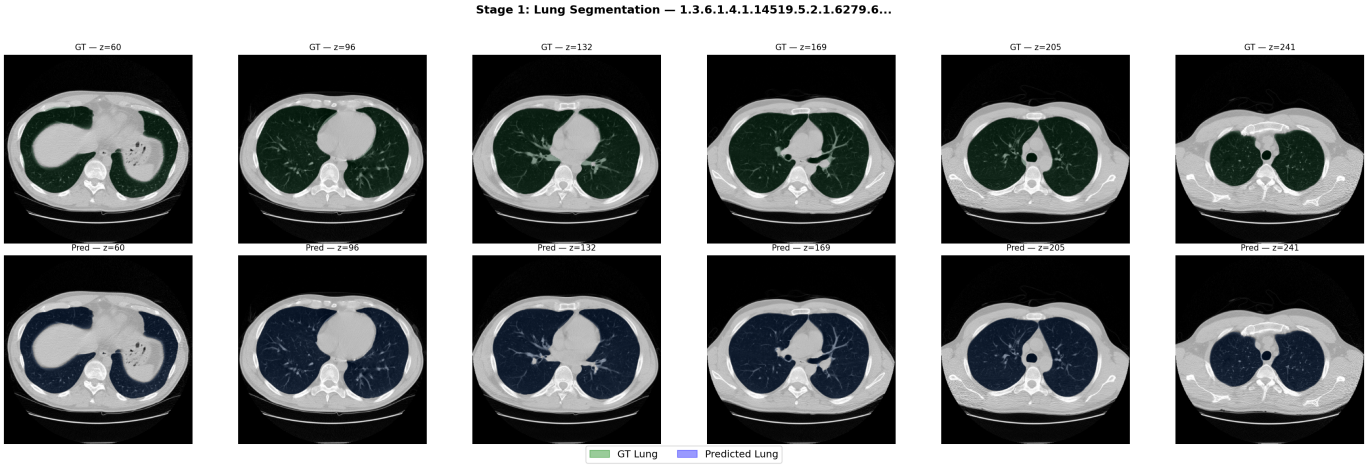

Fig. 12. Stage 1 segmentation: Scan 1 — GT (top) vs Prediction (bottom).

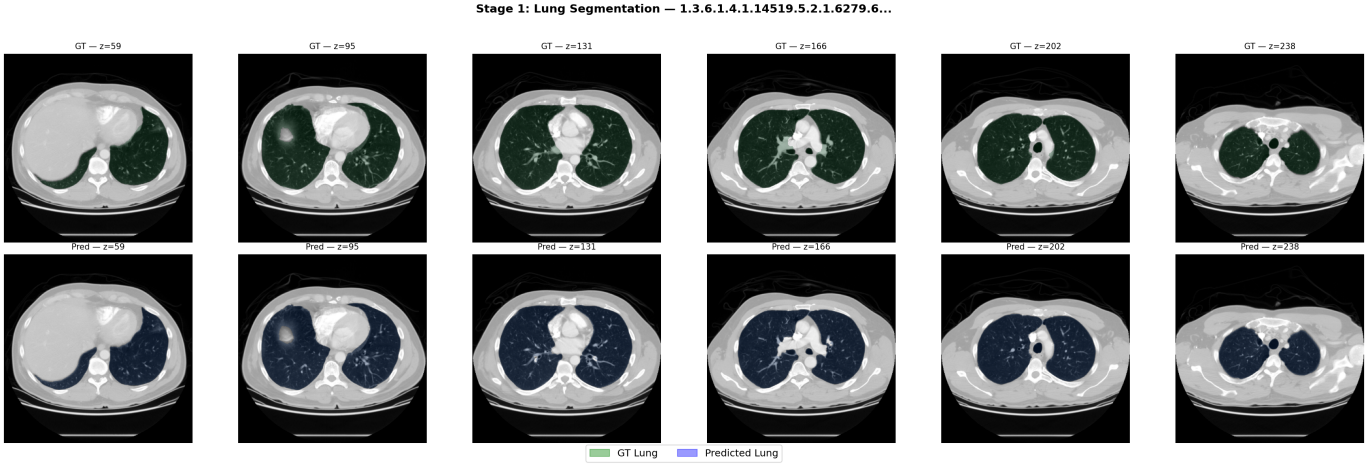

Fig. 13. Stage 1 segmentation: Scan 2 — GT (top) vs Prediction (bottom).

