Thesis Title: Explainable AI for Lung Cancer Detection with LLM-Driven Clinical Narratives.

Supervised By

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Daffodil International University

Submitted By

Shah Nafis Mohammad Kavi
ID:213-35-754
Department of Software Engineering
Daffodil International University

This thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

This thesis titled on "**Explainable AI for Lung Cancer Detection with LLM-Driven Clinical Narratives**", submitted by **Shah Nafis Mohammad Kavi (ID: 213-35-754)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Md Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Habibur Rahman
Associate Professor
Department of Computer Science and Engineering
Islamic University, Bangladesh

Explainable AI for Lung Cancer Detection with LLM-Driven Clinical Narratives.

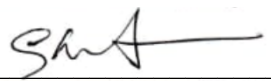
SHAH NAFIS MOHAMMAD KAVI

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR'S DECLARATION

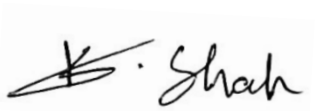
I hereby declare that I have reviewed this thesis entitled **"Explainable AI for Lung Cancer Detection with LLM-Driven Clinical Narratives"**, and in my opinion, it is adequate in terms of scope and quality for the award of the degree of Bachelor of Science in Software Engineering.



Full Name : Md. Shohel Arman
Position : Assistant Professor
Date : 15 September 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



Full Name : SHAH NAFIS MOHAMMAD KAVI

ID Number : 213-35-754

Date : 15 September 2025

Explainable AI for Lung Cancer Detection with LLM-Driven Clinical Narratives.

SHAH NAFIS MOHAMMAD KAVI

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

SEPTEMBER 2025

ACKNOWLEDGEMENTS

I would like to express appreciation to Almighty Allah for guiding me and providing me with the knowledge and strength necessary for being able to work on this thesis. This study would not have been feasible without His graces. My research was solely motivated by my desire to learn more and expand my expertise.

I am extremely thankful to my parents for their continuous support and prayers. Additionally, I want to express my gratitude to Prof. Dr. Imran Mahmud, Head of the Software Engineering Department, for creating a positive learning environment. I would especially like to thank my honorable supervisor, Md. Shohel Arman, for his unwavering support, insightful criticism, and heartfelt supervision during this project. His advice was helpful in developing this piece of work.

I would like to conclude by expressing my gratitude to my fellow DIU students and batchmates for their cooperation, support, and upbeat attitude, all of which enabled me to successfully finish this study.

DEDICATION

This thesis is dedicated towards the dreamers, the doubters, and everyone who dares to keep going when the road feels endless. This is proof that persistence can light the way.

ABSTRACT

Lung cancer remains one of the leading causes of cancer-related mortality worldwide. This paper presents a comprehensive AI-driven framework for early and accurate detection of lung cancer using the LIDC-IDRI dataset, integrating explainable AI (XAI) techniques and large language model (LLM)-generated clinical narratives to enhance trust and interpretability. The proposed system preprocesses DICOM series and XML annotations to generate pseudo-3D inputs from three adjacent CT slices centered on radiologist-annotated nodules, storing malignancy scores as averaged floating-point values. Three deep learning models — EfficientNetV2-S, DenseNet201, and MobileViT-XXS — are trained using 5-fold stratified cross-validation with binary cross-entropy loss and label smoothing. A Multi-Attention Stacked Ensemble (MASE) fuses base model predictions for improved performance. Grad-CAM explanations are generated per model and aggregated for robust visualization, while an LLM transforms model outputs and CAM data into concise, radiologist-style justifications.

Experimental results show an ensemble accuracy of 94.9%, AUC of 0.9821, sensitivity of 0.9560, specificity of 0.9459, and F1-score of 0.9074, outperforming individual models and demonstrating robust classification ability. The integration of XAI and automated clinical narratives addresses the critical need for transparency in medical AI, potentially improving adoption in clinical workflows. This work contributes novel methodologies for pseudo-3D processing, malignancy score preservation, multi-attention ensembling, and combined visual-textual explainability.

Keywords: Lung cancer, LIDC-IDRI, pseudo-3D, explainable AI, Grad-CAM, multi-attention ensemble, large language models, clinical trust.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

- τ — Decision threshold for converting probabilities to class labels
- TP, FP, TN, FN — Confusion-matrix counts (true/false positives/negatives)
- TPR, FPR — True Positive Rate and False Positive Rate (ROC axes)
- I(z) — Central axial slice at index z for CAM overlay (pseudo-3D full-slice inputs)
- R×R — Common input resolution for normalization/overlay of CAMs
- F — Feature tensor at a chosen layer for Grad-CAM (channels × spatial size)
- z — Malignancy logit (pre-sigmoid) used for Grad-CAM gradients
- Φ — Structured descriptor vector driving the LLM narrative (risk bin, CAM stats, etc.)
- μ — CAM-derived location descriptor (qualitative location derived from μ)
- H — Entropy-based focality measure of normalized saliency (lower H means focal)
- d — Proximity (distance) of CAM centroid to annotated nodule center (if available)
- T — Temperature parameter for probability calibration (temperature scaling)
- θ — Parameter vector of the LLM used in the narrative generation prompt formalization
- k×k — Convolutional kernel size (used in pseudo-3D through-plane discussion)
- W_1, W_2, b_1, b_2 — Weights and biases in the position-wise MLP (Transformer block)
- φ — SiLU activation
- ρ — ReLU nonlinearity symbol used in attention stacker description

LIST OF ABBREVIATIONS

- LIDC-IDRI — Lung Image Database Consortium and Image Database Resource Initiative
- CAD — Computer-Aided Diagnosis
- CT — Computed Tomography (context: CT slices / axial slices)
- DL — Deep Learning
- CNN — Convolutional Neural Network(s)
- XAI — Explainable Artificial Intelligence
- LLM — Large Language Model
- MASE — Multi-Attention Stacked Ensemble
- ROC — Receiver Operating Characteristic (as ROC curve / ROC–AUC context)
- PR — Precision–Recall (as PR curve / PR–AUC context)
- AUC — Area Under the ROC Curve (also used simply as "AUC")
- AUPRC — Area Under the Precision–Recall Curve (PR–AUC)
- CV — Cross-Validation (patient-level 5-fold CV)
- Grad-CAM / CAM — Gradient-weighted Class Activation Mapping (and "CAM" shorthand
- NLST — National Lung Screening Trial (dataset reference)
- PHI — Protected Health Information ("non-PHI numeric tuples")
- ReLU — Rectified Linear Unit (nonlinearity)
- SiLU — Sigmoid Linear Unit (nonlinearity)
- MLP — Multi-Layer Perceptron (position-wise MLP in Transformer block)
- AdamW — Adam with decoupled weight decay (optimizer)
- BCEWithLogitsLoss — Binary cross-entropy with logits (loss)

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, accounting for an estimated 1.8 million deaths annually [1]. Despite significant advances in treatment and screening, early diagnosis remains difficult, primarily due to the heterogeneous appearance of pulmonary nodules and the high inter-observer variability among radiologists [2]. The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset has become a cornerstone for developing and benchmarking computer-aided detection and diagnosis (CAD) systems [3].

Deep learning (DL) has revolutionized medical image analysis, achieving state-of-the-art performance across tasks such as tumor detection, segmentation, and classification [4,5]. Convolutional neural networks (CNNs) and transformer-based architectures have demonstrated strong capabilities in capturing complex radiological features. However, individual models are prone to overfitting and may fail in edge cases due to dataset imbalance or lack of generalization [6]. Ensemble approaches, which integrate predictions from multiple models, have been shown to enhance robustness and accuracy [7].

In this research, we propose a multi-architecture ensemble framework for lung cancer detection, leveraging pseudo-3D nodule representations from the LIDC-IDRI dataset. Each input sample consists of three consecutive axial slices centered on a radiologist-annotated nodule. This strategy provides volumetric context without the computational burden of full 3D networks. Unlike prior works that crop nodules into small patches [11], we retain entire axial slices, ensuring preservation of surrounding anatomical structures such as vasculature, parenchyma, and airway context. This design better reflects the clinical reasoning process, where radiologists rarely evaluate nodules in isolation.

