electronics

MDPI

Article

Multi-Task Deep Learning for Lung Nodule Detection and Segmentation in CT Scans

Runhan Li [1,2] **and Barmak Honarvar Shakibaei Asli** [2,*]

[1] College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2] Centre for Life-Cycle Engineering and Management, Faculty of Engineering and Applied Sciences, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK
* Correspondence: barmak@cranfield.ac.uk

Abstract

The early detection of pulmonary nodules in chest CT scans is critical for improving lung cancer outcomes. While existing computer-aided diagnosis (CAD) systems have shown promise, most treat detection and segmentation as separate tasks, leading to fragmented pipelines and limited representation sharing. This study proposes a 2.5D multi-task learning (MTL) framework that integrates both tasks within a unified Mask R-CNN architecture. The framework incorporates a tailored preprocessing pipeline—including Hounsfield Unit (HU) normalisation, CLAHE enhancement, and lung parenchyma masking—to improve input consistency and task-relevant contrast characteristics. To enhance sensitivity for small or ambiguous nodules, an auxiliary RoI classifier is introduced. Additionally, a nodule-level evaluation strategy aggregates slice-wise predictions across the *z*-axis, supporting a clinically meaningful assessment that approximates 3D diagnostic workflows. Experiments on the LUNA16 dataset demonstrate that the proposed framework achieves a favourable trade-off between detection and segmentation performance under a unified 2.5D multi-task setting. These results highlight the potential of integrated MTL approaches to advance CAD systems for early lung cancer screening.

Keywords: lung cancer; pulmonary nodules; CT imaging; multi-task learning; mask R-CNN; computer-aided diagnosis

1. Introduction

Lung cancer remains a leading cause of cancer mortality, where outcomes depend strongly on detecting small nodules at an early stage [1–5]. Low-dose chest CT is the modality of choice, but manual reading is time-consuming and susceptible to inter-reader variability and missed subtle findings [6,7]. Computer-aided (CAD) pipelines have progressed from hand-crafted features to deep learning, delivering large gains for both detection and segmentation [4,8,9]. Yet, most systems still treat these tasks in isolation, leading to fragmented pipelines as well as limited representation sharing and leaving potential synergies untapped [10–12]. Multi-task learning (MTL) offers a principled way to share representations and potentially reduce architectural redundancy, but practical designs for lung nodules face two persistent hurdles: task imbalance during training and the high computational cost of full-3D processing [13–16]. A typical CAD system is illustrated in Figure 1. The pipeline includes CT acquisition, lung segmentation, lung nodule detection and segmentation, and high-level feature extraction (e.g., nodule volume, shape, and texture) for malignancy assessment.

Check for updates

Academic Editor: Hamed Mozaffari

Received: 13 January 2026
Revised: 5 February 2026
Accepted: 6 February 2026
Published: 9 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

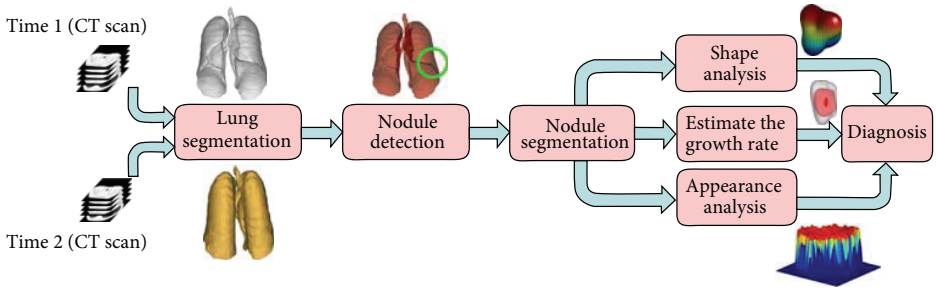


Figure 1. A typical CAD pipeline for lung cancer. The system takes CT scans as input, performs lung segmentation to define the region of interest, and applies nodule detection and segmentation. Key features such as volume, shape, and texture are extracted and used to support diagnostic classification [8].

This work addresses these issues with a 2.5D instance-segmentation framework that jointly detects and segments pulmonary nodules. The approach couples a tailored pre-processing pipeline with a Mask R-CNN backbone adapted to stacked slices, and adds a small auxiliary RoI classifier to boost recall for borderline proposals. Because nodules are volumetric findings, we also score predictions at the nodule level by merging slice-wise outputs into 3D entities. The aim is a practical, design-oriented solution that balances contextual modelling and architectural complexity while preserving clinical relevance in nodule analysis.

1.1. Objectives

The main objective of this study is to design a compact MTL for nodule analysis on CT with three goals:

- Unify detection and segmentation in a single architecture that shares features while preserving task-specific heads.
- Adopt a 2.5D formulation that captures limited through-plane context while avoiding volumetric convolutions required by full-3D models.
- Improve sensitivity to small/low-contrast nodules and evaluate performance at the clinically relevant nodule level.

1.2. Contributions

This work offers:

- A data pipeline tailored for LUNA16 (HU normalisation, CLAHE enhancement, lung masking, and slice-level packaging) for stable 2.5D inputs;
- A 2.5D Mask R-CNN-based MTL model with anchors tuned for small nodules and an auxiliary RoI classifier to retain borderline true positives;
- A nodule-level evaluation protocol that merges slice-wise predictions across z, reporting Precision/Recall/F_1-score alongside Dice/IoU;
- Evidence that the proposed design attains a favourable precision–recall/segmentation trade-off, while avoiding volumetric convolutions, reflecting a structural trade-off between through-plane context and architectural complexity.

2. Related Work

Recent advances in deep learning have significantly improved the performance of pulmonary nodule detection and segmentation. However, most prior studies have addressed these tasks independently, often using object detection networks such as Faster R-CNN or SSD for localisation, and segmentation architectures like fully convolutional networks (FCNs), U-Net, or V-Net. Although effective, such task-specific pipelines neglect shared

representations and frequently introduce redundant computation [7,11,17]. The following sections review representative deep learning models for detection, segmentation, and MTL in this domain.

2.1. Lung Nodule Detection

Table 1 presents a selection of representative detection models. While 2D CNN-based methods such as DetectNet achieved high performance on individual slices, 3D CNN variants provided richer spatial context at the cost of increased memory and computational demand. Hybrid methods combining volumetric CNNs with handcrafted or radiomic features have also been proposed to enhance model robustness [18–23]. More recent one-stage detectors such as YOLO-MSRF [24] introduce multi-scale receptive fields and dedicated small-object layers to improve LUNA16 performance. More broadly, recent advances in the object detection literature have focused on improving robustness and optimisation stability under complex visual conditions. Beyond medical CT, robustness-oriented detection research has also addressed degraded inputs and modality discrepancy (e.g., RGB–Thermal SOD) [25]; meanwhile, general-purpose one-stage detectors such as YOLOv9 continue to advance optimisation stability and efficiency [26].

Table 1. Representative deep learning-based models for pulmonary nodule detection. Metrics reported as available (ACC: accuracy, Sens: sensitivity, AUC: area under curve) [18].

Type	Model/Method	Dataset	ACC (%)	Sens. (%)	AUC	Notes
2D CNN-based	DetectNet [19]	LIDC-IDRI	93.00	89.00	–	F_1-score = 90.96, single-stage
	Faster R-CNN Variant [20]	LUNA16	–	86.42	95.4	Deconv + dual RPN
	YOLO-MSRF [24]	LUNA16	95.41	94.02	–	Single-stage, MSRF
3D CNN-based	Multilevel CNN [21]	LUNA16	–	94.40	–	Multi-context fusion
	Point-supervised [22]	LUNA16	–	80.00	–	Weak supervision
Hybrid	3D CNN + VGG + Radiomics [23]	NLST	76.79	–	78	Feature-level fusion

2.2. Lung Nodule Segmentation

In parallel, segmentation models have evolved from basic FCNs to increasingly sophisticated designs leveraging attention mechanisms, multi-scale fusion, and transformers. Table 2 compares several well-known architectures. U-Net variants, attention mechanisms and transformer-based encoders have reported strong Dice scores on LIDC-IDRI/LUNA16, whereas 3D designs typically require significantly higher memory [18,27–33]. Recent studies have explored more advanced attention-based and feature-fusion strategies for medical image segmentation. Representative examples include APU-Net [34], MCAT-Net [35], and recent attention-driven aggregation segmentation models [36], which share a common focus on enhancing multi-scale feature interaction, boundary awareness, and global contextual modelling to improve segmentation robustness under varying nodule appearances.

Table 2. Representative deep learning-based models for pulmonary nodule segmentation. Reported metrics include Dice similarity coefficient (DSC) and sensitivity [18].

Type	Model/Method	Dataset	DSC (%)	Sens. (%)	Notes
2D CNN-based	Improved U-Net [27]	LUNA16	73.6	–	Residual connections, batch normalisation, skip connections
3D CNN-based	Improved Dig-CS-VNet [28]	LUNA16, LNDb	94.9, 81.1	92.7, 76.9	Feature separation, 3D attention blocks
ResNet-based	DB-ResNet [29]	LIDC-IDRI	89.40	–	Dual-branch structure, intensity pooling, modular ResBlocks

Table 2. *Cont.*

Type	Model/Method	Dataset	DSC (%)	Sens. (%)	Notes
Multi-view CNN	MV-CNN [30]	LIDC-IDRI	82.74	–	Axial/coronal/sagittal branches, late fusion
Multi-scale + view	MSMV-Net [31]	LUNA16, MSD	55.60, 59.94	–	2D view fusion, attention-weighted deep supervision
GAN-based	CSE-GAN [32]	LUNA16, ILND	80.74, 76.36	85.46, 82.56	3D U-Net + CSE attention, sScE discriminator
Transformer-based	DEHA-Net [33]	LIDC-IDRI	87.91	90.84	Dual-encoder, transformer modules, hard attention

2.3. Multi-Task Learning Approaches

Multi-task learning (MTL) combines related objectives to share representations and reduce redundancy; cascaded or parallel designs (e.g., joint detection–segmentation) consistently improve either sensitivity or boundary quality relative to single-task baselines [13,37–39]. Table 3 summarises the selected MTL-based frameworks for nodule analysis. Most adopt either cascaded or parallel architectures, often augmented with attention, deep supervision, or dynamic loss weighting. However, most of these models rely on full 3D processing, which substantially increases architectural complexity and training cost. This has motivated alternative MTL designs that explore different trade-offs between contextual modelling and architectural simplicity, such as formulations based on limited 3D context (e.g., 2.5D representations).