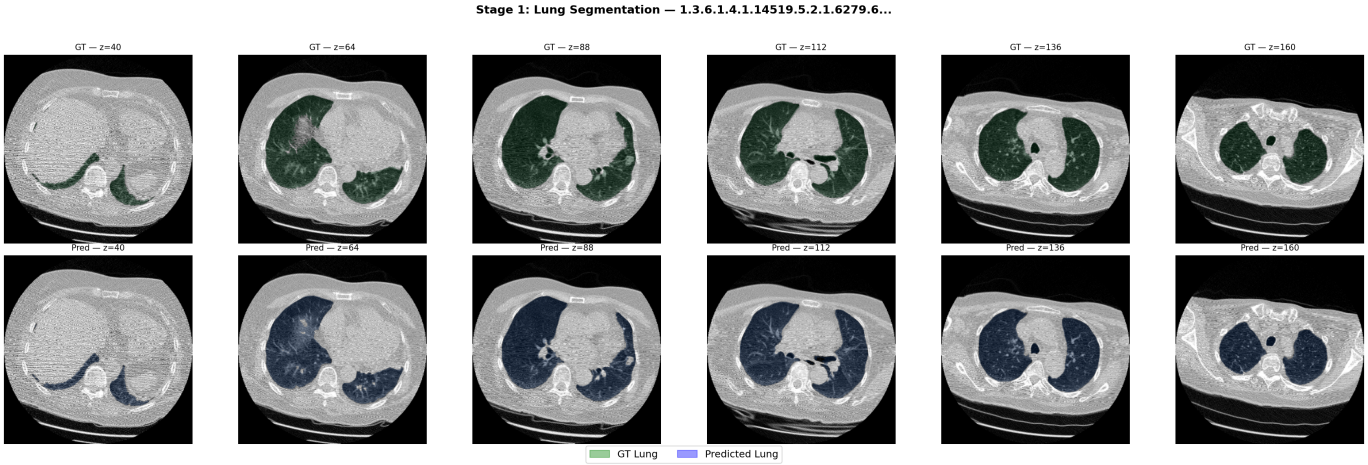

Fig. 14. Stage 1 segmentation: Scan 3 — GT (top) vs Prediction (bottom).

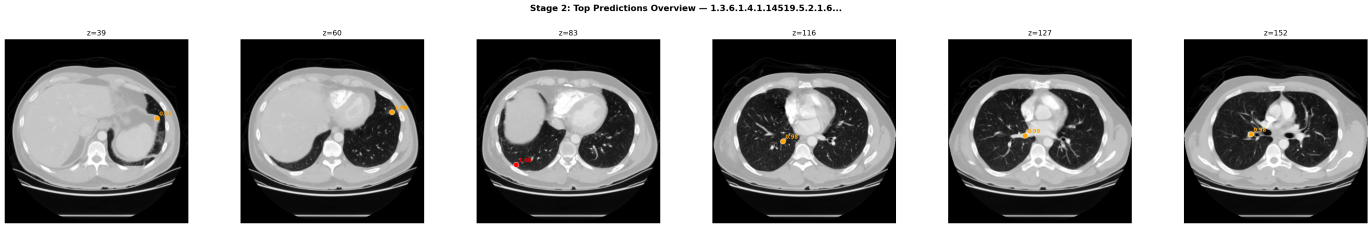

Fig. 15. Stage 2 detection overview: Scan 1 — top predictions across slices.

