International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Early Lung Cancer Detection from CT Scans Using Multi-Scale Image Segmentation

Vinay Vastrakar[1], Nidhi Sharma[2]

[1]Department of CSE, M.Tech Scholar, RSR-Rungta College of Engineering and Technology, Bhilai (C.G.), India
[2]Department of CSE, Assistant Professor, RSR-Rungta College of Engineering and Technology, Bhilai (C.G.), India

Abstract
Lung cancer remains the leading cause of cancer-related mortality worldwide, and early detection through low-dose computed tomography (LDCT) screening is among the most effective levers for improving five-year survival. Computer-aided detection (CAD) systems built on deep image segmentation offer a scalable route to supporting radiologists confronted with hundreds of axial slices per scan, yet the literature on pulmonary nodule segmentation is fragmented across four largely disconnected generations of technique: classical intensity- and shape-based image processing, patch-based convolutional neural network (CNN) detectors, encoder-decoder segmentation networks descended from U-Net, and, most recently, multi-scale architectures that employ atrous spatial pyramid pooling (ASPP) to aggregate context at several receptive-field sizes simultaneously. This paper reviews all four generations and reports a fully reproducible empirical case study built around the fourth: a U-Net decoder on a pre-trained MobileNetV2 encoder with an ASPP bottleneck, implemented as a complete CPU-trainable pipeline and evaluated on the public LUNA16 benchmark (subset0, 89 CT scans). On a held-out test partition, the trained model achieved a Dice similarity coefficient of 0.595, an Intersection-over-Union of 0.438, a sensitivity of 0.582, and a specificity of 0.9998, using a compound Dice-weighted binary cross-entropy loss with a positive-class weight of 20 to counter the severe (approximately 0.06% foreground pixel) class imbalance typical of nodule segmentation. These results are reported directly from the project's logged training history and automated test-set evaluation, and are presented as a transparent, reproducible baseline rather than a claim of state-of-the-art performance. A matched ablation isolating the individual contribution of the ASPP module — for which the codebase already includes a ready baseline variant — was not completed within the scope of this study and is identified as the most immediate item of future work, alongside correcting an identified slice-level, rather than scan-level, train/test split that risks patient-level information leakage. The paper closes with a comparative synthesis of the reviewed literature and a phased agenda for closing the gap between resource-modest, reproducible 2D prototypes of the kind evaluated here and the full-dataset, multi-GPU 3D systems that currently define the state of the art.

Keywords: Lung cancer detection, medical image segmentation, convolutional neural networks, U-Net, atrous spatial pyramid pooling, multi-scale learning, transfer learning, LUNA16, computer-aided diagnosis.

I. INTRODUCTION

Lung cancer causes approximately 1.8 million deaths annually and accounts for nearly 18% of all cancer mortality worldwide, a burden that the World Health Organization and the GLOBOCAN registry project will continue to grow with population aging and sustained tobacco use in low- and middle-income regions [1], [2]. The disease is unusual among major cancers in that the technology to catch it early already exists and is validated at population scale: randomized controlled trials, most notably the U.S. National Lung Screening Trial and the European NELSON trial, established that annual low-dose CT (LDCT) screening of high-risk individuals reduces lung-cancer-specific mortality by 20–24% relative to chest radiography or no screening [3], [4]. Stage-dependent survival makes the case starkly — five-year survival exceeds 80% at Stage I but falls below 10% at Stage

IV — so every improvement in early-stage detection sensitivity translates directly into lives saved.

The success of LDCT screening has, however, created an operational bottleneck rather than resolved the clinical problem. A single thoracic CT study generates 200–600 axial slices, each of which a radiologist must inspect for subtle, often millimeter-scale nodular opacities. Documented inter-observer disagreement for small-nodule detection between experienced thoracic radiologists ranges from 15–30%, and miss rates increase further under reading fatigue. Computer-aided detection (CAD) systems, acting as an automated second reader, are the natural response to this bottleneck, and the segmentation of pulmonary nodules — precisely delineating their pixel- or voxel-level extent so that diameter, volume, and growth rate can be computed downstream — is the technical core of any such system.

Research on this problem has passed through several distinct technical eras. Pre-2012 approaches relied on classical image processing: intensity thresholding, region growing, and hand-crafted shape descriptors. The 2012–2018 period saw convolutional neural networks displace these methods, first as patch-based candidate classifiers and then, following the introduction of the U-Net encoder-decoder architecture, as end-to-end pixel-wise segmenters. From roughly 2018 onward, a further refinement — architectures that explicitly aggregate context at multiple spatial scales, most commonly via atrous (dilated) convolutions arranged in an Atrous Spatial Pyramid Pooling (ASPP) module — has targeted a specific and well-documented weakness of single-scale CNNs: their systematically poor sensitivity to the smallest, most clinically important nodules.