Table 3. Representative multi-task learning (MTL) models for joint nodule detection and segmentation. FROC: free-response ROC; DSC: Dice similarity coefficient [18].

Type	Model/Method	Dataset	FROC (%)	DSC (%)	Key Features
Cascaded	NoduleNet [37]	LIDC-IDRI	+10.27% vs. single-task	83.1	Decoupled tasks, FP reduction, segmentation refinement module
Parallel (Weakly sup.)	Multi-branch U-net + ConvLSTM [38]	LIDC-IDRI	+6.89% vs. single-task	82.26	Sequential context, dynamic loss weighting, weak labels
Hybrid	ECANodule [13]	LIDC-IDRI	91.1	83.4	Dense skip connections, attention, OHEM training
Hybrid (Deep-attention)	MANet [39]	LIDC-IDRI	88.11	82.74	Deep supervision, multi-branch attention, boundary enhancement

2.4. Datasets and Evaluation Metrics

A wide range of public datasets have been used in prior studies as summarised in Table 4. The most frequently adopted include LIDC-IDRI, which provides CT scans with multi-reader annotations, and LUNA16, a curated subset optimised for nodule detection.

Table 4. Representative public datasets for lung nodule analysis. Abbreviations: CT (computed tomography), DX (digital radiography), CR (computed radiography) [18].

Dataset	Modality	Annotation Type	Size	Notes
LIDC-IDRI [40]	CT, DX, CR	Bounding boxes, malignancy ratings (4 readers)	1018 scans	Multi-reader annotated; basis for many derived datasets
LUNA16 [7]	CT	Filtered nodules $\geq$3 mm from LIDC-IDRI	888 scans	High-quality subset of LIDC-IDRI; used for 10-fold CV
NLST [41]	CT, X-ray	Patient outcome, lesion info	54,000+ participants	Large-scale clinical trial; mortality-focused

Table 4. *Cont.*

Dataset	Modality	Annotation Type	Size	Notes
ANODE09 [42]	CT	True nodules and irrelevant findings	55 scans	Partially annotated; used for CAD evaluation
ELCAP [43]	CT	Nodule locations	50 scans	Early low-dose CT dataset; used for CAD benchmarking
Tianchi [44]	CT	3D nodule boxes, clinical labels	1000+ scans	5–30 mm nodules annotated by 3 doctors

In parallel, Table 5 outlines common evaluation metrics used to assess detection and segmentation models. For detection, precision, recall, F_1-score, and FROC are frequently used, often computed at the nodule level. For segmentation, Dice similarity coefficient (DSC) and Intersection over Union (IoU) are the most widely adopted. However, many prior studies report slice-level metrics, which may not fully reflect 3D clinical accuracy.

Table 5. Representative evaluation metrics for pulmonary nodule analysis [18].

Metric	Description	Formula
Accuracy (ACC)	Correct predictions among all cases.	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity (Recall)	True positive rate.	$\frac{TP}{TP+FN}$
Specificity	True negative rate.	$\frac{TN}{TN+FP}$
Precision (PPV)	Positive predictions that are correct.	$\frac{TP}{TP+FP}$
F_1-Score	Harmonic mean of precision and recall.	$2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
Dice Similarity Coefficient (DSC)	Similarity between prediction and GT.	$\frac{2\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$
Intersection over Union (IoU)	Overlap ratio of prediction and GT.	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
AUC	Area under ROC curve.	–
ROC Curve	True positive rate vs. false positive rate.	–
FROC Curve	Sensitivity vs. false positives per scan.	–
CPM	Average sensitivity at 7 FPs/scan.	–
Mean Average Precision (mAP)	Mean of AP over all classes.	$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} AP_i$

Notes: TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives. A and B are the predicted and true binary masks. AUC = Area Under the Curve, ROC = Receiver Operating Characteristic, FROC = Free-response ROC, CPM = Competition Performance Metric (used in LUNA16).

In addition to downstream detection and segmentation performance, image quality assessment (IQA) metrics are also used to evaluate the effect of preprocessing enhancements. Table 6 summarises representative IQA metrics, including full-reference (FR) metrics such as PSNR, SSIM, and FSIM, which compare enhanced images against a reference, and no-reference (NR) metrics like BRISQUE and NIQE, which evaluate quality without ground truth. These measures help quantify the perceptual and structural improvements introduced by contrast enhancement techniques such as CLAHE.

2.5. Summary

Despite these advancements, prior MTL approaches face several limitations. Full 3D models are resource intensive and prone to overfitting on small datasets. Slice-level evaluation metrics remain dominant, providing limited insight into true 3D detection performance. In addition, task imbalance and recall sensitivity, which are critical in early cancer detection, are often under-addressed.

To overcome these challenges, this study introduces a 2.5D instance-segmentation formulation with task-specific heads, anchor tuning for small nodules and an auxiliary RoI classifier to improve recall under a unified 2.5D setting.

Table 6. Representative image quality assessment (IQA) metrics used in medical imaging. Abbreviations: MSE (Mean Squared Error), SSIM (Structural Similarity Index), PSNR (Peak Signal-to-Noise Ratio), PIQE (Perception-based Image Quality Evaluator), NIQE (Naturalness Image Quality Evaluator), BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) [18].

Type	Metric	Description	Formula
FR	MSE	Calculates the mean squared error between images. A smaller MSE indicates higher image similarity. A lower value indicates better quality.	$\mathrm{MSE}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$, where x_i and y_i are corresponding pixel values in the two images, and N is the total number of pixels.
	SSIM	Assesses perceptual similarity based on luminance, contrast, and structure.	$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$, where μ_x, μ_y are the mean intensities of images x and y, respectively; σ_x, σ_y are their standard deviations; σ_{xy} is the cross-covariance. C_1 and C_2 are constants to stabilise the division.
	PSNR	Measures signal fidelity between two images. A higher value indicates better image quality.	$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right)$, where MAX is the maximum possible pixel value of the image.
NR	PIQE	Evaluates perceptual quality based on block-level distortion. It analyses local blocks for artefacts and noise. Lower value indicates better quality.	$\mathrm{PIQE} = f(\mathrm{BM},\mathrm{NM})$, where BM is blockiness measure and NM is noise measure.
	NIQE	Estimates deviation from natural image statistics. A lower value indicates better quality.	$d = \sqrt{(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)}$, where μ_1, μ_2 and Σ_1, Σ_2 are the mean vectors and covariance matrices of the test and natural.
	BRISQUE	Predicts image quality using natural scene statistics. A lower score means better quality.	$\hat{I}(i,j) = \frac{I(i,j)-\mu(i,j)}{\sigma(i,j)+C}$, where $\mu(i,j)$ and $\sigma(i,j)$ are local mean and standard deviation from MSCN coefficients. Features are modelled using AGGD, and quality is predicted via SVR.
	Entropy	Measures intensity distribution complexity. Higher values indicate greater contrast and detail.	$H = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$, where $p(x_i)$ is the normalized histogram probability of pixel value x_i.

3. Methodology

This study adopts a 2.5D design that stacks five adjacent CT slices as input. This provides limited through-plane context while retaining a standard 2D instance-segmentation backbone. The formulation avoids volumetric convolutions used in full 3D models, which is treated as a structural difference rather than an empirically measured efficiency advantage.

3.1. Data Collection

LUNA16 [7] is a widely used benchmark dataset derived from the LIDC-IDRI dataset [40]. It contains annotated chest CT scans with 3D volumes and labelled nodule locations. The dataset is divided into 10 subsets (subset0–subset9) for cross-validation. All experiments in this study are conducted on the LUNA16 subset0, which is intentionally selected as a controlled experimental subset. Rather than serving as a large-scale benchmarking dataset, subset0 is used to facilitate systematic analysis of architectural design choices and training strategies under a fixed and reproducible setting. For each scan, the dataset provides two types of annotations:

- **Lung Masks:** Binary segmentation masks identifying the lung parenchyma. These are provided as separate volumetric files aligned with each CT scan and are used to suppress non-lung regions during preprocessing.
- **Nodule Annotations:** Structured CSV files provide nodule locations in world coordinates (in millimetres), including centre coordinates (x, y, z) and nodule diameter. These annotations are converted to voxel coordinates using the scan's spatial metadata (origin and spacing). For segmentation supervision, approximate 2D masks are generated by projecting the annotated nodule diameter onto axial slices, resulting in simplified geometric masks rather than voxel-level delineations of true nodule morphology.

In total, the LUNA16 subset0 comprises 89 scans (67 with nodules, 22 without). The split preserves class balance as shown in Table 7. The decision to use only subset0 is

motivated by computational constraints and the goal of developing and validating a prototype framework on a manageable subset before scaling to full cross-validation.

Table 7. Train/validation split and sample counts for 2.5D stacks.

	Scans (Nodule/Non-Nodule)	Samples Total	Pos.	Neg.
Training	70 (53/17)	317	266	51
Validation	19 (14/5)	75	60	15

Because of memory and time constraints, we adopted a 2.5D sampling approach: from each volume, we extract five adjacent axial slices centred on the nodule (for positive cases) or on a randomly selected z-position in the middle third of the lung volume (for negative cases). The five slices are stacked channel-wise to form a 5-channel input tensor. While this strategy retains some 3D context, it does not fully capture the volumetric extent of elongated nodules, and limits the diversity of negative examples. We return to this limitation in the discussion.

3.2. Preprocessing Pipeline

To ensure high-quality and consistent input, the following preprocessing steps were applied to each CT scan (Figure 2). The pipeline comprises: (1) HU clipping to $[-1000, 400]$ and linear normalisation to $[0, 1]$; (2) slice-wise CLAHE to enhance local contrast; (3) multiplication by the provided lung masks to suppress non-lung regions; (4) conversion of nodule annotations to voxel space and slice association; and (5) packaging five-slice stacks $\{z-2, \ldots, z+2\}$ for model input (positives centred on nodules; negatives sampled within the mid-lung region).

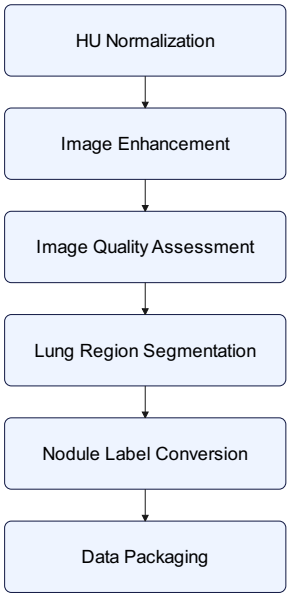


Figure 2. Overview of the data preprocessing pipeline.