The pipeline integrates three complementary architectures: EfficientNetV2-S, DenseNet201, and MobileViT-XXS. Their outputs are combined using a Multi-Attention Stacked Ensemble (MASE) that assigns adaptive weights to each base model, enhancing predictive stability. To address the black-box nature of deep learning, we incorporate explainable AI (XAI) via Gradient-weighted Class Activation Mapping (Grad-CAM), which highlights imaging regions driving malignancy predictions. Finally, we introduce a large language model (LLM)-based narrative generator that transforms predictions and visual evidence into radiologist-style textual justifications.

This combination of context-aware input, robust ensemble learning, visual explanations, and narrative outputs establishes a clinically relevant, interpretable, and high-performing framework for lung cancer classification.

1.2 Motivation

The motivation for this research stems from three critical challenges in lung cancer diagnostics:

(1) Early detection barriers: Pulmonary nodules are often small and visually ambiguous, complicating their classification as benign or malignant. Radiologists face difficulties distinguishing subtle imaging cues, and inter-observer variability further exacerbates diagnostic uncertainty [8]. Automated methods capable of consistently detecting malignancy can significantly improve clinical outcomes.

(2) Limitations of existing deep learning methods: Single-model CNNs or ViTs, though powerful, are vulnerable to overfitting and biased toward dominant imaging features [9]. Ensemble methods have been shown to improve robustness by combining diverse feature representations [7]. However, many existing lung cancer AI studies restrict analysis to cropped nodule patches, neglecting the broader thoracic context [11]. This simplification may lead to unrealistic performance in controlled settings but hinders clinical generalization. Our approach instead processes full slices with pseudo-3D context, preserving anatomical cues that radiologists rely on [13].

(3) Trust and interpretability: High accuracy alone is insufficient for clinical adoption. Radiologists demand transparency and actionable explanations [12]. Prior work shows that clinicians are hesitant to trust black-box models without interpretable reasoning [14]. By combining Grad-CAM visualizations with LLM-generated clinical narratives, our pipeline addresses this trust gap, producing outputs aligned with radiological decision-making.

(4) Clinical usability: Importantly, our system is designed with medical professionals in mind. Studies have shown that interpretability improves physician confidence and willingness to use AI systems [12,14]. By offering both visual (CAM heatmaps) and textual justifications, the proposed pipeline provides richer context than conventional models, making it more useful in real-world diagnostic workflows.

1.3 Research Objective

1. **Develop a pseudo-3D representation of nodules** by extracting three adjacent slices centered on annotated nodules from the LIDC-IDRI dataset.

2. **Preserve full-slice context rather than cropped patches**, ensuring the model learns from both nodule morphology and surrounding anatomical structures.

3. **Train multiple deep learning architectures** (EfficientNetV2-S, DenseNet201, MobileViT-XXS) for binary classification of pulmonary nodules (benign vs malignant).

4. **Design a Multi-Attention Stacked Ensemble (MASE)** to integrate base model outputs with adaptive weighting, improving robustness and accuracy.

5. **Incorporate explainability via Grad-CAM**, highlighting image regions most relevant to malignancy classification.

6. **Generate radiologist-style textual justifications** using LLMs, enhancing the interpretability and clinical usability of model outputs.

7. **Evaluate performance comprehensively** using metrics such as accuracy, AUC, sensitivity, specificity, F1-score, and precision, benchmarking against state-of-the-art lung cancer detection models.

1.4 Research Question

I. Can pseudo-3D slice-based inputs capture sufficient volumetric information to improve nodule classification compared to single-slice methods?

II. Does a Multi-Attention Stacked Ensemble (MASE) outperform individual models and traditional ensembles in terms of predictive stability?

III. How does preserving full-slice context (rather than cropping nodules) influence classification performance and clinical relevance?

IV. How does the proposed pipeline improve clinical interpretability and usability compared to prior black-box or patch-based approaches?

CHAPTER 2

LITERATURE REVIEW

2.1 Preliminaries

Computer-aided diagnosis (CAD) in thoracic imaging has developed rapidly in the last two decades, supported by the availability of benchmark datasets and advances in deep learning. The LIDC-IDRI dataset [2,15], with over 1,000 annotated CT cases, has emerged as the primary source for pulmonary nodule classification studies. Its richness lies not only in the imaging data but also in radiologist-provided malignancy scores and descriptive annotations, capturing inter-observer variability that reflects real-world diagnostic uncertainty.

CT imaging is inherently volumetric, but training 3D convolutional neural networks (3D-CNNs) requires immense computational resources and large datasets, often unavailable in medical imaging. As a practical alternative, pseudo-3D approaches stack adjacent 2D slices to approximate volumetric context without the prohibitive cost of 3D models [16].

Another foundational element is ensemble learning, which reduces variance and exploits model complementarity. In medical imaging, ensembles often outperform individual architectures [17]. Yet, ensemble methods must balance complexity, interpretability, and efficiency. Finally, explainability has emerged as a decisive factor in clinical translation. Techniques like Grad-CAM [19] visualize model attention, while narrative explanations align predictions with clinician reasoning [14,20]. These four pillars—datasets, input representation, ensemble learning, and explainability—frame the literature on lung cancer detection.

2.2 Previous studies on Lung Cancer Detection

Early Radiomics and Handcrafted Features: Before the deep learning era, CAD systems relied on radiomic and handcrafted descriptors. Han et al. [21] analyzed textural features from CT nodules and applied SVM classifiers, achieving moderate classification accuracy but struggling with robustness across different scanners. Aerts et al. [3] demonstrated the potential of radiogenomics by linking radiomic features to tumor

phenotypes and patient survival. However, such methods required manual feature engineering, often missing subtle imaging cues. Their reproducibility was further undermined by scanner variability, as pointed out by McWilliams et al. [8], who developed a malignancy probability model but acknowledged its dependence on consistent feature extraction. These works laid the groundwork by highlighting that quantitative imaging was feasible but insufficiently generalizable.

Patch-Based CNN Approaches: The advent of CNNs marked a paradigm shift. Shen et al. [11] applied a multi-scale CNN to cropped nodule patches from LIDC-IDRI, reporting significantly improved accuracy over radiomic models. Their design exploited CNNs' hierarchical feature extraction but restricted inputs to small cropped patches, discarding parenchymal context. Setio et al. [22] addressed false-positive reduction by training multi-view CNNs on patches extracted from multiple orientations. Their method achieved sensitivity of 85.4% at one false positive per scan, a notable benchmark at the time.

Patch-based CNNs gained popularity due to their simplicity and efficiency. Kumar et al. [23] applied transfer learning with ResNet50 on cropped nodules, reporting ~89% accuracy. Ciompi et al. [24] developed a patch-based malignancy classification system validated on LIDC-IDRI, demonstrating CNNs' strong discriminative power. However, a recurring limitation was their neglect of surrounding anatomy. For instance, vessels or lobular positioning, which radiologists often use as malignancy cues, were excluded. Consequently, models risked overfitting to local textural patterns.

Context-Preserving and Volumetric Methods: Recognizing this limitation, researchers moved toward context-preserving and volumetric approaches. Causey et al. [25] integrated radiomic features with CNN features from larger ROIs, reporting AUC 0.91 on LIDC-IDRI. Their hybrid system outperformed patch-only CNNs, underscoring the value of anatomical context. Ardila et al. [13] advanced the field by training an end-to-end 3D CNN on the NLST dataset. Their model achieved radiologist-level performance (AUC 0.94) in cancer screening, a milestone that demonstrated the power of volumetric modeling. However, training required enormous computational resources and was difficult to replicate in routine research environments.

Huang et al. [16] proposed a multi-view pseudo-3D CNN that stacked adjacent slices to approximate volumetric context more efficiently. Hancock and Magnan [15] similarly showed pseudo-3D architectures could achieve comparable performance to 3D CNNs with lower computational burden. Ciompi et al. [24] also highlighted that including contextual slices improved malignancy prediction. These studies collectively validated pseudo-3D as a pragmatic compromise, retaining context while avoiding the impracticality of full 3D CNNs.

Ensemble Learning Approaches: As single architectures reached performance plateaus, ensembles emerged. Dietterich [7] formalized ensemble theory, later applied in medical imaging. Liao et al. [26] developed a CNN ensemble for malignancy classification, achieving higher AUC than any individual network. Wang et al. [27] fused features from multiple CNNs, reporting ~87% accuracy. Hancock and Magnan [15] also observed performance gains when averaging predictions across models.