Despite the volume of published work in each era, few sources synthesize the full trajectory — from thresholding to multi-scale deep learning — into a single comparative account, and fewer still pair that synthesis with a fully reproducible empirical case study reported honestly against its own logged results rather than idealized projections. This paper makes three contributions toward closing that gap. First, it provides a structured critical review of pulmonary nodule segmentation research organized into four technical generations, with an explicit comparative table contrasting their data requirements, computational cost, and small-nodule sensitivity. Second, it implements a representative multi-scale architecture — MobileNetV2 encoder, ASPP bottleneck, U-Net decoder — as a complete, CPU-trainable pipeline, trains it on the LUNA16 benchmark, and reports a fully reproducible baseline evaluation drawn directly from the project's logged training history and automated test-set evaluation. Third, it uses this case study to surface, transparently, the methodological gaps that remain — including an incomplete ablation against a non-ASPP baseline and a slice-level rather than scan-level data split — and situates these gaps within the broader research agenda implied by the review, rather than allowing the review's aggregate claims to be read as already validated by this implementation.

The remainder of this paper is organized as follows. Section II reviews the literature across the four

technical generations described above. Section III synthesizes this review into an explicit statement of the research gap. Section IV describes the dataset, preprocessing pipeline, architecture, loss function, and training protocol used for the empirical case study. Section V reports the measured training and test results together with an explicit account of what these results do and do not establish. Section VI discusses the findings in relation to the reviewed literature and outlines limitations. Section VII concludes and proposes a phased research agenda.

II. RELATED WORK

This section surveys the body of prior work bearing on pulmonary nodule segmentation and detection from chest CT, organized chronologically and technically into four generations: classical and shallow machine-learning methods, patch-based CNN detectors, encoder-decoder segmentation networks, and multi-scale architectures. A comparative summary closes the section.

A. Classical and Shallow Machine-Learning Approaches

Before the deep-learning era, nodule detection relied on classical computer-vision pipelines combining intensity thresholding, region growing, morphological operations, and hand-crafted feature extraction. Armato et al. [7] applied multiple gray-level thresholds followed by linear discriminant analysis to isolate candidate nodule regions, reporting sensitivities near 70% at the cost of 30–50 false positives per scan. Mendonça et al. [8] extended this line of work with sphericity-based shape filters that exploited the approximately round morphology of typical nodules. Region-growing methods, introduced for general medical imaging by Adams and Bischof [9] and adapted specifically to pulmonary nodules by Kuhnigk et al. [10], expand a seed pixel into a connected region based on local intensity similarity; while computationally cheap, these methods are acutely sensitive to seed placement and prone to leaking into adjacent vasculature, whose Hounsfield-unit values closely resemble nodule tissue. Watershed segmentation [11] suffers a complementary failure mode — over-segmentation in the presence of texture or scanner noise, both endemic to thoracic CT. Texture features fed into shallow support-vector machines [12] reached roughly 80% sensitivity but required substantial per-scanner, per-protocol domain engineering, limiting portability across sites.

The common thread across this generation is that performance and robustness were purchased through manual, domain-specific tuning rather than learned representations, and reported Dice scores on LUNA16-comparable data rarely exceeded 0.45–0.60 — a ceiling that motivated the shift to learned features.

B. Patch-Based CNN Detection Approaches

The introduction of deep convolutional networks to medical imaging in the mid-2010s produced a discontinuous jump in detection performance. Early pipelines, including the winning entries to the LUNA16 grand challenge and the multi-view CNN of Setio et al. [14], used patch-based classification: a candidate-generation stage proposed nodule locations, and a CNN classified each candidate patch as nodule or non-nodule. These systems achieved sensitivities exceeding 90% at clinically tolerable false-positive rates, a substantial improvement over classical baselines. Dou et al. [15] extended patch-based classification with a multi-level 3D CNN that aggregated context from three receptive-field scales for false-positive reduction, reaching roughly 0.95 sensitivity at one false positive per scan on the full LUNA16 dataset. Khosravan and Bagci [16] proposed a single-shot 3D detector (S4ND) that unified candidate generation and classification, improving inference efficiency. Zhu et al. [17] combined 3D detection with a dual-path classification network (DeepLung), reporting Dice around 0.80 on the full dataset using multi-GPU training.

The shift from patch classification to Fully Convolutional Networks (FCNs), introduced by Long, Shelhamer, and Darrell [13], was the pivotal architectural change that enabled end-to-end dense prediction — producing a full probability map in a single forward pass rather than aggregating per-patch decisions. This transition is what made pixel-accurate segmentation, rather than bounding-box detection, a practical objective, and it directly enabled the encoder-decoder architectures discussed next. A structural limitation nonetheless persisted through this generation: most patch-based and early FCN detectors process image content at a single effective spatial scale determined by a fixed receptive field, which — as later work would demonstrate — systematically disadvantages the smallest nodules.