This preprocessing pipeline is designed to produce visually enhanced and anatomically focused inputs while preserving structural consistency. To assess whether CLAHE introduces perceptual trade-offs (e.g., noise amplification or over-enhancement) and to quantify its effect on image characteristics, we conduct a lightweight IQA on all scans. We adopt one full-reference metric (PSNR) and two no-reference metrics (BRISQUE and entropy), which jointly capture fidelity, naturalness and contrast/detail. To assist visual inspection, a slice-level viewer is implemented. It displays CT slices in grayscale, keeps

the lung area unchanged, and renders the background in white or light blue for contrast. Example visualisations are provided in Section 4.

3.3. Multi-Task Learning Architecture

The proposed framework performs lung nodule detection and segmentation using a unified instance segmentation architecture based on Mask R-CNN. It integrates a shared feature backbone with dual-purpose task heads to achieve simultaneous object-level localisation and pixel-wise segmentation. The overall architecture, including the detailed layer-wise structure and data flow across shared and task-specific branches, is illustrated in Figure 3. This figure reflects the final implementation of the model, encompassing ResNet backbone stages, FPN output levels, region proposal generation, and parallel detection, segmentation, and auxiliary classification branches.

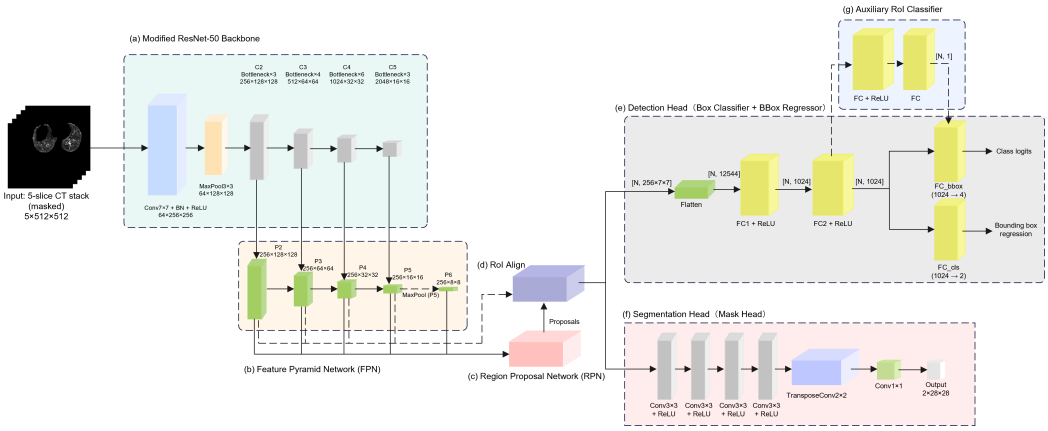
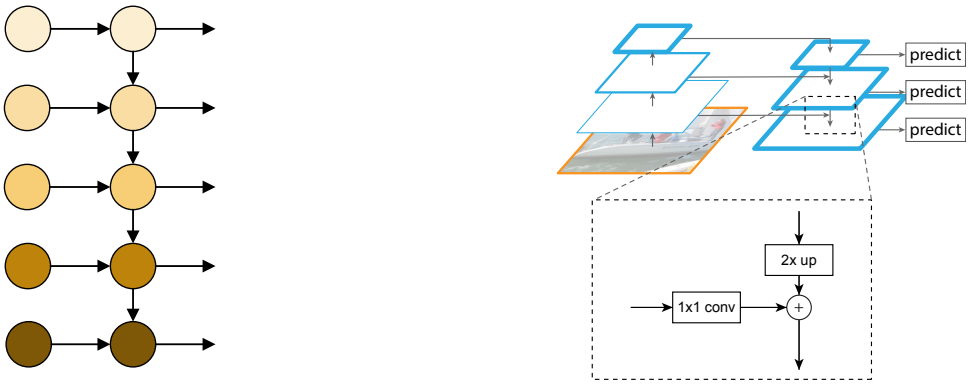


Figure 3. Detailed architecture of the proposed multi-task learning framework based on 2.5D Mask R-CNN. The figure illustrates the data flow through backbone, FPN, RPN, and parallel detection, segmentation, and auxiliary classification branches.

3.3.1. Shared Backbone and FPN

The model is built upon the `maskrcnn_resnet50_fpn` implementation in `torchvision`, using a ResNet-50 backbone pretrained on ImageNet combined with a Feature Pyramid Network (FPN) to produce multi-scale feature maps. An overview of the FPN architecture and its lateral connection building block is shown in Figure 4, highlighting how multi-scale features from different backbone stages are fused via a top–down pathway. Following the FPN, a Region Proposal Network (RPN) [45] operates on each pyramid level to generate candidate object regions. The RPN slides over the feature maps, predicting for each predefined anchor box an objectness score and bounding box offsets. Top-scoring proposals from all pyramid levels are then collected, ranked, and passed to the RoI Align layer for feature extraction in the downstream detection and segmentation heads. The architecture of the RPN is illustrated in Figure 5. To accommodate the 2.5D input design, the initial convolutional layer (`conv1`) is replaced with a 5-channel variant. The pretrained RGB weights are copied into the first three channels, while the remaining two channels are initialised with the mean of the RGB weights. This preserves the benefits of transfer learning while allowing the model to exploit the cross-slice context.



(**a**) Overview of the FPN structure (**b**) Building block of FPN

Figure 4. Feature Pyramid Network (FPN) architecture and building block. (**a**) The traditional FPN fuses multi-scale features from the backbone in a top–down pathway via lateral connections [46]. (**b**) Each building block upsamples a coarser feature map, merges it with the corresponding bottom–up map via a 1×1 convolution and element-wise addition, and applies a 3×3 convolution to reduce aliasing [47].

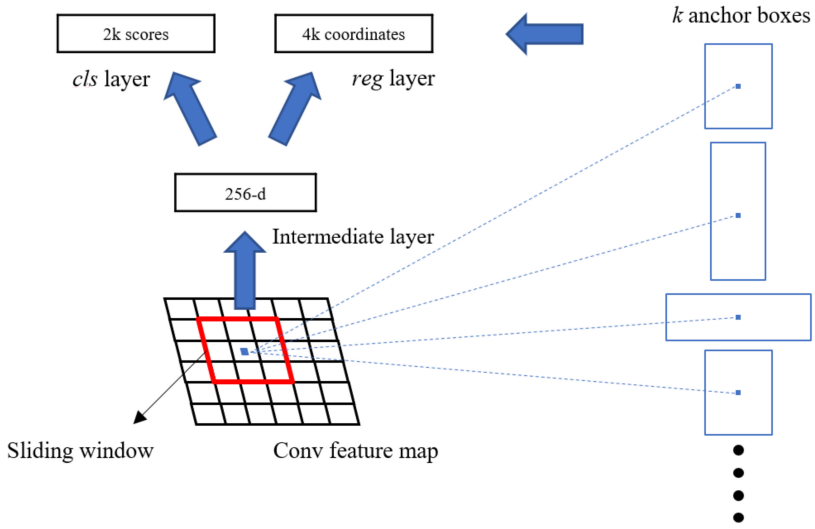


Figure 5. Region Proposal Network (RPN): Each sliding window location on the convolutional feature map produces intermediate features that are fed into two sibling layers for classification (objectness) and regression (bounding box refinement) across multiple anchors [45].

3.3.2. Detection and Segmentation Heads

The detection branch follows the Faster R-CNN paradigm, performing classification and bounding-box regression on each region proposal. The segmentation branch uses a small fully convolutional network (FCN) to predict a binary instance mask that approximates the spatial extent of each detected nodule, conditioned on box-derived supervision. The segmentation head consists of four consecutive 3×3 convolution layers with ReLU activation, followed by a transposed convolution for upsampling to 28×28 pixels per mask, which are resized to the original image resolution during inference.

3.3.3. Design Motivation and Model Evolution

The architectural choices in this study were driven by empirical observations from initial experiments and a progressive refinement process. Preliminary experiments with 3-slice input stacks yielded poor detection performance (maximum F_1-score = 0.4186). This limitation motivated the transition to 5-slice stacks, which capture richer anatomical context

across adjacent slices and significantly improve proposal localisation and recall. To address these limitations, a series of iterative design adjustments was introduced:

- **Anchor Generator Optimisation.** The default RPN anchor scales were manually redefined as $(4, 6, 8)$, $(8, 12, 16)$, $(16, 24, 32)$, $(32, 48, 64)$, and $(64, 96, 128)$ pixels, each with aspect ratios $(0.5, 1.0, 2.0)$. This reconfiguration improves anchor coverage for small nodules often under 10 mm in diameter, which are frequently missed under the default setup in LUNA16.
- **Increased Proposal Count.** The maximum number of RPN proposals was raised to 2000 during training and 300 during inference. This allows more candidate regions to be passed to the RoI heads, increasing the likelihood of capturing true nodules.
- **Auxiliary RoI Classifier.** A fully connected classifier was added after the RoI feature extraction stage to predict whether a proposal corresponds to a true nodule. During training, proposals with IoU ≥ 0.3 to any ground truth were labelled positive, and a binary cross-entropy loss $\mathcal{L}_{\text{aux}}$ was added to the total loss to supervise this auxiliary classifier. The complete loss formulation is presented in Section 3.3.4. In inference, the auxiliary logits were fused with the primary classification scores to re-rank predictions. The auxiliary RoI classifier was designed as a complementary decision head that focuses on local RoI-level discrimination, while the primary classification head prioritises recall at the proposal level.

The auxiliary classifier was introduced to complement the main classification head, which tends to under-score ambiguous or low-contrast nodules, by providing an additional confidence signal for borderline proposals. By adding an auxiliary binary classifier trained with looser IoU thresholds, the model gained a secondary decision path to better retain borderline true positives. This fusion mechanism also helps mitigate the risk of over-suppressing uncertain regions during inference.

These modifications specifically targeted the primary causes of false negatives in the baseline model: insufficient anchor coverage for small lesions, overly aggressive suppression of low-scoring proposals, and weak classification of borderline nodules. By expanding the proposal search space and introducing a secondary decision path, the enhanced architecture became more capable of retaining true nodules while maintaining controlled false positive (FP) rates. These adjustments collectively improved both proposal sensitivity and overall detection performance. Further refinements at the training strategy level are discussed in Section 3.5.

Although most of the published detection, segmentation and MTL models do not report parameter count or floating-point operations (FLOPs), we report the model size here for transparency and to characterise the architectural scale of the proposed framework. The implementation contains 43.85M trainable parameters, which is comparable to standard 2D Mask R-CNN variants. By adopting a 2.5D formulation, the design avoids volumetric convolutions and duplicated encoders commonly used in full 3D or cascaded MTL pipelines.