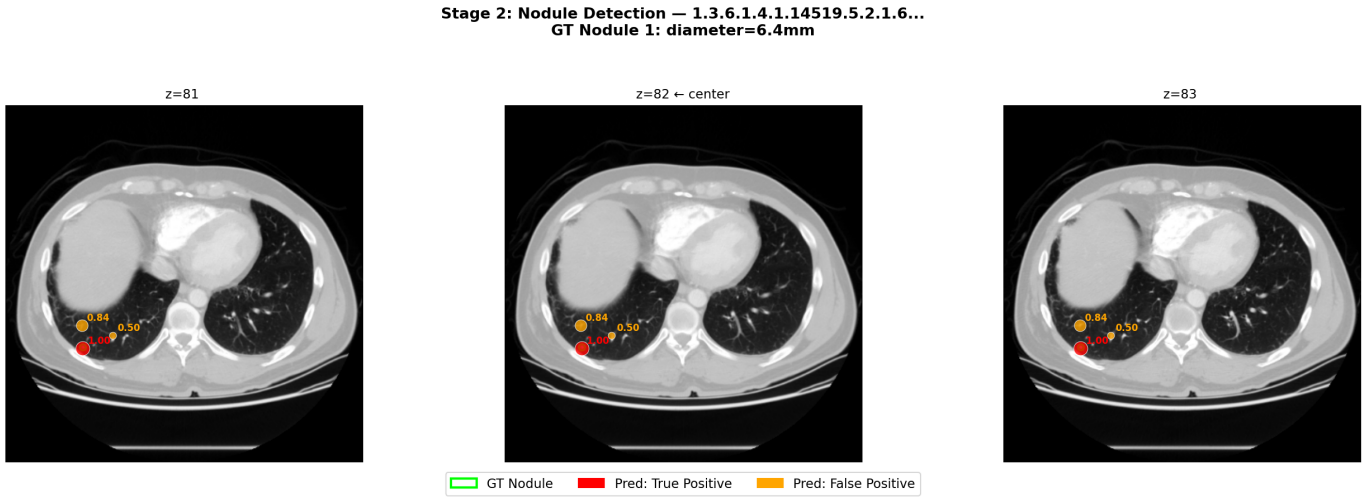

Fig. 16. Stage 2 detection: Scan 1, Nodule 1 close-up.

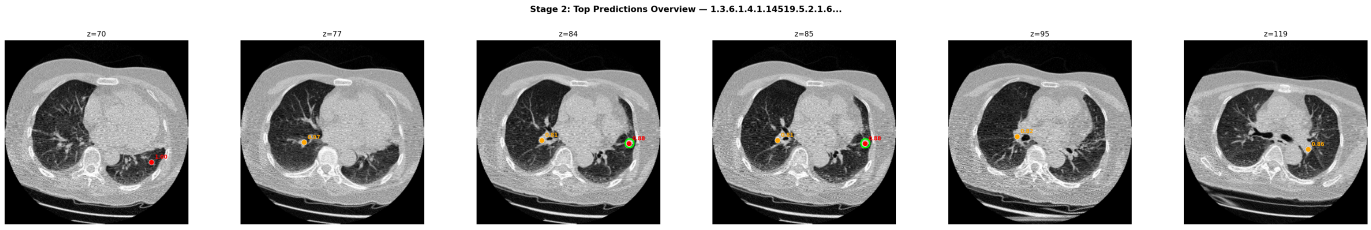

Fig. 17. Stage 2 detection overview: Scan 2 — top predictions across slices.