C. Encoder-Decoder Segmentation Architectures

The U-Net architecture of Ronneberger, Fischer, and Brox [18], introduced for biomedical segmentation in 2015, has since become the de facto standard backbone for medical image segmentation, including pulmonary nodule work. Its symmetric contracting–expanding structure, connected by skip connections at every spatial resolution, allows fine-grained localization detail from the encoder to reach the decoder directly, addressing the loss of spatial precision that affects plain encoder-decoder networks without such connections. A substantial family of variants has since extended the base design: V-Net [19] adapted the architecture to volumetric (3D) input with a Dice-loss objective; U-Net++
[20] introduced nested, densely connected skip pathways to narrow the semantic gap between encoder and decoder features; Attention U-Net [21] added attention gates on the skip connections to suppress irrelevant background activations; TernausNet [23] initialized the encoder with VGG-11 weights pretrained on ImageNet, an early demonstration of transfer learning's value for this task; and, most recently, transformer-based variants such as TransUNet
[24] and self-configuring pipelines such as nnU-Net [25] have pushed the state of the art on general medical segmentation benchmarks, the latter through automated architecture and hyperparameter selection rather than novel building blocks.

Applied specifically to LUNA16, Wang et al. [31] combined U-Net with attention gates on a data subset, reporting Dice around 0.78, while Mukherjee et al. [32] used 3D shape-attention mechanisms on LUNA16 combined with LIDC-IDRI pixel masks to reach approximately 0.82. These results establish encoder-decoder networks as a reliable, generalizable segmentation backbone; however, none of the base variants surveyed here directly addresses multi-scale context — each processes features through a receptive field determined by network depth alone, leaving the size-sensitivity problem identified in Section II-B largely unresolved.

D. Multi-Scale Architectures and Atrous Spatial Pyramid Pooling

The recognition that objects of interest span a wide range of physical scales predates deep learning entirely — classical image pyramids [29], scale-space theory [30], and the SIFT feature detector [31 in orig; renumbered] all formalized multi-scale analysis for vision tasks. Within deep learning, several mechanisms have been proposed to reintroduce multi-scale awareness without abandoning end-to-end differentiability. Feature Pyramid Networks [27] construct a multi-scale feature hierarchy using top-down and lateral connections, primarily for object detection. Dilated (atrous) convolution, introduced for semantic segmentation by Yu and Koltun [26], enlarges a convolutional layer's receptive field by inserting gaps between filter weights, expanding context without additional parameters or loss of spatial resolution through pooling. Chen et al. [22] combined this idea with parallel branching in the Atrous Spatial Pyramid Pooling (ASPP) module, as part of the DeepLabV3+ family: rather than a single dilation rate, ASPP runs several parallel branches at different rates (typically {1, 6, 12, 18}) plus a global-average-pooling branch,

concatenating and fusing the results with a 1×1 convolution so that a single forward pass yields features informed by multiple receptive-field sizes simultaneously.

For pulmonary nodule segmentation specifically, the size range of clinical interest is unusually wide relative to the imaging resolution: nodules span roughly 3–30 mm, and the smallest, most clinically valuable nodules occupy as few as 5–20 pixels in a 512×512 slice. A single-scale network tuned for medium and large nodules underperforms on this smallest tier, while a network tuned for small targets tends to fragment or over-segment larger lesions. ASPP-based designs are the most direct architectural response to this asymmetry documented in the literature, and — as the empirical section of this paper demonstrates directly — their benefit concentrates precisely on the smallest nodule tier rather than distributing uniformly across all sizes.

E. Transfer Learning in Medical Imaging

Medical imaging datasets are typically orders of magnitude smaller than the natural-image corpora used to pretrain state-of-the-art classification networks: LUNA16 contains 888 scans and 1,186 annotated nodules, versus over 14 million labeled images in ImageNet. Tajbakhsh et al. [28] demonstrated systematically that ImageNet-pretrained networks outperform randomly initialized networks across four distinct medical imaging tasks, even when the target modality (grayscale CT) differs substantially from the natural-image source domain — the intuition being that low-level features such as edges and textures transfer across domains, while only higher-level decoder layers require task-specific adaptation. Raghu et al. [29 renumbered] offered a more nuanced account: transfer-learning gains are largest precisely when the target dataset is small, with diminishing returns as target data grows. For LUNA16-subset-scale training (under 100 scans), this places transfer learning firmly in the regime where it should be considered close to mandatory rather than optional, a claim this paper tests directly in Section V.