3.3.4. Loss Functions

A unified loss function is based on the native implementation of the Mask R-CNN model from `torchvision`. The full training objective integrates supervision signals from both detection and segmentation tasks, as well as an auxiliary classification branch. Formally, the total loss $\mathcal{L}_{\text{total}}$ is computed as shown in Equation (1):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \tag{1}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss for each proposed region, computed using cross-entropy loss over the softmax outputs for foreground/background (nodule vs. non-nodule); $\mathcal{L}_{\text{box}}$ is the bounding box regression loss, defined using smooth ℓ_1 loss, which balances numerical stability and robustness to outliers. The element-wise formulation is illustrated in Equation (2); and $\mathcal{L}_{\text{mask}}$ is the pixel-wise binary cross-entropy loss applied to predicted masks and their corresponding GT masks, computed only for positively classified regions, which are RoIs classified as nodules. $\mathcal{L}_{\text{aux}}$ is the binary cross-entropy loss for the auxiliary classifier, computed on proposals with IoU ≥ 0.3, while $\lambda_{\text{aux}} = 0.3$ is the weighting factor for auxiliary loss:

$$\text{smooth}_{\ell_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}. \tag{2}$$

Each component is weighted equally in the official implementation, and the framework automatically balances these components without requiring additional manual hyperparameter tuning.

The built-in loss computation in the `torchvision` Mask R-CNN class automatically performs anchor matching, positive/negative proposal sampling, and per-instance mask supervision based on the predicted RoIs and GT. This enables a unified training of both detection and segmentation tasks with minimal manual setup.

In this study, an additional auxiliary classification branch is introduced after the RoI feature extraction stage to improve recall. The auxiliary classifier predicts whether each proposal corresponds to a true nodule, with binary labels assigned based on an IoU threshold of 0.3. The corresponding auxiliary loss $\mathcal{L}_{\text{aux}}$ is defined as the binary cross-entropy loss between predicted logits and these labels. The final composite loss, including $\mathcal{L}_{\text{aux}}$ and its weighting, is detailed in Section 3.5.

3.4. Evaluation Metrics

The evaluation of the proposed multi-task framework includes both IQA during preprocessing and task-specific performance metrics for detection and segmentation as described in Section 2.4.

3.4.1. Image Quality Assessment (IQA)

To assess the visual quality of preprocessed CT images, both NR and FR metrics are applied. Specifically, BRISQUE is computed to quantify the perceptual quality at the slice level, while PSNR measures pixel-wise fidelity between the original and CLAHE-enhanced images. These evaluations are conducted slice-wise and averaged across the dataset to determine whether enhancement introduces visual trade-offs and how it changes structure visibility.

3.4.2. Detection and Segmentation Metrics

Detection performance is evaluated with precision, recall, and F_1-score, using nodule-level true positive (TP), false positive (FP), and false negative (FN) counts after 3D prediction merging. Unlike the official LUNA16 evaluation protocol, which evaluates instance-level nodule detection performance under scan-normalized false-positive operating points (CPM/FROC), these metrics are adopted to support fixed-threshold nodule-level evaluation within a unified detection–segmentation framework.

Segmentation performance is measured using DSC and IoU with respect to the box-derived geometric masks. These metrics quantify the spatial consistency between predicted instance masks and their approximate supervision, rather than evaluating the fine-grained anatomical delineation of true nodule morphology. These are calculated by matching

predicted masks with ground truth masks using the Hungarian algorithm based on IoU thresholding as implemented in the `match_masks` function. Only matched pairs with IoU ≥ 0.25 contribute to the averaged Dice and IoU metrics. All formulas and detailed explanations of these metrics can be found in Tables 5 and 6.

3.4.3. Nodule-Level Evaluation Strategy

Although the model performs detection and segmentation on 2.5D slices, final evaluation is conducted at the 3D nodule level. To achieve this, slice-level predictions are merged across adjacent slices based on spatial overlap and z-axis proximity. Specifically, bounding boxes with sufficient spatial overlap and centre slice indices within a 3-slice adjacent window are grouped as a single nodule prediction. Here, the 3-slice tolerance is adopted as a pragmatic heuristic, considering the typical CT slice spacing in LUNA16 (approximately 1–1.25 mm) and the limited through-plane extent of many small nodules.

A merged prediction is considered a TP if its centre slice lies within three slices of a GT nodule and the 2D bounding box has an IoU greater than 0.2 with the annotated location. This tolerance is dataset- and protocol dependent and is not claimed to be a universally optimal setting. This post-processing strategy mitigates redundant detections across slices and yields a more clinically meaningful evaluation. Slice-level metrics are not reported separately, as all predictions are aggregated before scoring. As a result, the reported nodule-level precision, recall, and F_1-score are not directly comparable to CPM values reported under the official LUNA16 evaluation protocol.

This strategy ensures a realistic evaluation that reflects clinical expectations: nodules are volumetric and not isolated 2D events. It also helps reduce the overcounting of TP when nodules appear in multiple slices.

During training, Dice and IoU are monitored for early stopping and model selection, and all detections are aggregated at the 3D nodule level before scoring to avoid inflating true positives. Classification metrics such as precision, recall, and F_1-score are primarily used to evaluate the model's detection performance at the nodule level, with predictions matched to ground truth nodules based on spatial proximity and bounding box overlap.

3.5. Training Strategy

The direct joint training of both detection and segmentation heads initially led to poor detection performance, as the model tended to prioritize mask learning. To address this imbalance, a three-phase training strategy was adopted. To mitigate this, the training was staged as follows:

- **Phase 1: Detection-Only Training.** The mask head was frozen, and only the detection components were trained. This phase established strong detection weights without being influenced by segmentation gradients.
- **Phase 2: Segmentation-Only Training.** The detection branch was frozen, and the mask head was fine-tuned to learn instance masks based on stable RoI proposals.
- **Phase 3: Joint Fine-Tuning.** Both branches were unfrozen and jointly optimised, starting from previously trained weights. The number of epochs in each phase was iteratively adjusted based on validation trends to achieve optimal performance.

In addition, the auxiliary loss weight λ_{aux} was manually adjusted based on observed precision–recall trade-offs across the training runs. The final setting of $\lambda_{\text{aux}} = 0.3$ offered a good balance between promoting recall and avoiding excessive FPs.

Lastly, ResNet-50 was selected as the backbone instead of ResNet-34, as it provides deeper and more expressive feature representations while remaining computationally feasible for 2.5D inputs. Empirical trials showed that ResNet-50 led to better detection performance and more stable training, especially when combined with the FPN structure.

Across the three phases, the optimiser and hyperparameters remain fixed; however, different subnetworks are enabled/disabled according to the staged schedule described above. In the final joint fine-tuning stage, all components are optimised together using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4. The slice-level dataset provides 2.5D CT inputs consisting of five adjacent slices stacked channel-wise. All components of the architecture, including the backbone, RPN, detection head, mask head, and auxiliary RoI classifier, are jointly optimised. Mixed precision training is enabled via PyTorch 2.7.1 AMP to improve training speed and reduce memory usage.

The loss function defined in Equation (1) is optimised jointly using the Adam optimiser. The auxiliary loss promotes the retention of borderline proposals to improve recall.

Training uses the modified anchor generator described in Section 3.3, with scales starting from 4 pixels to improve small-nodule coverage. The RPN is configured to generate up to 2000 proposals per image during training to maximise candidate coverage before suppression.

During validation, low-confidence predictions are filtered using a dynamic score threshold determined by the 30th percentile of the predicted confidence scores, clipped to the range $[0.1, 0.5]$. This adaptive thresholding retains a sufficient number of potential nodules while suppressing obvious FPs, complementing the auxiliary classifier's ability to re-rank borderline detections.

Both detection (F_1-score) and segmentation (Dice) metrics are monitored on the validation set after each epoch. Training is stopped early if neither metric improves for 5 consecutive epochs. The best-performing models for each metric are checkpointed separately.

4. Results and Discussions

To validate the effectiveness of the proposed framework, this section presents a series of experiments designed to evaluate image quality changes, detection and segmentation performance, and the contribution of individual architectural components. Results are structured to reflect the logical progression from preprocessing evaluation to end-task performance and error analysis. First, the effectiveness of the preprocessing pipeline, especially the use of CLAHE enhancement, is assessed using both FR and NR image quality metrics. Second, the performance of the model is quantitatively analysed across detection and segmentation tasks on the validation set, with a particular focus on the interplay between precision, recall, and segmentation accuracy during different training phases. Finally, qualitative case studies are provided to illustrate typical successes and failure modes, highlighting the strengths and limitations of the proposed method in realistic clinical scenarios.

4.1. Image Enhancement and Quality Assessment

To assess the impact of image enhancement on CT image quality prior to model training, CLAHE was applied to each 2D slice in the dataset. Both NR and FR IQA metrics were computed to evaluate their effectiveness. These scores are averaged across 89 CT scans in the subset0. Visual inspection, quantitative comparisons, and statistical testing were conducted. As shown in Figure 6, CLAHE redistributes local intensity contrast and can make certain boundaries visually more salient. However, it remains uncertain whether such visual changes translate into improved semantic signal for nodule analysis. Importantly, higher entropy does not necessarily imply improved semantic information, as CLAHE may also amplify noise or vascular textures in homogeneous lung regions. This risk is explicitly examined in downstream detection behaviour rather than assumed a priori.

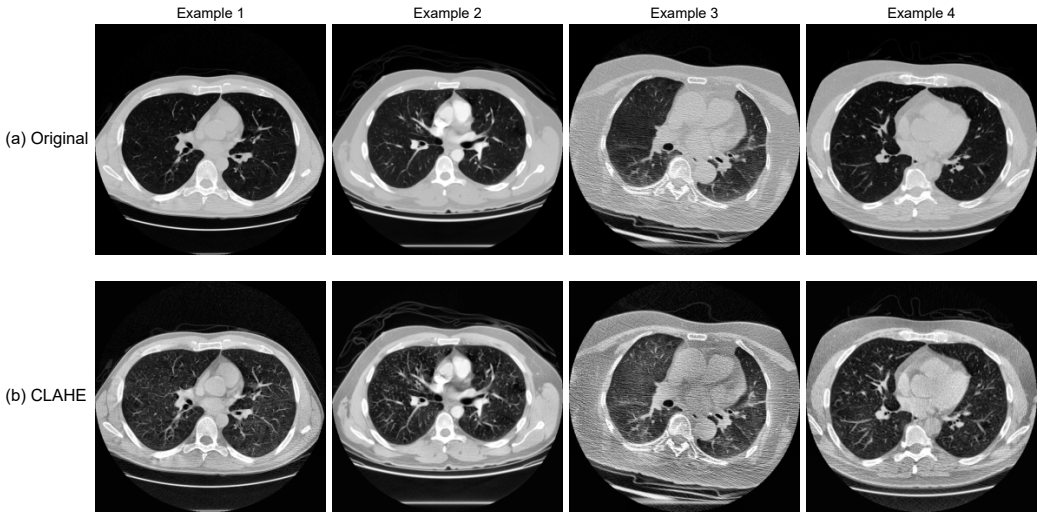


Figure 6. CT image enhancement visual comparison: (**a**) original slice (PNG-normalised) and (**b**) CLAHE-enhanced slices.