While effective, these ensembles often relied on simple averaging or majority voting, limiting their ability to capture complementary strengths. Suk and Shen [18] showed that feature-level ensemble learning improved robustness in Alzheimer's disease imaging, suggesting similar benefits could apply to lung cancer. However, until recently, lung nodule ensembles lacked advanced mechanisms to adaptively weight model contributions.

Explainability and Clinical Adoption: Interpretability has been a persistent concern. Selvaraju et al. [19] introduced Grad-CAM, widely adopted to localize salient regions in medical images. Applied to lung nodules, it revealed whether models attended to nodules versus irrelevant regions. Tjoa and Guan [10] argued that explainability is critical for physician trust, while Topol [12] emphasized its necessity for AI in clinical practice. Holzinger et al. [14] expanded the discussion to "causability," stressing explanations that align with clinical reasoning.

Attempts to generate textual explanations also emerged. Cai et al. [29] created tools for human-centered interaction with imperfect algorithms. Zhang et al. [30] combined CNNs with RNNs to generate radiology-style reports for chest X-rays, though not yet applied to CT nodules. Lundervold and Lundervold [28] reviewed deep learning in medical

imaging, concluding that trustworthiness requires interpretable pipelines. However, most lung nodule studies stopped at Grad-CAM heatmaps, leaving a gap in narrative justification that radiologists find essential.

Multi-Attention Stacked Ensemble (MASE): A significant leap came with the work of Saha and Prakash [31], who proposed a Multi-Attention Stacked Ensemble (MASE) tailored for lung nodule classification. They fine-tuned EfficientNetV2-S, DenseNet201, and MobileViT-XXS on 96×96 cropped patches from LIDC-IDRI, then introduced a two-stage attention mechanism: model-wise attention to weight backbone contributions and class-wise attention to refine predictions. A lightweight meta-learner further improved accuracy. Their system incorporated dynamic focal loss, MixUp augmentation, and test-time augmentation, achieving 98.09% accuracy and 0.9961 AUC, surpassing all prior state-of-the-art. Performance was balanced across sensitivity (98.73%) and specificity (98.96%), and robustness was confirmed in challenging cases with high radiologist disagreement.

This work set a new benchmark, but its reliance on cropped patches remained a limitation. By discarding contextual anatomy, it risked misalignment with clinical reasoning. Furthermore, while technically impressive, it did not integrate advanced interpretability beyond predictive performance metrics. The present research directly builds on MASE, adapting it to full-slice pseudo-3D inputs to preserve context and extending it with dual interpretability mechanisms (Grad-CAM + LLM narratives). This progression not only advances accuracy but also strengthens clinical usability, addressing both performance and trust.

2.3 Comparative Analysis

The trajectory of prior research shows a clear evolution. Early radiomic methods established feasibility but lacked robustness. Patch-based CNNs boosted performance but discarded context. Volumetric models captured rich features but were computationally burdensome. Pseudo-3D emerged as a balanced compromise. Ensemble models

improved robustness but often used simplistic averaging. Explainability frameworks addressed trust but rarely went beyond saliency maps.

Saha and Prakash's MASE framework [31] marked a high point by combining advanced backbones with attention-based stacking, achieving record-breaking accuracy. Yet, its dependence on cropped patches limited clinical realism. Our research extends this by retaining full-slice pseudo-3D context and adding dual interpretability, positioning the pipeline as clinically aligned in usability.

Table 2.1: Comparative Analysis of Previous Studies

Ref	**Dataset**	**Methodology**	**Best Model**	**Results**	**Limitations**
[11] Shen et al. (2015)	LIDC-IDRI	Multi-scale CNN on cropped patches	CNN	86% accuracy	Context discarded
[22] Setio et al. (2016)	LIDC-IDRI	Multi-view CNN	Multi-view CNN	Sensitivity 85.4%	High FP; patch-based
[13] Ardila et al. (2019)	NLST	End-to-end 3D CNN	3D CNN	AUC 0.94	Very high compute cost
[23] Kumar et al. (2020)	LIDC-IDRI	Transfer learning CNN	ResNet50	89% accuracy	Limited generalization
[25] Causey et al. (2018)	LIDC-IDRI + radiomics	Hybrid CNN + radiomics	Hybrid CNN	AUC 0.91	No full-slice preservation
[26] Liao et al. (2019)	LIDC-IDRI	Ensemble of CNNs	CNN Ensemble	AUC 0.94	Averaging only; no XAI
[27] Wang et al. (2017)	LIDC-IDRI	Deep feature fusion	CNN Fusion	87% accuracy	Limited interpretability
[31] Saha & Prakash (2023)	LIDC-IDRI	Multi-Attention Stacked Ensemble on cropped nodules	EffNetV2-S + DenseNet201 + MobileViT-XXS	Acc. 98.09, AUC 0.9961	Patch-based; no narratives

CHAPTER 3

METHODOLOGY

3.1 Workflow

This study adopts an end-to-end workflow that integrates context-preserving data preparation, complementary representation learning, attention-based stacking, and multi-modal explainability. The pipeline is designed to align with radiological practice, emphasize reproducibility, and produce outputs that are both performant and clinically interpretable.

Data acquisition and labelling: The de-identified LIDC-IDRI collection serves as the data source, given its status as the reference dataset for pulmonary nodule research and its rich expert annotations [2,15]. Only nodules associated with explicit malignancy scores are included; studies lacking such scores are excluded to avoid ambiguous supervision. Original XML annotations are parsed to map each nodule to its DICOM series via SeriesInstanceUID and to recover the axial position (z-index). Where multiple readers provide malignancy assessments, a continuous, reader-aggregated malignancy value is retained and subsequently thresholded to derive the binary training label, as detailed in 3.3.1. This approach preserves information about inter-observer variability that characterizes LIDC-IDRI [2].

Preprocessing and context-aware sample construction: DICOM images are converted to Hounsfield Units, clipped using a lung-appropriate window (e.g., −1000,400) and normalized on a per-scan basis to reduce inter-scanner variability frequently noted in secondary analyses of LIDC-IDRI [15]. Slices are ordered by spatial metadata to ensure consistent axial sequencing. To admit volumetric cues without incurring the computational burden of full three-dimensional networks, each training instance is formed as a pseudo-3D stack of three consecutive full axial slices centered on the annotated nodule, {z−1, z, z+1}. In contrast to the prevalent patch-based paradigm, cropping around the nodule is intentionally avoided so that vascular, lobar, and parenchymal context remains available to the model—an alignment with the way radiologists reason in practice and a strategy supported by prior evidence on pseudo-3D efficacy [15,16] and the cost of full 3D screening systems [13]. Stacks are resized to a unified resolution R×R (specified in 3.8.4) and stored with minimal metadata to support traceability (patient identifier, series UID, z-index, continuous malignancy score, binary label).

Evaluation protocol and reproducibility: Model selection and performance estimation follow patient-level, stratified 5-fold cross-validation to prevent subject overlap between training and validation partitions and to mirror clinical deployment, where a patient is

encountered once. The same folds are used across all base learners and the ensemble to ensure paired comparisons [7,26]. Random seeds, library versions, and preprocessing parameters are fixed and documented to enable exact reproduction (see 3.8.4).

Base learners: Three complementary image backbones are employed, EfficientNetV2-S, DenseNet-201, and MobileViT-XXS. Each consumes the pseudo-3D tensor [3,R,R]and is equipped with a single-logit classification head for the binary task. Training uses AdamW with a scheduled learning rate under mixed-precision (AMP), and early stopping based on validation AUC. Class imbalance is addressed without oversampling or synthetic duplication by combining loss-level reweighting—BCEWithLogitsLoss with per-fold $pos_{weight} = \frac{N_{neg}}{N_{pos}}$ and WeightedRandomSampler to obtain approximately balanced mini-batches (3.4). Image augmentations (geometric and mild photometric) are applied during training; **test-time augmentation is not used** in the final reporting, as exploratory trials did not yield consistent improvements (3.6, 3.8).