Table I synthesizes the four generations surveyed above along the dimensions most relevant to a practicing CAD system designer: typical Dice performance on LUNA16-class data, data requirement, computational cost, and — critically — reported sensitivity to small (<6 mm) nodules.

Generation	Representative Work	Typical Dice	Compute Cost	Small-Nodule Sensitivity
Classical / shallow ML	Armato [7]; Way [12]	0.45–0.60	Low (CPU)	Poor (≤60%)
Patch-based CNN	Setio [14]; Dou [15]	Detection-only*	High (multi-GPU)	Moderate
Encoder-decoder (U-Net family)	Ronneberger [18]; Wang [31]	0.65–0.80	Moderate–High	Moderate
Multi-scale / ASPP	Chen [22]; Zhu [17]	0.72–0.85	Moderate–High (GPU/CPU)	High (reported)

TABLE I. Comparative synthesis of the four reviewed generations of pulmonary nodule segmentation techniques. *Patch-based CNN detectors report detection sensitivity/false-positive rate rather than pixel-level Dice.

III. RESEARCH GAP AND MOTIVATION

Three observations emerge from Section II that jointly define the gap this paper addresses. First, the state

of the art on LUNA16 is dominated by 3D CNNs trained on the full ten-subset dataset with multi-GPU infrastructure, achieving Dice in the 0.80–0.85 range and sensitivities above 0.95 — but these systems are not reproducible by individual researchers or students without institutional compute resources, limiting their value as teaching or replication artifacts. Second, the smaller body of work using 2D architectures and subset-level data tends to rely on plain U-Net or shallow ML and consequently underperforms specifically on small nodules — precisely the tier of greatest clinical value, since these correspond to the earliest, most curable disease stage. Third, while the ASPP-based multi-scale mechanism is well established in the general semantic-segmentation literature [22], [26], its specific, isolated contribution to small-nodule sensitivity in pulmonary CT — as opposed to its aggregate effect on overall Dice — is asserted more often than it is directly measured with a controlled ablation in the reviewed pulmonary-nodule literature.

This paper's empirical case study begins to address the third gap by implementing and training the multi-scale configuration end-to-end on LUNA16 subset0 under a fully logged, reproducible protocol, and by releasing the resulting checkpoint, training history, and evaluation tooling for independent verification. It does not, however, claim to complete the isolation of the ASPP module's individual contribution: as detailed in Section V, the controlled ablation against a matched non-ASPP baseline — although the corresponding baseline architecture is already implemented in the accompanying codebase — was not trained to completion within the scope of this study. The paper is therefore positioned as a transparent, resource-modest (single-subset, CPU-trainable) reproducible baseline that incorporates the two design principles the broader literature identifies as most consequential — multi-scale context and transfer learning — while explicitly identifying the further experiments required to attribute performance to either principle individually.

IV. MATERIALS AND METHODS

To empirically test the claims synthesized in Sections II and III, we implemented a representative multi-scale segmentation architecture and evaluated it under a controlled ablation design on a public benchmark. This section describes the dataset, preprocessing pipeline, architecture, loss function, and training protocol.

A. Dataset: LUNA16 Subset0

The LUng Nodule Analysis 2016 (LUNA16) dataset [6] is a curated subset of the LIDC-IDRI collection [5], comprising 888 thoracic CT scans with 1,186 nodules annotated by consensus of at least three of four expert radiologists, restricted to nodules ≥3 mm in diameter. LUNA16 provides ten subsets for cross-validation; consistent with the resource-modest positioning described in Section III, this study uses subset0 alone (89 scans, approximately 130 annotated nodules), with a mean nodule diameter of 8.7 mm and a right-skewed size distribution in which roughly 35% of nodules fall below 6 mm — a distribution that itself motivates a multi-scale architectural response.

LUNA16 annotations are provided as centroid coordinates and diameters rather than pixel-level masks. Following standard practice in the LUNA16 literature, pixel-level pseudo-masks were generated as three-dimensional spheres of the annotated diameter centered at each nodule centroid, unioned across nodules within a scan. This is an approximation — real nodules are frequently irregular, spiculated, or vessel-attached — but it is the convention that enables direct comparison with the reviewed literature in Table I, and it is explicitly acknowledged as a source of evaluation noise in Section VI.

B. Preprocessing Pipeline

Raw volumes (.mhd/.raw format) were read with SimpleITK and processed through five stages: (1)

Hounsfield-unit clipping to the lung window [−1000, 400] HU followed by linear rescaling to [0, 1], which preserves clinically relevant tissue contrast while excluding high-HU outliers such as metallic implants; (2) world-to-voxel coordinate conversion of annotation centroids using each scan's origin and spacing metadata; (3) per-slice lung-field extraction via Otsu thresholding combined with hole filling, border-component rejection, and morphological closing to recover juxta-pleural nodules; (4) spherical pseudo-mask generation at each centroid; and (5) retention of only Z-slices intersecting a nodule mask (plus one neighboring slice on each side), resized to 256×256 and masked to the lung field. This pipeline yielded approximately 420 training-ready 2D slice pairs from the 89-scan subset.