4.1.1. PSNR-Based Evaluation

To quantify the structural similarity between the original and enhanced images, PSNR was calculated on a per-scan basis. The mean PSNR was 22.63 dB, indicating that although contrast is altered, some local structure differences are introduced. The distribution is shown in Figure 7. While a PSNR above 20 dB generally indicates that major anatomical structures are preserved, this also reflects that CLAHE introduces moderate pixel-level changes. This outcome is expected, as contrast enhancement inevitably alters local intensity distributions, even when global structures are preserved.

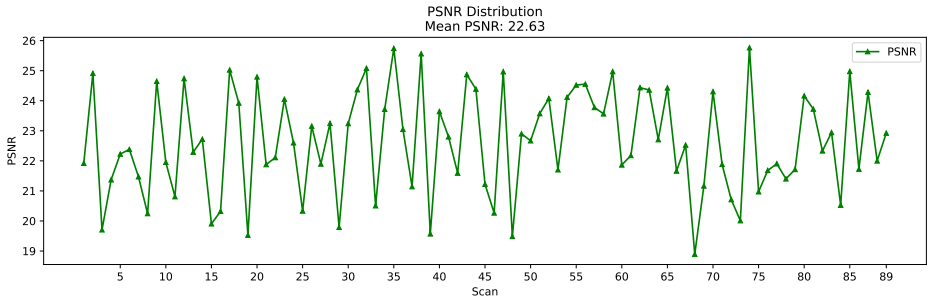


Figure 7. Per-scan PSNR comparison: original vs. CLAHE-enhanced slices.

4.1.2. BRISQUE-Based Evaluation

For perceptual quality evaluation, BRISQUE scores were computed before and after CLAHE enhancement. A per-scan comparison is provided in Figure 8, where the average BRISQUE for original slices is 47.10 and 49.93 for CLAHE-enhanced slices, indicating a decline in perceptual quality. This result highlights the perceptual trade-off of CLAHE, as local contrast enhancement may introduce noise-like textures and over-enhancement artefacts that are penalised by BRISQUE. However, BRISQUE penalises local over-enhancement and noise, which may explain the higher scores. This motivated the addition of an alternative NR metric to cross-validate this observation.

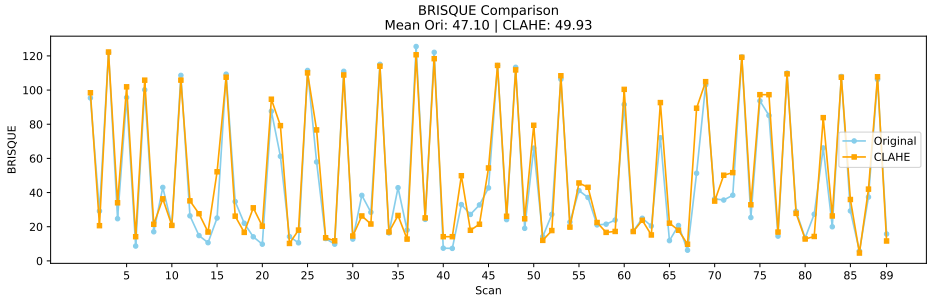


Figure 8. Per-scan BRISQUE comparison: Original vs. CLAHE-enhanced slices.

4.1.3. Entropy-Based Evaluation

Entropy was introduced as a complementary NR metric to assess global intensity variation. As shown in Figures 9 and 10, CLAHE significantly increased entropy from 4.10 to 4.51 on average, reflecting increased intensity variability. However, higher entropy is not sufficient evidence of improved semantic information, and may also be consistent with noise amplification or enhanced vascular textures.

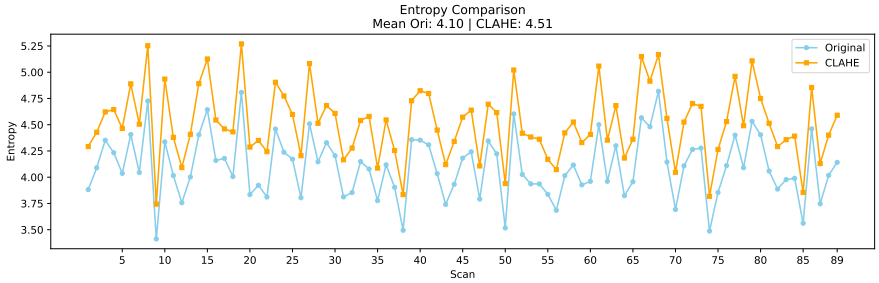


Figure 9. Per-scan entropy comparison: original vs. CLAHE-enhanced slices.

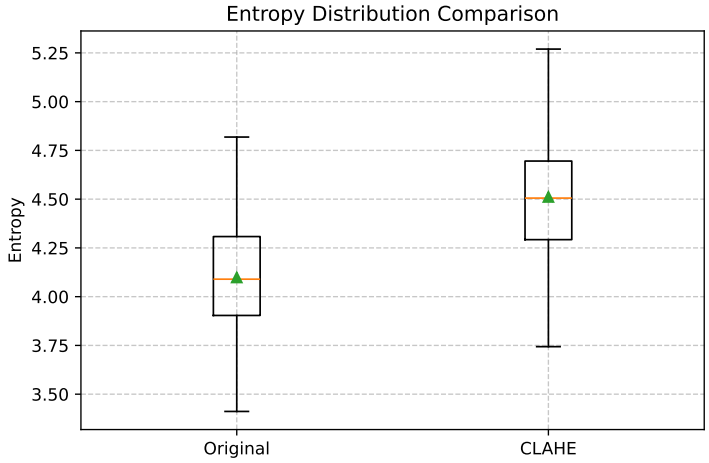


Figure 10. Entropy distribution across all scans.

4.1.4. Paired *t*-Test Results and Conclusion

Paired *t*-tests were conducted to evaluate the statistical significance of the observed differences:

- **BRISQUE**: $t = -3.0053$, $p = 3.46 \times 10^{-3} \rightarrow$ significant
- **Entropy**: $t = -58.5689$, $p = 2.83 \times 10^{-72} \rightarrow$ significant

While BRISQUE scores increased slightly after CLAHE enhancement, entropy analysis indicated a redistribution of gray-level intensities, reflecting increased variability but

also potential noise amplification, while PSNR confirmed that structural information was largely preserved. Overall, the IQA results indicate a clear perceptual trade-off: BRISQUE worsened after CLAHE, while entropy increased due to greater intensity variability, which may also reflect noise amplification or enhanced non-nodular structures. Therefore, CLAHE is not presented as a general-purpose image quality improvement.

Importantly, the decision to retain CLAHE in subsequent experiments is based on the task-level validation results in Section 4.3.2, where CLAHE yields higher recall (+0.08) and F_1-score (+0.04) under the same training and evaluation protocol. Accordingly, CLAHE is treated as a recall-oriented, task-specific preprocessing option, and its potential contribution to false positives is explicitly acknowledged.

4.2. Preprocessing Visualisation

To illustrate the preprocessing steps prior to model training, Figure 11 shows two sample CT slices at three key stages. Figure 11a is the original CT slices, displayed without any enhancement, to represent the raw input before preprocessing. Figure 11b presents the enhanced results of the original images. Figure 11c shows the corresponding results after applying the lung mask, which removes irrelevant anatomical structures and focuses the input on the lung parenchyma.

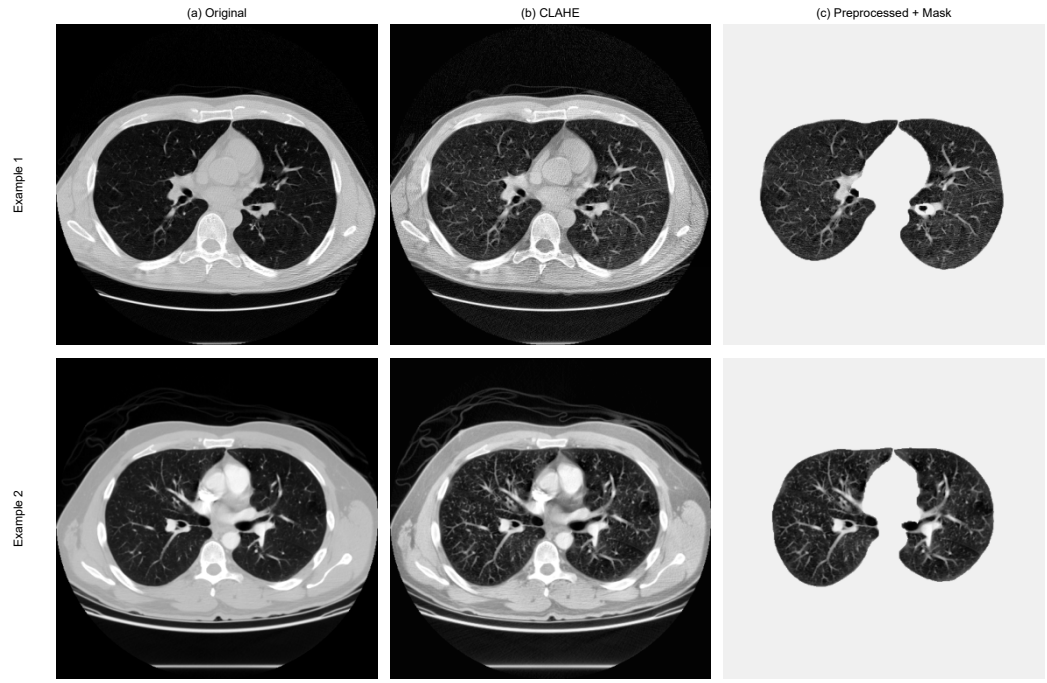


Figure 11. CT preprocessing stages: (**a**) original CT slice, (**b**) CLAHE-enhanced image, and (**c**) final input after applying lung mask to the enhanced image.