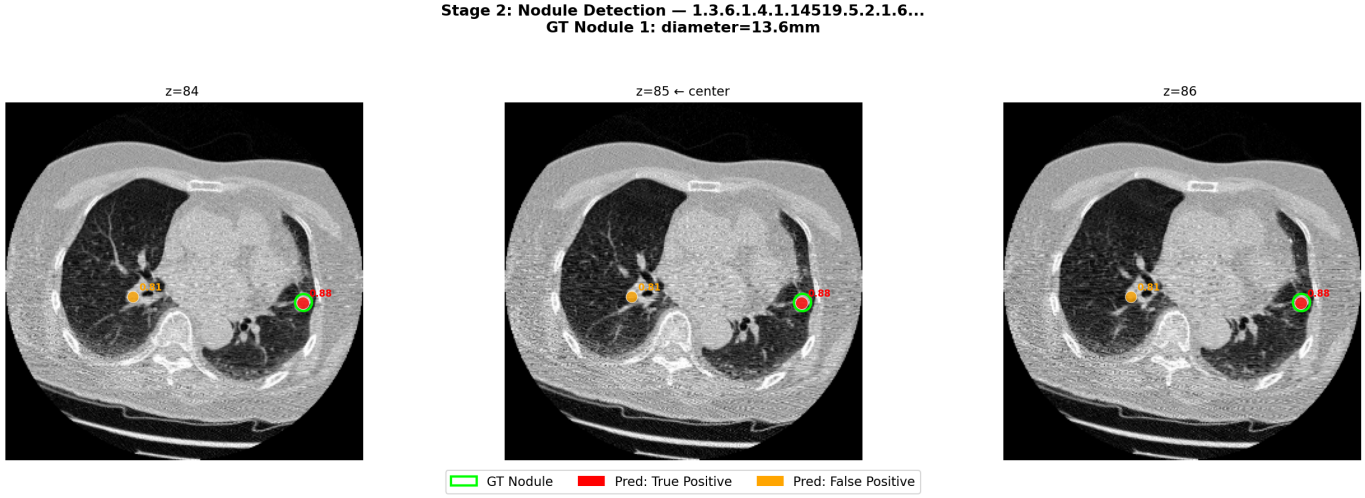

Fig. 18. Stage 2 detection: Scan 2, Nodule 1 close-up.

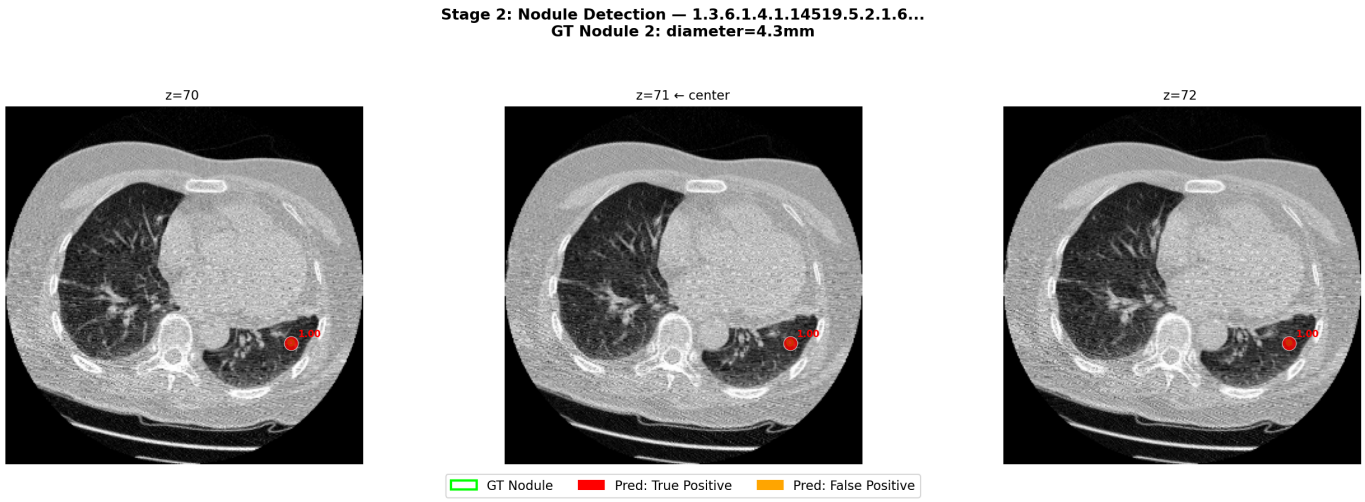

Fig. 19. Stage 2 detection: Scan 2, Nodule 2 close-up.