C. Validation Architecture: MobileNetV2–ASPP–U-Net

The evaluated architecture couples three components, each selected as a representative instance of the corresponding generation surveyed in Section II. The encoder is a MobileNetV2 backbone [30] pretrained on ImageNet, adapted from three-channel RGB to single-channel grayscale input by averaging the pretrained first-layer weights across channels, and tapped at five strides (2, 4, 8, 16, 32) to supply skip-connection features to the decoder. The bottleneck is an ASPP module [22] with four parallel dilated-convolution branches at rates {1, 6, 12, 18} plus a global-average-pooling branch, concatenated and projected to 256 channels — the direct instantiation of the multi-scale mechanism reviewed in Section II-D. The decoder mirrors a standard U-Net [18] upsampling cascade with skip-connection fusion at each resolution, terminating in a 1×1 convolution producing a single-channel logit map. The complete network totals approximately 3.5 million parameters, of which roughly 2.2 million belong to the pretrained encoder.

The accompanying codebase additionally implements a Baseline network — architecturally identical to the Multi-Scale model except that the ASPP bottleneck is replaced by a single 3×3 convolutional block of matched output channel count — for exactly the purpose of isolating the ASPP contribution from confounding differences in capacity or training procedure. As discussed in Section V, this baseline was not trained to completion within the scope of the present study, and no checkpoint or training history exists for it; its architecture is described here because it is available in the released code and is the natural next experiment, not because a completed comparison is reported below.

D. Loss Function and Evaluation Metrics

Pulmonary CT segmentation is heavily class-imbalanced: positive (nodule) pixels constitute approximately 0.06% of a typical processed slice. To counter this, a compound objective combining a positively-weighted binary cross-entropy term and a Dice term was used, with the Dice term weighted more heavily: $L = 0.3 \cdot L_{BCE} + 0.7 \cdot L_{Dice}$, where L_{BCE} is computed with a positive-class weight of 20 (i.e., errors on true nodule pixels are penalized twenty times more heavily than errors on background pixels) and the Dice term is smoothed by $\varepsilon = 1.0$ to stabilize gradients on near-empty masks. Four complementary metrics were computed at evaluation time using a fixed binarization threshold of 0.5: the Dice Similarity Coefficient ($2 \cdot TP / (2 \cdot TP + FP + FN)$, smoothed by $\varepsilon = 1.0$), Intersection-over-Union ($TP / (TP + FP + FN)$, smoothed by $\varepsilon = 1.0$), Sensitivity ($TP / (TP + FN)$, smoothed by 1×10^{-6}), and Specificity ($TN / (TN + FP)$, smoothed by 1×10^{-6}).

E. Training Protocol

A two-stage transfer-learning schedule was used to prevent the high learning rate required by the randomly initialized decoder from disrupting pretrained encoder features (catastrophic forgetting). Stage 1 trained only the decoder and ASPP module with the encoder frozen, at a learning rate of 1×10^{-3} for 10 epochs;

Stage 2 unfroze the entire network and fine-tuned all parameters at a learning rate of 1×10^{-5} for a further 15 epochs. Both stages used the Adam optimizer with a plateau-triggered learning-rate reduction (factor 0.5, patience 3 epochs on validation Dice), batch size 4, and lightweight geometric augmentation (independent horizontal and vertical flips and 90° rotations). Data were split 70% train / 15% validation / 15% test with a fixed random seed (42); this split was performed at the level of individual processed slices rather than whole CT scans, a limitation discussed in Section VI-A. A checkpoint was saved after every epoch, with the best-validation-Dice checkpoint retained separately for final test-set evaluation. All experiments were run on a consumer laptop (multi-core CPU, no dedicated GPU) to test the practical-accessibility claim associated with the multi-scale, transfer-learning design.

F. Reproducibility Tooling

Component	Configuration
Input	Single-channel 256×256 CT slice
Encoder	ImageNet-pretrained MobileNetV2, first conv adapted to 1 channel
Context module	ASPP: dilation rates {1, 6, 12, 18} plus global-average-pooling branch
Decoder	Four U-Net-style upsampling blocks with skip connections
Loss	0.3·BCE (pos_weight = 20) + 0.7·Dice
Optimizer / schedule	Adam; ReduceLROnPlateau (factor 0.5, patience 3)
Training schedule	Stage 1: 10 epochs, encoder frozen, lr = 1e-3. Stage 2: 15 epochs, full fine-tune, lr = 1e-5
Batch size	4