Attention-based stacked ensembling: To improve robustness and exploit complementary representations, predictions are combined using a Multi-Attention Stacked Ensemble (MASE) inspired by recent advances in lung CT classification [31]. For each validation instance within a fold, the concatenated logits from the three trained backbones serve as inputs to a lightweight meta-learner equipped with two attention pathways: model-wise attention, which assigns input-conditioned weights to each backbone, and class-wise attention, which refines the decision boundary for the binary outcome. The stacker is fitted solely on training-fold outputs and is evaluated on the corresponding validation partition, ensuring strict fold integrity. Relative to static averaging, attention-based stacking has been shown to better capitalize on heterogeneity among learners in medical imaging [7,26], and to yield state-of-the-art results in closely related settings [31].

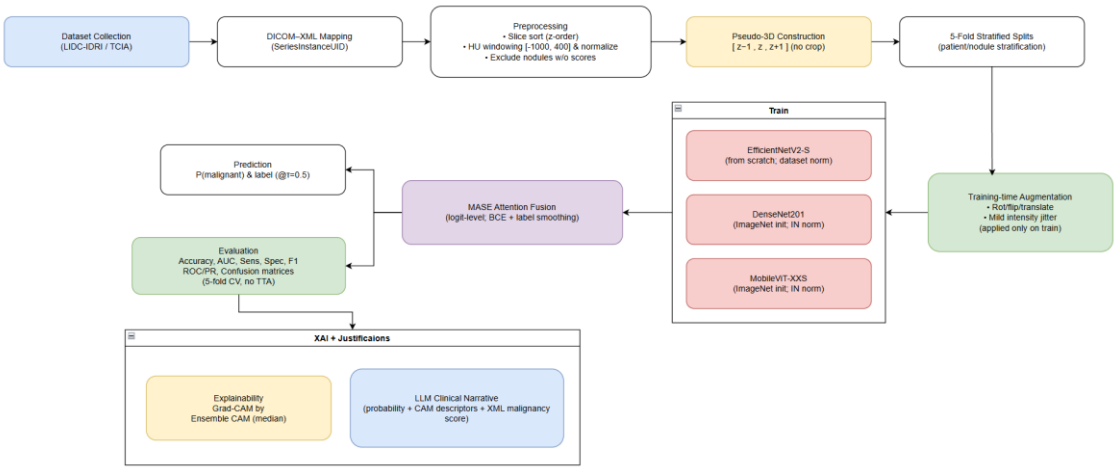
Evaluation and reporting: For each backbone and for the ensemble, the following metrics are computed per fold, area under the ROC curve (AUC), accuracy, sensitivity, specificity, F1-score, and precision. AUC is emphasized for its threshold-agnostic property and resilience to class imbalance in diagnostic tasks. Unless otherwise stated, a sigmoid threshold of 0.5 is used for class decisions. Per-sample logits, probabilities, and predicted labels are retained together with identifiers to facilitate auditing, error analysis, and linkage to explainability artefacts.

Explainability and clinician-oriented outputs: Model transparency is addressed at two levels. First, Grad-CAM heatmaps are computed for each backbone at a task-appropriate late feature layer, resampled to the input resolution, and overlaid on the axial image to reveal salient regions [19]. Second, because the deployed prediction originates from the ensemble rather than a single model, an ensemble-level attention map is produced by normalizing per-model CAMs to [0,1], resampling to the common grid, and median-stacking across the three backbones (and, when presenting aggregate visuals, across folds). Median aggregation attenuates outliers and highlights consensus foci, thereby offering a faithful visual rationale for the final decision (3.9.2). To complement visual evidence with structured justification, outputs are converted into concise clinical narratives using a large language model conditioned on the predicted probability, the final

label, succinct descriptors of the ensemble CAM (e.g., focal versus diffuse saliency; central versus peripheral location), and basic case metadata. Prompting is constrained to avoid speculation and to reflect domain-appropriate reasoning, addressing recognized barriers to adoption of black-box systems in healthcare [10,12,14,20].

Collectively, the workflow (i) preserves clinical context via full-slice, pseudo-3D inputs rather than nodule crops [15,16] while avoiding the computational demands of volumetric 3D screening networks [13]; (ii) improves generalization through an attention-based stack that adaptively exploits complementary learners [7,26,31]; and (iii) enhances interpretability by coupling Grad-CAM with clinician-readable narrative explanations [10,12,14,19,20]. These design choices are consistent with contemporary evidence and are intended to produce outputs that are performant, reproducible, and suitable for review within routine radiological workflows.

Figure 3.1 The proposed workflow



3.2 Dataset Collection

The study utilizes the publicly available, de-identified LIDC–IDRI thoracic CT collection, which provides multi-reader annotations for pulmonary nodules and has become the canonical benchmark for lung nodule analysis in computer-aided diagnosis research [2,15]. LIDC–IDRI offers per-study DICOM series together with XML files containing reader delineations and categorical ratings (e.g., malignancy, margin, sphericity), with up to four thoracic radiologists contributing assessments. The dataset's design purposefully captures inter-observer variability, a characteristic that mirrors real-world diagnostic uncertainty and is pertinent to model development and evaluation [2].

Eligibility criteria: Inclusion was restricted to nodules with explicit malignancy scores recorded in the LIDC annotations, ensuring clinically meaningful supervision. Studies or annotations lacking a malignancy score were excluded. Consistent with the dataset's annotation protocol, malignancy ratings are typically provided for nodules ≥3 mm; "non-

nodules" and small nodules without malignancy ratings were not considered for training or evaluation [2]. No additional external datasets were used.

Acquisition and mapping: The dataset was obtained in its native structure (patient folders containing one or more CT DICOM series plus XML annotations). For each annotated nodule, the corresponding XML file was parsed to extract (i) the SeriesInstanceUID to identify the correct CT series, (ii) the axial slice location (z-index) for the nodule-centered slice, and (iii) the reader-specific malignancy ratings. Where multiple readers were available for a given nodule, the malignancy assessments were retained as a **continuous value** by aggregating across readers (e.g., mean), postponing dichotomization to the label construction step (see 3.3.1). This preserves information about disagreement while enabling a consistent binarization policy for supervised learning on the classification task [2,15].

Data integrity and completeness: Prior to preprocessing, each referenced DICOM series was validated for essential metadata needed to convert to Hounsfield Units (e.g., rescale slope and intercept) and to reliably order slices (spatial position tags). Series with missing critical tags or with corrupted files were discarded. For transparency and reproducibility, the full accession log (patient/series identifiers retained in hashed form), counts of included/excluded nodules, and reasons for exclusion are maintained alongside the codebase and will be provided upon request.

Cohort construction and traceability: After eligibility filtering and XML–DICOM mapping, each retained nodule defines a case anchored at its annotated axial index. Downstream steps construct pseudo-3D, full-slice inputs by stacking adjacent slices around this index; no cropping to the nodule is performed (details in 3.3). For every case, a compact metadata record is stored—comprising a patient-level identifier, series UID, axial index, the aggregated malignancy score, and the derived binary label—so that predictions, Grad-CAM overlays, and narrative explanations can be traced back to their source scans without exposing PHI.

Partitioning policy: Although partitioning is operationalized during training (see 3.8.3), the dataset collection step enforces a patient-level identity key to guarantee that all subsequent splits are performed per patient rather than per slice or per nodule. This prevents inadvertent cross-contamination between training and validation folds and reflects intended clinical deployment, where a single patient should be evaluated once per encounter [2,15].

Ethical considerations: LIDC–IDRI is fully de-identified and publicly released for research use; therefore, this work did not involve human subjects research as defined by institutional policy and did not require additional IRB review. All processing adhered to the dataset's terms of use and community best practices for handling medical images.

This collection protocol produces a curated cohort of nodule-centered CT studies with consistent malignancy supervision, reliable XML–DICOM linkage, and patient-level identifiers to support stratified cross-validation. The resulting cohort is expressly tailored to the subsequent context-preserving pseudo-3D representation (3.3), class-balanced

training regime (3.4), and attention-based ensembling (3.7.5), while retaining traceability for explainability analyses and clinical narrative generation.

3.3 Dataset Preprocessing

Preprocessing converts raw DICOM series into standardized, context-preserving inputs suitable for supervised learning while maintaining traceability to the underlying scans and annotations. All operations are deterministic and version-locked (software and parameters detailed in 3.8.4).