Note: The image shown in (c) is the CLAHE-enhanced CT slice after lung masking. This version was selected as the final model input, following the quality assessment results discussed in Section 4.1.

This visualisation illustrates the preprocessing pipeline and the contrast redistribution introduced by CLAHE. Nevertheless, CLAHE may also enhance non-nodular structures (e.g., vessels), and its potential impact on false positives is considered in the downstream error analysis. These processed volumes are then used for training the multi-task model described in the next section.

4.3. Multi-Task Model Performance

This section presents the performance of the proposed MTL framework on both lung nodule detection and segmentation tasks. Evaluation is performed on the validation subset derived from LUNA16 subset0, using the evaluation metrics described in Section 3.

4.3.1. Detection and Segmentation Performance

Prior to reporting validation metrics, the staged optimisation schedule is verified using the total training loss in Figure 12. Vertical markers at epochs 10 and 15 denote transitions from detection-only to segmentation-only and then to joint training. The loss decreases steadily once the segmentation phase begins and stabilises during joint training, indicating that optimisation behaves as designed. Consequently, metrics recorded in epochs 1–14 reflect single-branch training and should not be interpreted as end-to-end MTL performance. This motivates the focus on epochs $\geq$15 for evaluating the full MTL configuration, including comparison between original and CLAHE-enhanced inputs and when selecting representative checkpoints.

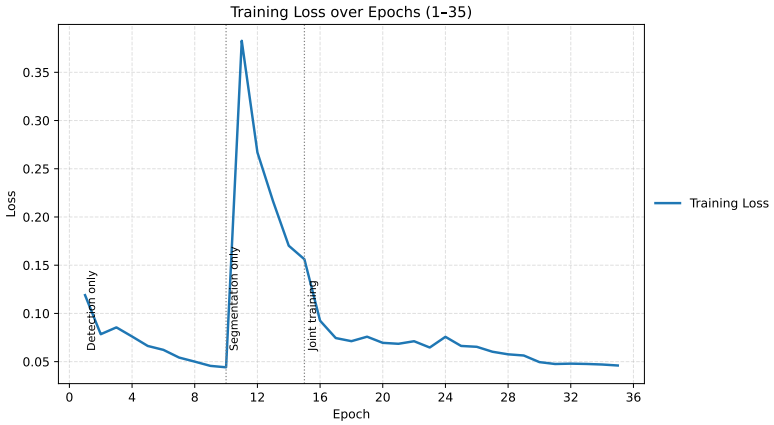


Figure 12. Total training loss across 35 epochs with stage boundaries (detection-only, segmentation-only, and joint training).

Figure 13 illustrates the evolution of precision, recall, F_1-score, Dice coefficient, and IoU for the CLAHE-enhanced input. An early peak in F_1-score ($\approx$0.77) appears around epoch 6 when only the detection branch is trained, while Dice and IoU remain low (<0.50) because the mask head is frozen. During the segmentation-only phase (epochs 10–14), Dice and IoU increase steadily to $\approx$0.78 and $\approx$0.67, whereas F_1-score drops due to reduced recall. Once joint training begins at epoch 15, detection- and segmentation-related metrics improve jointly, reaching a balanced performance at epoch 19. After epoch 25, recall steadily declines while precision increases, indicating a shift towards more conservative predictions. By the final epoch, Dice reaches 0.82, and IoU 0.72, but the F_1-score decreases to 0.43 due to low recall.

Table 8 summarises validation performance at representative epochs; epoch 19 provides the best balance (F_1-score = 0.70; *Dice* = 0.81). Qualitative examples in Figure 14 show precise boundaries across sizes and challenging locations.

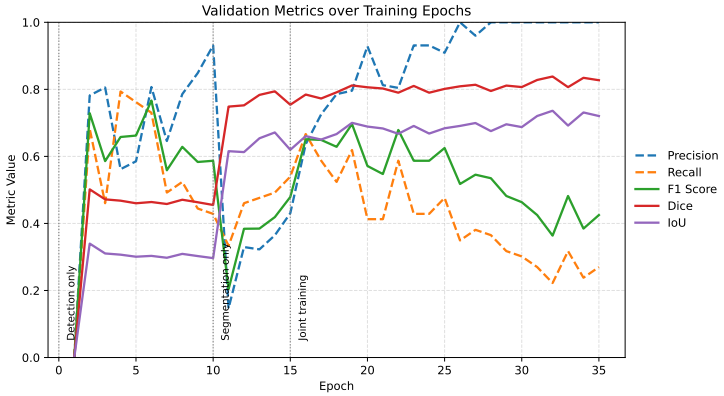


Figure 13. Validation metrics over training epochs for CLAHE-enhanced CT images (lung region only).

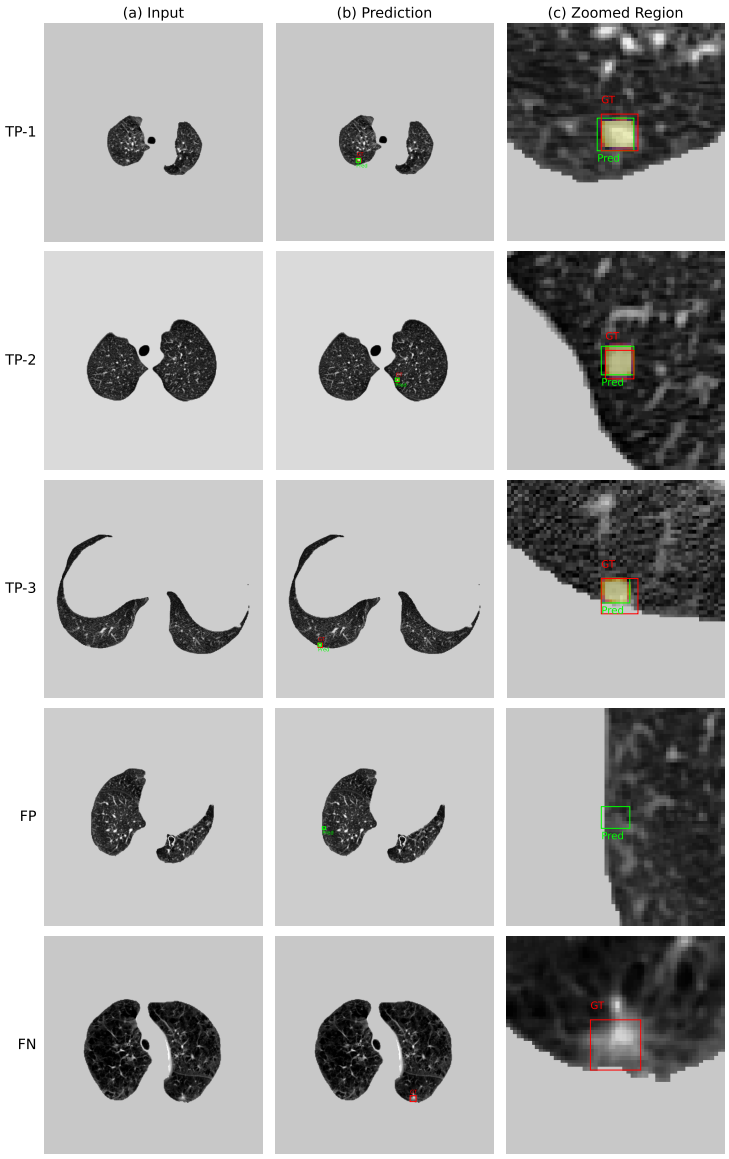


Figure 14. Qualitative results of the proposed Mask R-CNN-based MTL framework for lung nodule detection and segmentation. (**a**) Input CT slice after preprocessing; (**b**) predicted bounding boxes (green) and masks with ground-truth boxes (red); (**c**) zoomed view of the nodule region. Rows show true positives (TP-1, TP-2, TP-3), a false positive (FP), and a false negative (FN).

Table 8. Validation performance of the proposed MTL model at selected epochs. (Epoch 19 exhibits the best balance between detection and segmentation).

Epoch	TP	FP	FN	Precision	Recall	F_1-Score	Dice	IoU
6	46	11	17	0.81	0.73	0.77	0.46	0.30
15	34	45	29	0.43	0.54	0.48	0.76	0.64
19	39	10	24	0.80	0.62	0.70	0.81	0.70
35	17	0	46	1.00	0.27	0.43	0.82	0.72

The first three rows correspond to true positives (TP-1, TP-2, and TP-3) and demonstrate successful detections across a variety of nodule sizes, positions, and appearances. Notably, TP-3 features a boundary-adjacent nodule, where shape irregularities and partial-volume effects often hinder both detection and segmentation; in this case, predicted and ground-truth boxes align closely, and the mask accurately follows the annotated region. The fourth row (FP) shows a vascular bifurcation misclassified as a nodule—an anatomically complex region where texture and intensity resemble true nodules. The fifth row (FN) depicts a very small sub-solid nodule with low contrast; no prediction met the detection criteria after merging, leaving the case undetected.

4.3.2. Validation of Enhancement Effectiveness

To evaluate whether CLAHE enhancement improves model performance, a comparative experiment was conducted using original (non-enhanced) CT inputs under the same preprocessing pipeline, data splits, and multi-stage training schedule described earlier. The only difference between the two settings was whether CLAHE was applied.

For a fair comparison, only epochs beyond 15 (i.e., the joint training stage) were considered, since earlier epochs exclusively trained either the detection or the segmentation branch. Figure 15 presents the validation dynamics for the original input.

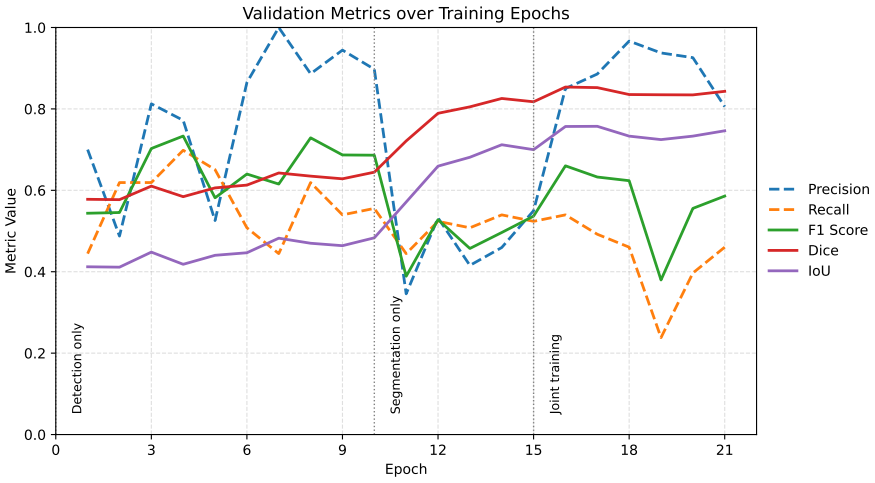


Figure 15. Validation metrics on original CT images (with lung mask) over epochs.