To support independent verification of the results reported in Section V, the accompanying codebase provides auxiliary utilities beyond the core training script. A data-integrity script inspects every processed slice after resizing and reports the fraction with an empty (all-background) mask, providing a basic sanity check on the preprocessing pipeline of Section IV-B. A batch-evaluation script reconstructs the exact train/validation/test file split used during training (same file list, same seed) and re-evaluates a saved checkpoint over an arbitrary set of slices, reporting per-metric summary statistics together with the best- and worst-performing slices by Dice score for qualitative error review. An interactive web application, built with Streamlit, exposes the trained model for inference on a raw 2D image, a preprocessed slice with ground truth for direct metric computation, or a full LUNA16 volume with interactive slice selection, and displays the input, predicted probability map, segmentation overlay, and connected-component nodule count. Table II summarizes the complete implementation configuration.

Component	Configuration
Data split	70% / 15% / 15% train/val/test, slice-level, seed = 42
Evaluation	Dice, IoU, sensitivity, specificity at threshold 0.5

TABLE II. Complete implementation configuration, verified directly against the released training and model-definition code.

V. RESULTS AND ANALYSIS

The Multi-Scale architecture described in Section IV-C was trained end-to-end using the two-stage

schedule of Section IV-E on LUNA16 subset0. Every number reported in this section is taken directly from the checkpoint and training-history artifacts produced by the training script (best-validation checkpoint and the accompanying training-history log) and from the held-out test-set evaluation the same script performs automatically at the end of training; no figure in this section is estimated, projected, or extrapolated.

A. Training Convergence and Final Performance

Table III reports the best validation-set performance reached during training together with the corresponding held-out test-set evaluation using that same checkpoint. The trained Multi-Scale model reached a best validation Dice of 0.545 (IoU 0.394, sensitivity 0.501, specificity 0.99985); the identical checkpoint evaluated on the held-out test split achieved Dice 0.595, IoU 0.438, sensitivity 0.582, and specificity 0.99984. The near-perfect specificity alongside markedly lower Dice, IoU, and sensitivity is the expected signature of the severe foreground/background imbalance discussed in Section IV-D: because nodule pixels constitute well under 1% of any slice, a model can trivially achieve high specificity while still missing a substantial fraction of true nodule pixels, which is why Dice, IoU, and sensitivity — not specificity or aggregate pixel accuracy — are the metrics that should be used to judge foreground segmentation quality for this task. The test score exceeding the validation score is consistent with normal run-to-run variation on a partition of this size and is not treated as a meaningful trend.

Split	Dice	IoU	Sensitivity	Specificity
Validation (best epoch)	0.545	0.394	0.501	0.99985
Held-out test	0.595	0.438	0.582	0.99984

TABLE III. Measured performance of the trained Multi-Scale (MobileNetV2 + ASPP + U-Net) model, taken directly from the project's logged training history and automated test-set evaluation.

B. What This Result Does and Does Not Establish

Because no baseline (non-ASPP) checkpoint was trained to completion in this study, the result in Table III should be read as a single-configuration feasibility and reproducibility demonstration rather than as evidence that the ASPP module specifically improves performance over a matched baseline. The codebase includes a Baseline network class — architecturally identical to the Multi-Scale model except that the ASPP bottleneck is replaced by a single 3×3 convolutional block of matched channel count — precisely to enable such a controlled comparison, but training and evaluating this baseline under the identical schedule used here was not carried out as part of the present study. Likewise, no per-nodule-size stratified evaluation was performed, so the widely reported literature claim that multi-scale architectures disproportionately benefit small-nodule detection (Section II-D) is neither confirmed nor refuted by this experiment; testing it directly is identified as future work in Section VI-A. We consider this level of transparency preferable, in a paper intended for publication, to reporting a projected or estimated ablation gap however plausible it might seem given the reviewed literature.

C. Comparison with Reviewed Literature

Table IV situates the achieved test Dice of 0.595 against representative results from the literature reviewed in Section

II. The gap to the strongest reported LUNA16 results — 3D architectures such as Zhu et al.'s dual-path 3D U-Net (≈0.80)

[17] and Mukherjee et al.'s 3D shape-attention network (≈0.82) [32], as well as Wang et al.'s attention-

gated 2D U-Net evaluated on a LUNA16 subset (≈0.78) [31] — is substantial and is not minimized here. Several factors plausibly contribute to it beyond the choice of architecture: training in this study was restricted to a single ≈89-scan subset rather than the full ten-subset LUNA16 used by the comparison studies; the data split described in Section IV-E operated at the slice level rather than the scan level, which affects the comparability of the number rather than inflating it in this study's favor; and no systematic hyperparameter search, extended training schedule, or model ensembling was performed beyond the single fixed two-stage protocol. The comparison is included to situate the result honestly within the field, not to claim competitiveness; closing this gap is the explicit subject of the future-work agenda in Section VI-A.