DICOM to Hounsfield Units and intensity standardization: Each axial slice is converted from vendor-specific stored values to Hounsfield Units (HU) using the DICOM rescale parameters (slope $\boldsymbol{a}$, intercept $\boldsymbol{b}$):

$$\boldsymbol{HU(p) = a \cdot raw(p) + b}$$

where $\boldsymbol{p}$ indexes pixels. Following conversion, slices are clipped to a lung-appropriate window [L,U] (empirically set to L=−1000, U=400) and linearly mapped to [0,1] to reduce inter-scanner variability frequently reported in secondary analyses of LIDC–IDRI [15]:

$$\tilde{\boldsymbol{x}}(\boldsymbol{p}) = min\{max[\boldsymbol{HU}(\boldsymbol{p}), \boldsymbol{L}], \boldsymbol{U}\}, \qquad \boldsymbol{x}(\boldsymbol{p}) = \frac{\tilde{\boldsymbol{x}}(\boldsymbol{p}) - \boldsymbol{L}}{\boldsymbol{U} - \boldsymbol{L}} \in [\boldsymbol{0}, \boldsymbol{1}].$$

This per-scan normalization preserves relative contrast within the lung window while ensuring comparable dynamic range across studies.

Axial ordering and spatial consistency: Slices are ordered using spatial tags (e.g., *Image Position (Patient)*), resolving ties by instance number when necessary. Series lacking essential metadata (rescale slope/intercept or positional tags) or exhibiting non-monotonic axial geometry are excluded (see 3.2). Orientation is preserved per DICOM convention; no left–right flips are applied.

Context-preserving pseudo-3D construction (no cropping): For each annotated nodule (identified in the XML with a consensus axial index z_i, 3.2), a pseudo-3D input is formed by stacking the full axial slice at z_i together with its immediate neighbors:

$$\boldsymbol{X^{(i)} = [I(z_i - 1), I(z_i), I(z_i + 1)] \in R^{H \times W \times 3}}$$

where $\boldsymbol{I}(\cdot)$ denotes the preprocessed slice and the channel axis encodes limited through-plane context. When a neighbor is unavailable at the volume boundary, the nearest available slice is replicated to maintain three channels. In contrast to prevalent patch-based methods, no spatial cropping around the nodule is performed; the entire slice is retained to preserve vascular, lobar, and parenchymal cues used in radiological reasoning [15,16]. This design yields contextual information comparable to limited-depth volumetrics at a fraction of the computational cost of full 3D pipelines [13].

Geometric standardization. Each stack $\boldsymbol{X}^{(i)}$ is resampled to a common in-plane resolution $\boldsymbol{R} \times \boldsymbol{R}$ using bicubic interpolation:

$$\widehat{\boldsymbol{X}}^{(i)} = \mathcal{R}_{\boldsymbol{R}}\left(\boldsymbol{X}^{(i)}\right) \in \boldsymbol{R}^{\boldsymbol{R}\times\boldsymbol{R}\times\boldsymbol{3}}$$

with $\boldsymbol{R}$ fixed for all backbones (reported in 3.8.4). Because the task is classification rather than morphometric measurement, in-plane resampling is preferred to full isotropic resampling; this avoids introducing additional interpolation artefacts through the slice thickness dimension.

Sample packaging and traceability. For each case iii, the preprocessed tensor $\hat{x}^{(i)}$ is stored with a compact metadata record:

$$\mathbf{meta}^{(\mathbf{i})} = \{\mathbf{patient\_id}, \mathbf{series\_uid}, \boldsymbol{z_i}, \boldsymbol{s_i}, \boldsymbol{y_i}\}$$

where $\boldsymbol{s_i}$ is the aggregated malignancy score (continuous) and $\boldsymbol{y_i}$ the derived binary label (see 3.3.1). This enables deterministic linkage between inputs, predictions, Grad-CAM overlays, and source scans, without exposure of PHI. The resulting dataset consists of nodule-anchored, full-slice pseudo-3D tensors suitable for patient-level stratified cross-validation (3.8.3).

3.3.1 Label Construction

LIDC–IDRI provides reader-specific malignancy ratings on a 1–5 ordinal scale for nodules ≥3 mm, with up to four thoracic radiologists contributing assessments [2]. To exploit the multi-reader design while providing a stable binary supervision signal, labels are constructed in two steps.

(i) Reader aggregation: For nodule i, let $\{r_{ij}\}_{j=1}^{m_i}$ be the available reader scores ($m_i \in \{1, \ldots, 4\}$). The continuous malignancy estimate is the simple mean

$$s_i = \frac{1}{m_i} \sum_{j=1}^{m_i} r_{ij}$$

which retains information about inter-observer disagreement typical of this dataset [2] and avoids premature discretization.

(ii) Dichotomization: The binary class label $y_i \in \{0, 1\}$ is obtained by thresholding s_i at a fixed cut-point τ chosen a priori to reduce label noise from "indeterminate" scores:

$$y_i = \mathbf{1}[s_i \geq \tau].$$

In this study, τ=3.5 is used, mapping means dominated by ratings 4–5 to the malignant class and 1–3 to the benign class. This choice follows common practice to down-weight borderline cases without discarding data, and was kept constant across folds to ensure comparability. (A sensitivity analysis around τ can be incorporated without affecting the remainder of the pipeline.)

For completeness, class prevalence is computed per training fold after dichotomization and used to derive the positive-class weight for the loss in 3.8.1. Specifically, letting $N_{\text{train}}^{\text{pos}}$ and $N_{\text{train}}^{\text{neg}}$ denote the number of positive and negative samples in the training partition of a fold, the scalar weight applied in **BCEWithLogitsLoss** is

$$\text{pos\_weight} = \frac{N_{\text{train}}^{\text{neg}}}{N_{\text{train}}^{\text{pos}}}$$

which counteracts imbalance without duplicating samples (the complementary use of a WeightedRandomSampler to balance mini-batches is described in 3.4).

This labelling protocol preserves the advantages of LIDC–IDRI's multi-reader design while yielding a reproducible binary target for supervised learning. It is aligned with the subsequent context-preserving pseudo-3D representation, class-balanced optimization, and patient-level validation (3.8.3), thereby supporting both methodological rigor and clinical plausibility.

3.4 Dataset Balancing

After dichotomization of the multi-reader malignancy scores at τ=3.5 the positive and negative classes in LIDC–IDRI are not evenly represented and the degree of imbalance varies per fold under patient-level stratified cross-validation. To address this without altering the empirical prevalence or duplicating samples, the study combines loss-level class reweighting with balanced mini-batch sampling, and does not employ oversampling, synthetic data generation, or MixUp in the final pipeline. This choice preserves data diversity, mitigates minority under-representation during optimization, and avoids the instability observed in preliminary experiments with mixup-style schemes.

Loss-level class reweighting: Let $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the training partition of a given fold, with $y_i \in \{0,1\}$ and model logit $z_i = f_\theta(x_i)$. For that fold, let $N_{\text{pos}} = \sum_i 1[y_i = 1]$ and $N_{\text{neg}} = \sum_i 1[y_i = 0]$. The empirical risk minimized during training is the **class-weighted** logistic (BCE-with-logits) loss

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(w_+ \, y_i . l_+(z_i) + w_- \, (1 - y_i) . l_-(z_i)\right),$$

where

$$\mathrm{l}_+(z) = log(1 + e^{-z}), \qquad \mathrm{l}_-(z) = \log(1 + e^{z}),$$

and the class weights are chosen to counteract imbalance,

$$w_+ = \frac{N_{\text{neg}}}{N_{\text{pos}}}, \qquad w_- = 1$$

In the PyTorch implementation (BCEWithLogitsLoss), this corresponds to setting $pos\_weight = N_{\text{neg}}/N_{\text{pos}}$ per fold, recomputed after label construction (3.3.1). This reweighting equalizes, in expectation, the gradient contribution of positive and negative instances, improving sensitivity without inflating the dataset.

Balanced mini-batch sampling: In parallel, batches are drawn with a WeightedRandomSampler to approximate class balance at the batch level (targeting ≈1:1 positives to negatives). Concretely, define sampling probabilities

$$\Pr(i \text{ selected}) = \begin{cases} \dfrac{\alpha}{N_{\text{pos}}}, & y_i = 1, \\ \dfrac{1-\alpha}{N_{\text{neg}}}, & y_i = 0, \end{cases} \quad \text{with} \quad \alpha \approx \tfrac{1}{2}.$$

This maintains a steady presence of minority-class examples in each update, stabilizing optimization in the presence of skew. Because loss reweighting is already compensating

for prevalence, the batch sampler is set to *approximately* balanced (rather than aggressively oversampling), avoiding over-correction.