For the original CT input, epoch 16 was selected as the point of comparison because it yielded the highest F_1-score (0.66) after the start of joint training, balancing precision and recall while segmentation metrics had stabilised. For the CLAHE-enhanced input, epoch 19 was chosen as the comparison point, as it represented the best trade-off between detection sensitivity and segmentation accuracy (F_1-score = 0.70, *Dice* = 0.81).

The comparison in Table 9 indicates a clear performance gain from CLAHE enhancement. Although the original input achieves slightly higher precision (0.85 vs. 0.80) and marginally higher Dice and IoU at its best joint-training checkpoint (epoch 16), the CLAHE-

enhanced input delivers a higher recall (+0.08) and a higher F_1-score (+0.04). In clinical contexts, where missing a true nodule is typically more consequential than an occasional FP, the recall improvement is particularly meaningful. Nevertheless, this improvement should be interpreted with caution, as CLAHE may also enhance non-nodular structures such as vessels, potentially contributing to false positives under certain conditions. We emphasise that this gain is observed under the current subset0 setting and evaluation protocol; CLAHE may not generalise to other datasets or acquisition conditions, and its tendency to enhance vascular structures may increase false positives.

Table 9. Comparison of best validation performance after epoch 15 between original and CLAHE-enhanced inputs.

Input Type	Epoch	Precision	Recall	F_1-Score	Dice	IoU
Original CT	16	0.85	0.54	0.66	0.85	0.76
CLAHE-enhanced CT	19	0.80	0.62	0.70	0.81	0.70

From a task perspective, the recall gain suggests that CLAHE increases the detectability of small or low-contrast nodules and those near anatomical boundaries. This aligns with the entropy-based contrast analysis in Section 4.1, and the improvement in F_1-score confirms that the visual changes translate into measurable benefits within the integrated detection–segmentation framework.

Based on these findings, CLAHE-enhanced images are used as the default input for all subsequent experiments. All ablations, qualitative visualisations, and error analyses are therefore reported using the CLAHE-enhanced setting.

4.3.3. Ablation Study

To overcome the limitations of the baseline configuration, several modifications were introduced: expanding the input to 5 slices, optimising anchor settings, incorporating an auxiliary classifier, and applying a staged training strategy. The baseline model—a standard Mask R-CNN with 3-slice input, default anchors, and no auxiliary classifier—achieved a maximum F_1-score of 0.42 at epoch 24, with recall limited to 0.29 despite relatively high precision (0.78). These results underscored the need for architectural and training adjustments.

Table 10 compares the baseline with the final model. At epoch 19, the final model reached an F_1-score of 0.70, with a precision of 0.80 and a recall of 0.62. Dice and IoU also remained high at 0.81 and 0.70, confirming that segmentation performance was not compromised. In contrast, the baseline model at epoch 17, despite similar precision, suffered from poor recall and low F_1-score.

Table 10. Ablation study comparing the baseline and final models.

Model	TP	FP	FN	Precision	Recall	F_1-Score	Dice	IoU
Baseline (3-slice, default)	18	5	45	0.78	0.29	0.42	0.84	0.73
Final Model (5-slice + anchor + aux)	39	10	24	0.80	0.62	0.70	0.81	0.70

The baseline results are taken from epoch 17, which yielded the best F_1-score (0.42) under the default configuration. The final model results are taken from epoch 19, identified as the performance peak in the main experiment (see Table 8).

All experiments used the same train/validation split, training schedule, and evaluation protocol to ensure fair comparison. The improvements observed are the result of cumulative refinements; no single change alone accounted for the performance gain. The final design

offers a well-balanced trade-off between detection sensitivity and segmentation accuracy, supporting its suitability for clinical deployment.

4.4. Limitations

Despite the improvements achieved by the proposed MTL framework, several limitations should be acknowledged. (1) All experiments in this study are conducted on the LUNA16 subset0, which is intentionally used as a controlled experimental subset rather than a large-scale benchmarking dataset. As a result, the reported performance should not be interpreted as statistically representative of the entire LUNA16 dataset but rather as reflecting the relative effectiveness of different architectural components and training strategies under a fixed experimental setting. In addition, the 2.5D contextual formulation may limit generalisation and recall for extremely small or low-contrast lesions, and may introduce inter-slice inconsistency (or "flickering") along the z-axis, as predictions are generated in a slice-wise manner. In this study, slice stacking and nodule-level aggregation partially mitigate this effect, but no explicit inter-slice smoothness modelling is incorporated. (2) Segmentation supervision is derived from nodule centre coordinates and diameter, resulting in simplified geometric masks rather than voxel-level annotations of true nodule morphology (e.g., spiculation or lobulation). Consequently, Dice and IoU scores may overestimate segmentation quality and should not be interpreted as measuring accurate anatomical delineation. In addition, contrast enhancement via CLAHE may amplify vascular textures or local noise in homogeneous lung regions, potentially contributing to false positives under certain conditions. (3) The nodule-level protocol is lightweight and not identical to CPM (Competition Performance Metric, the official evaluation metric of LUNA16). This choice prioritises fixed-threshold instance-level consistency analysis rather than scan-normalized multi-operating-point benchmarking, and therefore does not support direct performance comparison or competitive benchmarking against CPM/FROC-based methods. Future work will scale to full LUNA16 and external cohorts, adopt voxel-level masks and standard FROC (Free-response ROC)/CPM, and explore hybrid 2D-3D designs. Moreover, most baseline methods do not report parameter count or FLOPs, preventing exhaustive numerical comparison.

4.5. Comparison with Published Baselines

To contextualise our results, we summarise representative published baselines spanning (i) detection-only, (ii) segmentation-only, and (iii) joint or cascaded pipelines. Most prior works report sensitivity/FROC for detection, whereas we adopt the F_1-score to reflect the balance between sensitivity and precision under imbalanced data. Scope notes (dataset coverage and input dimensionality) are provided with the entries. Where detection metrics differ, we qualitatively align conclusions by focusing on performance trends (e.g., whether a method's strength lies in a high true positive rate or balanced accuracy) rather than numerical values, ensuring fair comparison while preserving each method's original reporting.

As summarised in Table 11, our 2.5D MTL achieves a balanced performance between detection and segmentation, reflecting a design trade-off between limited through-plane contextual aggregation across adjacent slices and architectural simplicity, compared with full volumetric 3D models. We also discuss in Section 4.6 why unified MTL can mitigate error propagation compared with cascaded detect-then-segment pipelines such as NoduleNet [37]. A controlled re-implementation of two-stage and single-shot detectors under an identical protocol on the same subset is planned as future work.

Table 11. Comparison with representative baselines. Each entry is reported in its native metric. Note: LUNA16 is curated from LIDC-IDRI; figures from LIDC-IDRI are provided for trend-level context (not direct numeric comparison). Detection metrics vary (sensitivity/FROC vs. F_1-score).

Model	Task	Dataset	Metric(s)	Reported Value
Faster R-CNN Variant [20]	Detection (2D)	LUNA16	Sens/AUC	86.4/95.4%
YOLO-MSRF [24]	Detection (2D)	LUNA16	Sens	94.02%
Multilevel 3D CNN [21]	Detection (3D)	LUNA16	Sens	94.4%
Improved U-Net [27]	Segmentation (2D)	LUNA16	Dice	73.6%
DEHA-Net [33]	Segmentation (2D)	LIDC-IDRI	Dice	87.91%
CSE-GAN [32]	Segmentation (3D)	LUNA16	Dice	80.7%
NoduleNet [37]	Cascaded (Detect→Seg)	LIDC-IDRI	FROC/Dice	+10.3% vs. single-task/83.1%
ECANodule [13]	Hybrid (Detect + Seg)	LIDC-IDRI	FROC/Dice	91.1/83.4%
Ours (2.5D MTL)	**MTL (Detect + Seg)**	**LUNA16 subset0**	**F_1-score/Dice**	**0.70/0.81**

4.6. Discussion and Implications

Unified MTL integrates detection and segmentation within a single optimisation process. This joint learning mitigates error propagation from detector mislocalisation and lets segmentation cues refine detection features.

Cascaded systems like NoduleNet [37] rely on sequential training and suffer from weak gradient coupling, while single-task U-Net-based or detect-specific models perform well individually but remain suboptimal for joint localisation and delineation.

Our 2.5D formulation retains limited volumetric context while reducing architectural complexity compared with full 3D models, as it avoids volumetric convolutions and duplicated encoders. This represents a structural trade-off between through-plane context and model complexity, rather than an empirically validated efficiency advantage. Future work will reproduce two-stage baselines and extend evaluation to full LUNA16.

5. Conclusions

This study proposes a 2.5D multi-task framework that jointly detects and segments pulmonary nodules on CT. A tailored preprocessing pipeline (HU windowing, CLAHE, and lung masking) feeds a Mask R-CNN adapted to stacked slices; small-nodule anchors and a compact auxiliary RoI classifier help retain borderline positives. Predictions are aggregated across slices and scored at the nodule level. On the LUNA16 subset0, the approach achieves a reasonable precision–recall and segmentation trade-off under a controlled experimental setting, supporting the effectiveness of the proposed design choices.

Limitations include restricted long-range context in the 2.5D setting, evaluation on a single experimental subset, dynamic score thresholds that can suppress weak but valid detections, and box-derived masks; we also do not report CPM under the official 7-point protocol. Future work will scale to multi-centre datasets, adopt voxel-level masks and standard FROC/CPM evaluation, and explore full-3D or hybrid 2D–3D architectures to further boost small-nodule sensitivity.