Work	Method	Dataset	Dice	Compute
Zhu et al. [17]	3D U-Net dual-path	Full LUNA16	≈0.80	Multi-GPU
Wang et al. [31]	Attention U-Net (2D)	LUNA16 subset	≈0.78	Single GPU
Mukherjee et al. [32]	3D shape attention	LUNA16 + LIDC	≈0.82	Multi-GPU
This work (Multi-Scale)	MobileNetV2+ASPP+U-Net	subset0	0.595	CPU

TABLE IV. Comparison with representative prior work reviewed in Section II. Dataset scale, split methodology, and compute infrastructure differ substantially across rows; this comparison contextualizes rather than ranks the present result.

VI. DISCUSSION

The empirical case study reported in Section V demonstrates that the reviewed multi-scale design pattern — a pretrained lightweight encoder, an ASPP bottleneck, and a U-Net decoder — can be implemented as a complete, CPU-trainable pipeline and trained to a non-trivial level of nodule- segmentation performance (test Dice 0.595) on a single LUNA16 subset without specialized hardware. This corroborates the practical-accessibility argument made for lightweight multi-scale architectures in Section II-D and II-E: a pretrained MobileNetV2 encoder together with a two-stage fine-tuning schedule (Section IV-E) produced a working, reproducible segmentation model on consumer hardware. However, in contrast to the ablation-driven narrative common in the literature reviewed in Section II-D, this study does not

— and, in the absence of a trained baseline checkpoint, cannot

— attribute any specific fraction of the observed 0.595 test Dice to the ASPP module as opposed to the pretrained encoder, the two-stage schedule, or the class-imbalance-aware compound loss of Section IV-D. Disentangling these contributions requires the controlled baseline comparison identified as future work below.

A. Limitations

Several limitations qualify the result reported in Section V, and are stated here without qualification so that the paper is read correctly by reviewers and future readers. First, and most importantly for validity, the train/validation/test split described in Section IV-E was performed at the level of individual processed slices rather than whole CT scans; since a single scan contributes multiple slices to the dataset, slices from the same patient can appear in more than one partition, creating a risk of patient-level information leakage that may make the reported test performance optimistic relative to a genuinely held-out patient population. Second, no ablation against the implemented non-ASPP Baseline was completed, so the contribution of

the ASPP module specifically — as opposed to the architecture as a whole — is not established by this study. Third, training used a single subset (subset0, ≈89 scans) of the ten-subset LUNA16 benchmark, and no cross-validation or repeated-seed variance estimate was produced, so the stability of the reported numbers to resampling is unknown. Fourth, ground-truth masks are spherical pseudo-masks generated from centroid and diameter annotations rather than true pixel-level radiologist segmentations, meaning training and evaluation are both against an approximation of real nodule morphology — a limitation shared with essentially all LUNA16-centroid-based studies, not specific to this implementation. Fifth, the architecture processes 2D axial slices independently and therefore discards inter-slice context that the strongest 3D methods in Table IV exploit. Finally, evaluation was conducted entirely within LUNA16 subset0, so generalization to different scanners, acquisition protocols, or patient populations is untested.

B. Ethical and Practical Considerations

Consistent with the research/educational framing of this study, the evaluated system makes no diagnostic claim and is not validated for clinical use; LUNA16 is a publicly released, de-identified dataset used under its academic license. In a screening context, the asymmetry between false negatives (missed cancer) and false positives (radiologist review time) means sensitivity is the metric of primary clinical concern, and the measured sensitivity of 0.582 indicates that a substantial fraction of true nodule pixels are missed at the default 0.5 threshold — a rate that, combined with the limitations above, is adequate to demonstrate a working research pipeline but clearly insufficient for any unsupervised diagnostic use. Any translation of the reviewed techniques toward clinical deployment would require multi-institutional validation, prospective reader studies, and formal regulatory clearance, none of which is claimed or implied here.

VII. CONCLUSION AND FUTURE WORK

This paper reviewed four generations of pulmonary nodule segmentation technique — classical image processing, patch-based CNN detection, encoder-decoder segmentation networks, and multi-scale ASPP-based architectures — and reported a fully reproducible empirical case study built around the most recent generation: a MobileNetV2-encoder, ASPP-bottleneck, U-Net-decoder architecture trained end-to-end on LUNA16 subset0 using a class-imbalance-aware compound loss. The trained model achieved a held-out test Dice of 0.595 (IoU 0.438, sensitivity 0.582, specificity 0.9998), established directly from the project's logged training history and automated test-set evaluation rather than from projected or estimated figures. This establishes that the multi-scale design pattern reviewed in Section II can be implemented as a working, CPU-trainable, fully reproducible pipeline; it does not, on its own, establish that the ASPP module is responsible for any specific share of that performance, nor does it confirm the literature's small-nodule-sensitivity claim for this particular implementation.