Effect on decision threshold and metrics: Class weighting modifies the optimization landscape but **does not** change the monotonic mapping from logit to probability; therefore, AUC—our primary model-selection criterion—remains threshold-agnostic. Unless otherwise stated, a decision threshold of 0.50.50.5 is applied to the sigmoid outputs at inference for reporting accuracy, sensitivity, specificity, F1-score, and precision (3.8). Because class-weighted training can influence probability calibration, emphasis is placed on AUC and operating-point metrics rather than uncalibrated probabilities; optional post-hoc calibration can be incorporated later without altering the training recipe.

3.5 Feature Representation

The feature representation is designed to preserve thoracic context while providing limited through-plane information at negligible computational overhead. Each training instance is encoded as a three-channel, full-slice pseudo-3D tensor that can be consumed by modern 2D backbones initialized from large-scale pretraining [4]. This section specifies the construction, normalization, and rationale of the representation.

Channel construction (pseudo-3D): For an annotated nodule with axial index z_i (3.2–3.3), the preprocessed, lung-windowed slice at position z is denoted $I(z) \in [0,1]^{H\times W}$. The input tensor is formed by stacking the entire axial slices at $(z_i - 1,\ z_i,\ z_i + 1)$ as channels:

$X^{(i)} = [\,I(z_i - 1),\ I(z_i),\ I(z_i + 1)] \in R^{H\times W\times 3}$,followed by in-plane resampling to a unified resolution R×RR\times RR×R (bicubic) and framework-compatible permutation to channel-first layout,

$$\hat{X}^{(i)} = \mathcal{R}_{\mathcal{R}}\left(X^{(i)}\right) \in R^{3\times R\times R}.$$

At volume boundaries $where\ z_i \pm 1$ is unavailable, the nearest valid slice is replicated to maintain three channels. No spatial cropping is applied; full axial slices are retained to preserve vascular, lobar, and parenchymal cues that radiologists use in practice [15,16], while avoiding the computational burden of full 3D CNNs [13].

Intensity normalization and backbone compatibility: Prior to stacking, DICOM values are converted to HU and clipped to a lung window [L,U]=[-1000,400], then

linearly mapped to [0,1] (3.3). To leverage ImageNet-initialized weights in EfficientNetV2-S, DenseNet-201, and MobileViT-XXS, inputs are further standardized using the channel-wise affine normalization employed by these backbones:

$$\hat{X}^{(i)}_{c,u,v} \leftarrow \frac{\hat{X}^{(i)}_{c,u,v} - \mu_c}{\sigma_c}, \quad c \in \{1,2,3\},$$

With (μ_c, σ_c) set to the backbone's default normalization constants (kept identical across channels). Although the three channels correspond to adjacent CT slices rather than RGB, this affine mapping aligns the dynamic range with the statistics expected by the pretrained convolutional filters, a practice shown to expedite convergence in medical-image transfer learning [4]. No histogram equalization or heavy gamma curves are applied, as such manipulations can distort attenuation cues.

Through-plane cues captured by early filters: Because channels encode adjacent axial positions, the first convolutional layer can approximate discrete through-plane derivatives. Let $x_c(u, v)$ denote the intensity at pixel (u,v) in channel c of $\hat{X}^{(i)}$, where c=1,2,3 corresponds to $z_i - 1$, z_i, $z_i + 1$. A 1×1 kernel with weights (−1,0,+1) across channels computes

$\Delta_z x(u, v) \approx x_3(u, v) - x_1(u, v) \approx I(z_i + 1, u, v) - I(z_i - 1, u, v)$, a central-difference proxy of $\partial I / \partial z$. More general k × k kernels combine this through-plane signal with local in-plane texture, allowing the network to encode subtle morphology changes across slices (e.g., spiculation continuity, vessel attachment), which are diagnostically relevant. Thus, the pseudo-3D encoding supplies volumetric cues at negligible extra cost compared with single-slice 2D, while remaining far lighter than volumetric 3D CNNs [13,16].

Global contextual field: Using full-slice inputs (rather than cropped patches) ensures that the effective receptive field of deeper layers includes anatomical context beyond the nodule itself (hilum–periphery gradients, lobar boundaries, vessel trajectories). For a convolutional stack with stride pattern $\{s_\ell\}$ and kernel sizes$\{k_\ell\}$, the theoretical receptive field after LLL layers is

$$\mathrm{RF_L} = 1 + \sum_{\ell=1}^{L} \left(\prod_{j=1}^{\ell-1} s_j \right) (k_\ell - 1),$$

which, on full slices, can encompass long-range structures that are completely absent in patch crops. This property is central to the proposed "no-crop" design, as it enables joint modeling of nodule appearance and its anatomic milieu [15,16].

Feature tensor and metadata binding: For each case iii, the model-ready tensor and its minimal metadata are stored as

$$\left(\hat{X}^{(i)} \in R^{\mathbb{3}\times R\times R}, \ \mathrm{meta^{(i)}} = \{\text{patient\_id, series\_uid, } z_i, s_i, y_i\} \right),$$

where s_i is the aggregated malignancy score and $y_i y$ the derived binary label (3.3.1). The metadata linkage supports traceability of predictions, Grad-CAM overlays, and clinical narratives back to source scans without exposing PHI, which is essential for qualitative review (3.9).

The representation is intentionally minimal yet expressive: (i) it preserves contextual information required by radiologists (full slices) [15], (ii) it encodes limited depth cues (three adjacent slices) shown to approximate volumetric information with far lower compute than 3D CNNs [16,13], and (iii) it remains fully compatible with pretrained 2D backbones and attention-based ensembling (3.7.5), avoiding architectural changes that would hinder reproducibility. The same tensor schema is used for all base learners and folds, simplifying cross-model comparison and facilitating ensemble-level Grad-CAM aggregation (3.9.2).

3.6 Applied Algorithms

The methodology employs three complementary 2D backbones—EfficientNetV2-S, DenseNet-201, and MobileViT-XXS—trained on full-slice pseudo-3D inputs (3.3–3.5), followed by an attention-based stacked ensemble that combines their outputs into a single prediction. The choice of backbones reflects diversity in inductive biases and computational profiles: compound-scaled efficient CNNs (EfficientNetV2) for strong accuracy/latency trade-offs, densely connected feature reuse (DenseNet) for gradient flow and parameter efficiency, and a lightweight CNN-Transformer hybrid (MobileViT) for long-range dependencies with mobile-scale capacity [33–35]. Each backbone is adapted with a single-logit head for binary classification and trained under the same optimization, balancing, and augmentation regime (3.4, 3.6, 3.8). Predictions are fused with a Multi-Attention Stacked Ensemble (MASE) that learns *model-wise* and *class-wise* importance, building on ensemble theory [7], medical-imaging ensembles [26], and recent attention-stacking results in lung CT [31]. This section formalizes the components and their integration; detailed architectural notes for each backbone follow in 3.7.2–3.7.4.

Let $X \in R^{3\times R\times R}$ denote a pseudo-3D input (three consecutive full axial slices; 3.5). For the k-th backbone k∈{1,2,3}), denote parameters θ_k, logit $z_k = f_k(X;\theta_k) \in R$ and probability $p_k = \sigma(z_k)$ with $\sigma(t) = 1/(1+e^{-t})$.The vector of base logits is

$$z = [z_1, z_2, z_3]^\top \in R^3.$$

The stacked ensemble consumes **z** (out-of-fold when training the meta-learner) and outputs an ensemble logit z_* used for the final decision (3.7.5).

All base learners are trained with class-weighted BCE-with-logits using per-fold $\text{pos\_weight} = N_{\text{neg}}/N_{\text{pos}}$ and WeightedRandomSampler for approximately balanced mini-batches (no oversampling, no MixUp in the final configuration; 3.4). Augmentations are conservative (rigid/mild photometric, slice-coherent) and applied only at training time (3.6). The same patient-level 5-fold partitions are used across all models to enable paired comparisons and a clean stacking protocol (3.8.3).