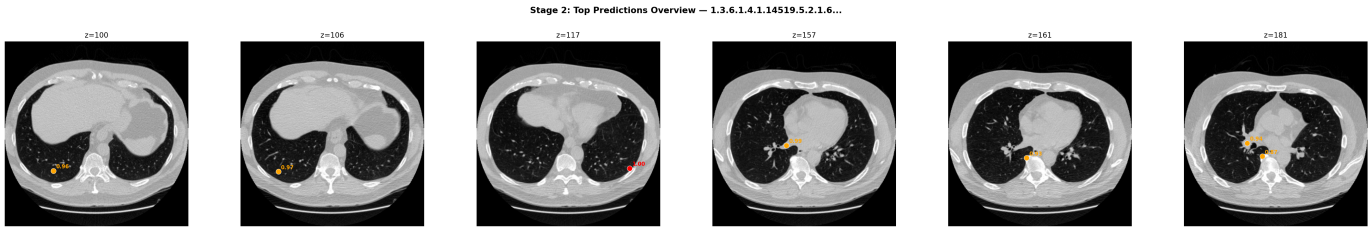

Fig. 20. Stage 2 detection overview: Scan 3 — top predictions across slices.

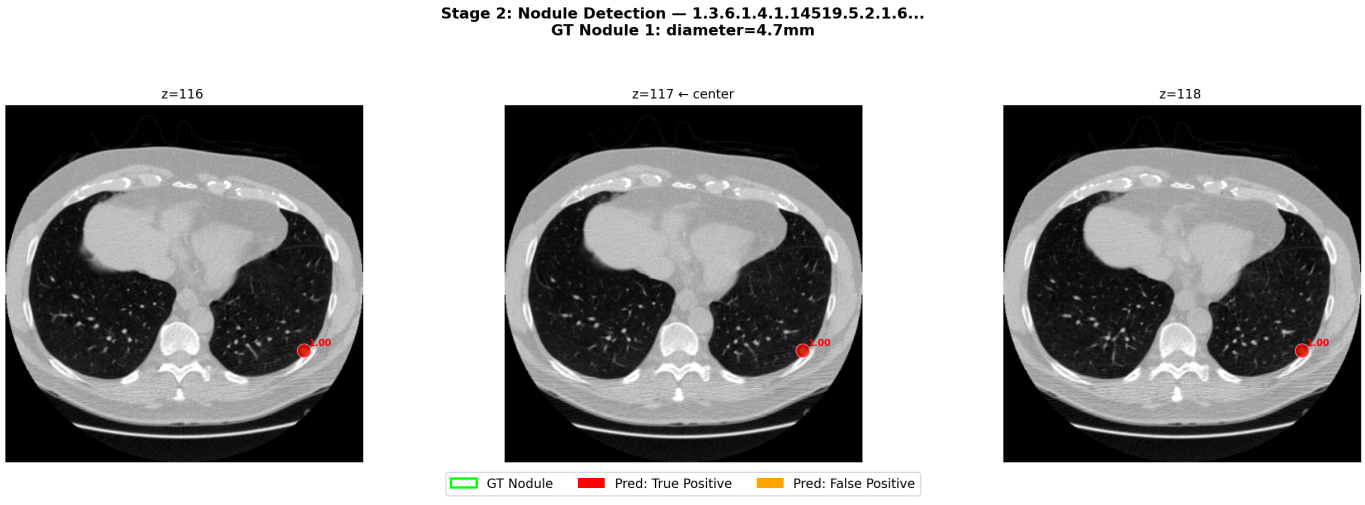

Fig. 21. Stage 2 detection: Scan 3, Nodule 1 close-up.