Author Contributions: Conceptualisation, R.L. and B.H.S.A.; methodology, R.L. and B.H.S.A.; resources, B.H.S.A. and R.L.; writing—original draft preparation, B.H.S.A. and R.L.; writing—review and editing, B.H.S.A. and R.L.; supervision, B.H.S.A.; visualisation, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant number: SJCX24_0130). The APC was not funded by any external sources.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cancer Research UK. Lung Cancer Statistics. 2025. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer#lung_stats1 (accessed on 5 June 2025).
2. The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **2011**, *365*, 395–409. [CrossRef]
3. MacMahon, H.; Naidich, D.P.; Goo, J.M.; Lee, K.S.; Leung, A.N.; Mayo, J.R.; Mehta, A.C.; Ohno, Y.; Powell, C.A.; Prokop, M.; et al. Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017. *Radiology* **2017**, *284*, 228–243. [CrossRef] [PubMed]
4. Riquelme, D.; Akhloufi, M.A. Deep learning for lung cancer nodules detection and classification in CT scans. *AI* **2020**, *1*, 28–67. [CrossRef]
5. Marinakis, I.; Karampidis, K.; Papadourakis, G. Pulmonary Nodule Detection, Segmentation and Classification Using Deep Learning: A Comprehensive Literature Review. *BioMedInformatics* **2024**, *4*, 2043–2106. [CrossRef]
6. Wang, Y.; Mustaza, S.M.; Ab-Rahman, M.S. Pulmonary Nodule Segmentation using Deep Learning: A Review. *IEEE Access* **2024**, *12*, 119039–119055. [CrossRef]
7. Setio, A.A.A.; Traverso, A.; De Bel, T.; Berens, M.S.; Van Den Bogaard, C.; Cerello, P.; Chen, H.; Dou, Q.; Fantacci, M.E.; Geurts, B.; et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Anal.* **2017**, *42*, 1–13. [CrossRef]
8. El-Baz, A.; Beache, G.M.; Gimel farb, G.; Suzuki, K.; Okada, K.; Elnakib, A.; Soliman, A.; Abdollahi, B. Computer-aided diagnosis systems for lung cancer: Challenges and methodologies. *Int. J. Biomed. Imaging* **2013**, *2013*, 942353. [CrossRef]
9. Zhang, G.; Jiang, S.; Yang, Z.; Gong, L.; Ma, X.; Zhou, Z.; Bao, C.; Liu, Q. Automatic nodule detection for lung cancer in CT images: A review. *Comput. Biol. Med.* **2018**, *103*, 287–300. [CrossRef]
10. Zheng, L.; Lei, Y. A review of image segmentation methods for lung nodule detection based on computed tomography images. *Matec Web Conf.* **2018**, *232*, 02001. [CrossRef]
11. Pehrson, L.M.; Nielsen, M.B.; Ammitzbøl Lauridsen, C. Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: A systematic review. *Diagnostics* **2019**, *9*, 29. [CrossRef]
12. Alshayeji, M.H.; Abed, S. Lung cancer classification and identification framework with automatic nodule segmentation screening using machine learning. *Appl. Intell.* **2023**, *53*, 19724–19741. [CrossRef]
13. Luo, D.; He, Q.; Ma, M.; Yan, K.; Liu, D.; Wang, P. ECANodule: Accurate Pulmonary Nodule Detection and Segmentation with Efficient Channel Attention. In *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
14. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. [CrossRef]
15. Thung, K.H.; Wee, C.Y. A brief review on multi-task learning. *Multimed. Tools Appl.* **2018**, *77*, 29705–29725. [CrossRef]
16. Zhao, Y.; Wang, X.; Che, T.; Bao, G.; Li, S. Multi-task deep learning for medical image computing and analysis: A review. *Comput. Biol. Med.* **2023**, *153*, 106496. [CrossRef]
17. Li, R.; Xiao, C.; Huang, Y.; Hassan, H.; Huang, B. Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: A review. *Diagnostics* **2022**, *12*, 298. [CrossRef]
18. Li, R.; Honarvar Shakibaei Asli, B. Multi-Task Deep Learning for Lung Nodule Detection and Segmentation in CT Scans: A Review. *Electronics* **2025**, *14*, 3009. [CrossRef]
19. George, J.; Skaria, S. Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans. In *Medical Imaging 2018: Computer-Aided Diagnosis*; SPIE: Bellingham, WA, USA, 2018; Volume 10575, pp. 347–355.
20. Xie, H.; Yang, D.; Sun, N.; Chen, Z.; Zhang, Y. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognit.* **2019**, *85*, 109–119. [CrossRef]
21. Dou, Q.; Chen, H.; Yu, L.; Qin, J.; Heng, P.A. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 1558–1567. [CrossRef]
22. Anirudh, R.; Thiagarajan, J.J.; Bremer, T.; Kim, H. Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data. In *Medical Imaging 2016: Computer-Aided Diagnosis*; SPIE: Bellingham, WA, USA, 2016; Volume 9785, pp. 791–796.
23. Paul, R.; Hawkins, S.H.; Schabath, M.B.; Gillies, R.J.; Hall, L.O.; Goldgof, D.B. Predicting malignant nodules by fusing deep features with classical radiomics features. *J. Med. Imaging* **2018**, *5*, 011021. [CrossRef] [PubMed]
24. Wu, X.; Zhang, H.; Sun, J.; Wang, S.; Zhang, Y. YOLO-MSRF for lung nodule detection. *Biomed. Signal Process. Control* **2024**, *94*, 106318. [CrossRef]
25. Tang, H.; Li, Z.; Zhang, D.; He, S.; Tang, J. Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 1958–1974. [CrossRef]

26. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. YOLOv9: Learning what you want to learn using programmable gradient information. In *Computer Vision—ECCV 2024, Proceedings of the 18th European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: Cham, Switzerland, 2024; pp. 1–21.
27. Tong, G.; Li, Y.; Chen, H.; Zhang, Q.; Jiang, H. Improved U-NET network for pulmonary nodules segmentation. *Optik* **2018**, *174*, 460–469. [CrossRef]
28. Ma, X.; Song, H.; Jia, X.; Wang, Z. An improved V-Net lung nodule segmentation model based on pixel threshold separation and attention mechanism. *Sci. Rep.* **2024**, *14*, 4743. [CrossRef]
29. Cao, H.; Liu, H.; Song, E.; Hung, C.C.; Ma, G.; Xu, X.; Jin, R.; Lu, J. Dual-branch residual network for lung nodule segmentation. *Appl. Soft Comput.* **2020**, *86*, 105934. [CrossRef]
30. Wang, S.; Zhou, M.; Gevaert, O.; Tang, Z.; Dong, D.; Liu, Z.; Jie, T. A multi-view deep convolutional neural networks for lung nodule segmentation. In *Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Republic of Korea, 11–15 July 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 1752–1755.
31. Liu, C.; Liu, H.; Zhang, X.; Guo, J.; Lv, P. Multi-scale and multi-view network for lung tumor segmentation. *Comput. Biol. Med.* **2024**, *172*, 108250. [CrossRef]
32. Tyagi, S.; Talbar, S.N. CSE-GAN: A 3D conditional generative adversarial network with concurrent squeeze-and-excitation blocks for lung nodule segmentation. *Comput. Biol. Med.* **2022**, *147*, 105781. [CrossRef]
33. Usman, M.; Shin, Y.G. Deha-net: A dual-encoder-based hard attention network with an adaptive roi mechanism for lung nodule segmentation. *Sensors* **2023**, *23*, 1989. [CrossRef] [PubMed]
34. Li, F.; Xu, L.; Ma, Z.; Zhao, Y.; Li, X. APU-Net: A U-Net Enhanced Network with Dynamic Feature Fusion and Pyramid Cross-Attention Mechanism for Polyp Segmentation. *Digit. Signal Process.* **2026**, *172*, 105879. [CrossRef]
35. Hu, T.; Lan, Y.; Zhang, Y.; Xu, J.; Li, S.; Hung, C.C. A lung nodule segmentation model based on the transformer with multiple thresholds and coordinate attention. *Sci. Rep.* **2024**, *14*, 31743. [CrossRef] [PubMed]
36. Yadav, D.P.; Sharma, B.; Webber, J.L.; Mehbodniya, A.; Chauhan, S. EDTNet: A spatial aware attention-based transformer for the pulmonary nodule segmentation. *PLoS ONE* **2024**, *19*, e0311080. [CrossRef]
37. Tang, H.; Zhang, C.; Xie, X. Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019, Proceedings, Part VI*; Springer: Cham, Switzerland, 2019; pp. 266–274.
38. Song, G.; Nie, Y.; Zhang, J.; Chen, G. Multi-task weakly-supervised learning model for pulmonary nodules segmentation and detection. In *Proceedings of the 2020 International Conference on Innovation Design and Digital Technology (ICIDDT), Zhenjing, China, 5–6 December 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 343–347.
39. Nguyen, T.C.; Nguyen, T.P.; Cao, T.; Dao, T.T.P.; Ho, T.N.; Nguyen, T.V.; Tran, M.T. MANet: Multi-branch attention auxiliary learning for lung nodule detection and segmentation. *Comput. Methods Programs Biomed.* **2023**, *241*, 107748. [CrossRef]
40. Armato, S.G., III; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef]
41. Trial Summary—Learn-NLST—The Cancer Data Access System. Available online: https://cdas.cancer.gov/learn/nlst/trial-summary/ (accessed on 14 July 2025).
42. Van Ginneken, B.; Armato, S.G., III; de Hoop, B.; van Amelsvoort-van de Vorst, S.; Duindam, T.; Niemeijer, M.; Murphy, K.; Schilham, A.; Retico, A.; Fantacci, M.E.; et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study. *Med. Image Anal.* **2010**, *14*, 707–722. [CrossRef] [PubMed]
43. ELCAP Public Lung Image Database. Available online: http://www.via.cornell.edu/lungdb.html (accessed on 14 July 2025).
44. A.T.C. Tianchi Medical AI Competition Dataset. 2017. Available online: https://tianchi.aliyun.com/competition/entrance/231601/information (accessed on 14 July 2025).
45. Hao, S.; Wang, P.; Hu, Y. Haze image recognition based on brightness optimization feedback and color correction. *Information* **2019**, *10*, 81. [CrossRef]
46. Ni, Y.; Zi, D.; Chen, W.; Wang, S.; Xue, X. Egc-yolo: Strip steel surface defect detection method based on edge detail enhancement and multiscale feature fusion. *J. Real-Time Image Process.* **2025**, *22*, 65. [CrossRef]
47. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



https://dspace.org/

Multi-task deep learning for lung nodule detection and segmentation in CT scans

Li, Runhan; Honarvar Shakibaei Asli, Barmak

2026-02-02
MDPI
https://doi.org/10.3390/electronics15040736
Li R, Honarvar Shakibaei Asli B. (2026) Multi-task deep learning for lung nodule detection and segmentation in CT scans. Electronics, Volume 15, Issue 4, February 2026, Article number 736

Downloaded from a DSpace repository. https://dspace.org/