3.6.1 Proposed Model

Base learners: Each backbone implements a mapping

$$f_k\colon R^{3\times R\times R} \to R, \qquad z_k = f_k(X;\theta_k),$$

comprising (i) a feature extractor ϕ_k (.) (EfficientNetV2-S / DenseNet-201 / MobileViT-XXS) operating on the pseudo-3D tensor and (ii) a single-logit classification head $g_k(\cdot), i.e., z_k = g_k\big(\phi_k(X)\big)$. Training minimizes the class-weighted empirical risk

$$\min_{\theta_k} \frac{1}{N} \sum_{i=1}^{N} \big(w_+ y_i \, log(1 + e^{-z_{k,i}}) + w_-(1 - y_i) log(1 + e^{z_{k,i}})\big),$$

with $w_+ = N_{\text{neg}}/N_{\text{pos}}$, $w_- = 1$ computed per fold (3.4). Optimization uses AdamW with a scheduled learning rate under mixed precision; early stopping monitors validation AUC (3.8).

Attention-based stacking (MASE): To fuse predictions, the ensemble learns (a) model-wise attention that allocates input-conditioned weights across backbones, and (b) class-wise attention that permits class-specific weighting of backbone evidence, followed by a lightweight meta-learner. Formally, let $z \in R^3$ be the concatenated logits for a case $\boldsymbol{X}$, a small bottleneck network produces a latent representation

$$\boldsymbol{h} \;=\; \rho(W_h\, z + b_h) \in R^d,$$

where $W_h \in R^{d\times 3}, b_h \in R^d, and\ \rho$ is a pointwise nonlinearity (ReLU).

- **Model-wise attention:** Attention coefficients over backbones are obtained as

$$\boldsymbol{\alpha} \;=\; \text{softmax}(W_a\, h + b_a) \;\in\; R^3,$$

with $W_a \in R^{3\times d}, b_a \in R^3$, so that $\sum_k \alpha_k = 1,\ \alpha_k \geq 0$. The attended backbone evidence is $\tilde{z} = \alpha \odot z$.

- **Class-wise attention:** For the binary task, class-specific scores are produced by a linear map

$$\boldsymbol{o} = W_c\,\tilde{z} + b_c \in R^{2}, \qquad W_c \in R^{2\times 3},\ b_c \in R^{2}.$$

Here, the rows of WcW_cWc constitute **class-wise attention over models** (one row per class), enabling the stack to emphasize different backbones for benign vs. malignant evidence.

- **Meta-learner / final logit:** A single ensemble logit compatible with **BCEWithLogitsLoss** is formed as the logit difference

$$z_* = o_{\text{mal}} - o_{\text{ben}} = \left(\mathbf{w}_{\text{mal}} - \mathbf{w}_{\text{ben}}\right)^{\top}\tilde{\mathbf{z}} + \left(b_{\text{mal}} - b_{\text{ben}}\right),$$

where $\mathbf{w}_{\text{mal}}, \mathbf{w}_{\text{ben}}$ are the rows of W_c. Equivalently, z_* can be viewed as a linear meta-learner operating on the model-wise attended logits. The ensemble probability is $p_*=\sigma(z_*)$

Training protocol for stacking: For each cross-validation fold, base learners are trained on the training partition only. The stacker is then fitted on out-of-fold base logits drawn from the same training partition (e.g., via internal K-fold on the training fold or by using the base models' validation logits), and evaluated on the fold's validation partition—ensuring that no case contributes its own base prediction to its ensemble target (strict fold integrity). The stacker optimizes the same class-weighted BCE loss as the bases, with the positive weight recomputed for the stacker's training subset (3.4). No test-time augmentation is applied at inference; the final decision is made from p_* at a default threshold of 0.5 (unless otherwise specified), with AUC as the primary model-selection criterion (3.8).

The architecture leverages (i) full-slice pseudo-3D inputs to expose contextual cues central to radiological reasoning (3.3–3.5; [15,16,13]); (ii) heterogeneous backbones to diversify representational biases [33–35]; and (iii) attention-based stacking to move beyond static averaging, consistent with ensemble theory [7], empirical advantages of medical-imaging ensembles [26], and prior lung-CT evidence for multi-attention stacking [31]. The formulation above is deliberately lightweight (few trainable parameters at the stacker) to minimize overfitting while capturing complementary backbone strengths.

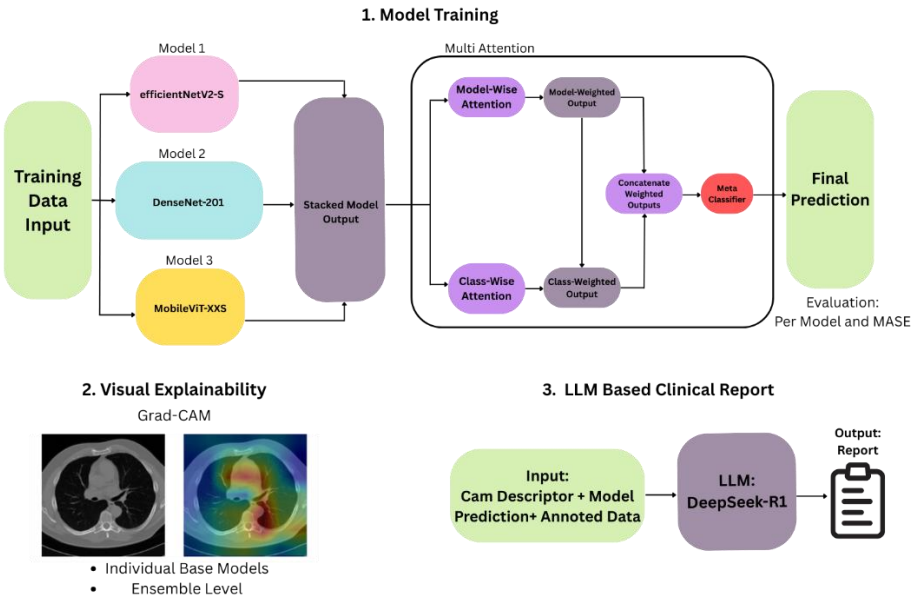


Figure 3.2 Proposed Framework

3.6.2 EfficientNetV2-S

Architecture and initialization: EfficientNetV2-S is adopted as a high-accuracy, compute-efficient backbone within the proposed pseudo-3D, full-slice framework. The network follows the EfficientNetV2 family's design that combines Fused-MBConv blocks in early stages for faster training with MBConv (inverted residual with depthwise separable convolution and squeeze-and-excitation) in later stages, together with improved compound scaling and progressive learning heuristics [33]. Models are initialized from ImageNet pretraining to leverage transferable low-level filters and regularization effects typical in medical imaging transfer learning [4]. Inputs are three-channel pseudo-3D tensors $\hat{X} \in R^{\mathbb{3}\times R\times R}(3.3-3.5)$, normalized with the backbone's affine channel statistics (applied identically across the three axial channels).

Block structure: Let $x \in R^{C \times H \times W}$ denote the input to a block. For an MBConv block with expansion ratio t, channel dimension is first expanded to tC via a pointwise 1×1 convolution, followed by a depthwise $k \times k$ times convolution, squeeze-and-excitation (SE), and a projection 1×1 back to C' channels, with a residual skip when shape-compatible. Using $\phi(\cdot)$ to denote the nonlinearity (SiLU), the main steps are:

$$\begin{aligned}
\mathbf{u} &= \phi\big(\mathrm{Conv}_{1\times1}^{C\to tC}(\mathbf{x})\big), \\
\mathbf{v} &= \phi\big(\mathrm{DWConv}_{k\times k}^{tC\to tC}(\mathbf{u})\big), \\
\mathbf{v}' &= \mathrm{SE}(\mathbf{v}), \\
\mathbf{y} &= \mathrm{Conv}_{1\times1}^{tC\to C'}(\mathbf{v}'), \\
\tilde{\mathbf{y}} &= \begin{cases} \mathbf{x}+\mathbf{y}, & \text{if } (C'=C)\wedge(\text{stride}=1), \\ \mathbf{y}, & \text{otherwise.} \end{cases}
\end{aligned}$$

The **SE** gating performs channel-wise re-weighting via global average pooling and a bottleneck MLP [33]:

$$z_c = \frac{1}{HW}\sum_{u,v} \mathrm{b}v_{cuv}\,, \quad s = \sigma\big(W_2\,\delta(W_1\,z)\big), \quad v'_{cuv} = s_c\,v_{cuv}$$

where δ and σ are ReLU and sigmoid, respectively. In Fused-MBConv, the expansion 1×1 and depthwise k×k steps are replaced by a single fused k×k convolution (with expansion), improving early-stage throughput while retaining representational capacity [33]. Stride and kernel configurations follow the standard EfficientNetV2-S specification.