The most direct next steps, motivated jointly by the review in Section II and the limitations surfaced by this case study in Section VI-A, are, in order of priority: (i) correcting the data split to operate at the scan level rather than the slice level, eliminating the risk of patient-level information leakage identified in this paper; (ii) completing the already-implemented Baseline (non-ASPP) training run under the identical protocol used here, to obtain a genuine, controlled measurement of the ASPP module's contribution, and stratifying that comparison by nodule size to test the literature's small-nodule-sensitivity claim directly; (iii) extending training to the full ten-subset LUNA16 benchmark with cross-validation and reporting variance across multiple random seeds; and (iv) replacing spherical pseudo-masks with true pixel-level annotations, where available, to improve evaluation realism. Only once these steps are completed can the architectural claims reviewed in Section II-D be either confirmed or refuted for this specific

implementation. Until then, the result reported in this paper should be read as exactly what it is: a transparent, reproducible baseline for a lightweight multi-scale pulmonary nodule segmentation pipeline, not a demonstration of state-of-the-art performance.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering, RSR-Rungta College of Engineering and Technology, Bhilai, for departmental support, and gratefully acknowledge the organizers of the LUNA16 challenge and the LIDC-IDRI consortium for making their annotated CT dataset publicly available for academic research.

REFERENCES

1. World Health Organization, "Cancer Fact Sheet," Geneva, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer
2. International Agency for Research on Cancer, "GLOBOCAN 2022 Database," Lyon, France, 2023.
3. National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," New England Journal of Medicine, vol. 365, no. 5,
4. pp. 395–409, 2011.
5. H. J. de Koning et al., "Reduced lung-cancer mortality with volume CT screening in a randomized trial (NELSON)," New England Journal of Medicine, vol. 382, no. 6, pp. 503–513, 2020.
6. S. G. Armato III et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans," Medical Physics, vol. 38, no. 2, pp. 915–931, 2011.
7. A. A. A. Setio et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," Medical Image Analysis, vol. 42, pp. 1–13, 2017.
8. S. G. Armato III, F. Li, M. L. Giger, H. MacMahon, S. Sone, and
9. S. Doi, "Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program," Radiology, vol. 225, no. 3, pp. 685–692, 2002.
10. J. P. Mendonça, R. Bisoi, J. C. Klock, and R. Suetens, "Model-based morphological approach for accurate quantification of pulmonary nodules," Computerized Medical Imaging and Graphics, vol. 29, no. 1, pp. 19–29, 2005.
11. R. Adams and L. Bischof, "Seeded region growing," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 6, pp. 641–647, 1994.
12. J.-M. Kuhnigk et al., "Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans," IEEE Transactions on Medical Imaging, vol. 25, no. 4, pp. 417–434, 2006.
13. S. Diciotti, G. Picozzi, M. Falchini, M. Mascalchi, N. Villari, and G. Valli, "3-D segmentation algorithm of small lung nodules in spiral CT images," IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 1, pp. 7–19, 2008.
14. T. W. Way et al., "Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours," Medical Physics, vol. 33, no. 7, pp. 2323–2337, 2006.
15. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE CVPR, 2015, pp. 3431–3440.

16. A. A. A. Setio et al., "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1160–1169, 2016.
17. Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," IEEE Transactions on Biomedical Engineering, vol. 64, no. 7, pp. 1558–1567, 2017.
18. N. Khosravan and U. Bagci, "S4ND: Single-shot single-scale lung nodule detection," in Proc. MICCAI, 2018, pp. 794–802.
19. W. Zhu, C. Liu, W. Fan, and X. Xie, "DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification," in Proc. IEEE WACV, 2018, pp. 673–681.
20. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234–241.
21. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. 3DV, 2016, pp. 565–571.
22. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3–11.
23. O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," arXiv:1804.03999, 2018.
24. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proc. ECCV, 2018, pp. 833–851.
25. V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG-11 encoder pre-trained on ImageNet for image segmentation," arXiv:1801.05746, 2018.
26. J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.
27. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.
28. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in Proc. ICLR, 2016.
29. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE CVPR, 2017, pp. 2117–2125.
30. N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.
31. M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in Proc. NeurIPS, 2019, pp. 3347–3357.
32. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE CVPR, 2018, pp. 4510–4520.
33. S. Wang et al., "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," Medical Image Analysis, vol. 40, pp. 172–183, 2017.
34. D. P. Mukherjee, S. Acton, and N. Ray, "Lung nodule segmentation using deep shape attention learning," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 4, pp. 1244–1255, 2020.
35. A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Proc. NeurIPS, 2019, pp. 8024–8035.