Spatial-context handling with full slices: Because inputs are full axial slices (no cropping), the effective receptive field compounds across stages to cover long-range anatomic context (lobar boundaries, vasculature) in addition to local nodule texture (3.5). With stride pattern $\{s_\ell\}$ and kernel sizes $\{k_\ell\}$, the theoretical receptive field after L layers is

$$\mathrm{RF_L} \;=\; 1 + \sum_{\ell=1}^{L}\left(\prod_{j=1}^{\ell-1} s_j\right)(k_\ell - 1),$$

which, on full slices, allows joint modeling of nodule and milieu—an advantage over patch-cropped inputs [15,16].

Classification head and output: Let $F \in R^{C^*\times H^*\times W^*}$ be the final feature map. A global average pooling produces $g \in R^{C^*}$ with components $g_c = \frac{1}{H^*W^*}\sum_{u,v} F_{cuv}$. The binary logit is

$$z \;=\; w^{\top g} + b, \qquad p = \sigma(z),$$

where σ is the logistic sigmoid. This single-logit head is used with BCEWithLogitsLoss and per-fold positive class weight $w_+ = N_{\text{neg}}/N_{\text{pos}}$ (3.4), ensuring consistent handling of imbalance across all backbones.

Optimization and training regime: EfficientNetV2-S is fine-tuned end-to-end (all layers trainable) with AdamW, a scheduled learning rate under mixed precision, early stopping on validation AUC, and WeightedRandomSampler to approximate class balance in mini-batches (3.4, 3.6, 3.8). Data augmentation is conservative and slice-coherent across the three channels (rigid, mild photometric; no MixUp; no TTA). The model is trained within the patient-level 5-fold protocol shared by all backbones (3.8.3), enabling paired comparisons and clean stacking.

Compatibility with explainability: For Grad-CAM analysis (3.9.1), the target layer is selected from the final MBConv/Fused-MBConv stage to maximize spatial resolution while remaining semantically deep. Given gradients $\partial z/\partial \mathbf{F}_{cuv}$ standard Grad-CAM weights α_c and the class-activation map MMM are computed as [19]:

$$\alpha_c = \frac{1}{H^* W^*} \sum_{u,v} \frac{\partial z}{\partial \mathbf{F}_{cuv}}, \qquad M(u,v) = \text{ReLU}\Big(\sum_c \alpha_c \, \mathbf{F}_{cuv} \Big),$$

which is then normalized, upsampled to R×R, and overlaid on the axial slice. These per-model CAMs feed the ensemble-level aggregation (3.9.2).

EfficientNetV2-S supplies a strong accuracy–latency baseline with modern training dynamics (fused blocks, SiLU, SE) and complements the representational biases of DenseNet-201 (dense feature reuse) and MobileViT-XXS (lightweight convolution-Transformer hybrid). Under attention-based stacking (3.7.5), the ensemble can emphasize EfficientNetV2-S features for cases where its inductive bias is most informative, while deferring to alternative backbones for other morphologies—consistent with ensemble theory [7], empirical gains in medical-imaging ensembles [26], and attention-stacking evidence in lung CT [31].

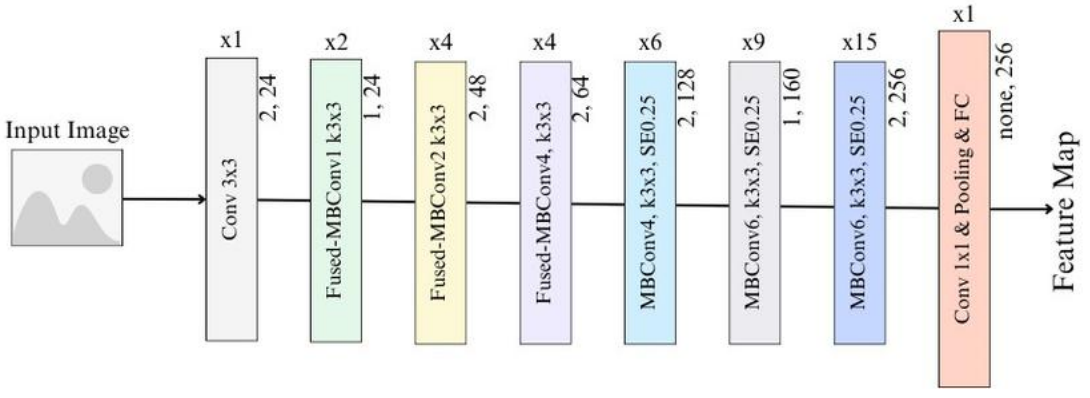


Figure 3.3 EfficientNetV2 Architecture

3.6.3 DenseNet201

Architecture and initialization: DenseNet-201 is employed to provide a complementary inductive bias based on dense feature reuse and enhanced gradient flow [34]. The network comprises a stem convolution followed by four dense blocks interleaved with transition layers. Models are initialized from ImageNet weights to leverage transferable low-level filters known to stabilize optimization in medical image transfer learning [4]. Inputs are the pseudo-3D, full-slice tensors $\hat{X} \in R^{3\times R\times R}$ (3.3–3.5), normalized with the backbone's channel-wise affine statistics applied identically across the three axial channels.

Dense connectivity and growth rate: Within a dense block, each layer receives as input a concatenation of all preceding feature maps and contributes k new feature channels (the **growth rate**). Let x_0 denote the block input and x_ℓ the output of the ℓ-th layer, ℓ=1,…,L. The dense connectivity is

$$, x_l \;=\; \mathcal{H}_\ell([x_0,\, x_1,\, \ldots,\, x_{l-1}]),$$

where [·] denotes channel concatenation, and $\mathcal{H}_\ell$ is a composite function (BatchNorm–ReLU–Conv) defined below. After ℓ\ell$\ell$ layers the block channel dimension is

$C\ell = C_0 + \ell\, k,$

with C_0 the incoming channel count. DenseNet-201 adopts bottleneck layers ("BN–ReLU– 1×1 Conv") before the 3×3 convolution to reduce computation:

$$\begin{aligned}
\tilde{\mathbf{x}}_\ell &= \mathrm{ReLU}\big(\mathrm{BN}([\mathbf{x}_0, \ldots, \mathbf{x}_{\ell-1}])\big),\\
\mathbf{u}_\ell &= \mathrm{Conv}_{1\times1}^{C_{\ell-1}\to 4k}(\tilde{\mathbf{x}}_\ell),\\
\hat{\mathbf{x}}_\ell &= \mathrm{ReLU}\big(\mathrm{BN}(\mathbf{u}_\ell)\big),\\
\mathbf{x}_\ell &= \mathrm{Conv}_{3\times3}^{4k\to k}(\hat{\mathbf{x}}_\ell).
\end{aligned}$$

This BN–ReLU–Conv (a.k.a. "pre-activation") arrangement improves optimization and reduces vanishing gradients by providing short paths from early to late layers via concatenations [34].

Transition layers and compression: Between dense blocks, a transition layer reduces spatial resolution and compresses channels:

$$\mathbf{y} \;=\; \mathrm{AvgPool}_{2\times2}\Big(\mathrm{Conv}_{1\times1}^{C\to\lfloor\theta C\rfloor}(\mathrm{ReLU}(\mathrm{BN}(\mathbf{x})))\Big)\,,$$

where $\theta \in (0,1]$ is the compression factor (DenseNet-201 uses θ=0.5) , striking a balance between capacity and parameter efficiency. This progressive downsampling grows the theoretical receptive field, which—on full slices—permits joint modeling of nodule appearance and broader anatomic context (3.5):

$$\mathrm{RF}_L \;=\; 1 + \sum_{\ell=1}^{L}\left(\prod_{j=1}^{\ell-1} s_j\right)(k_\ell - 1).$$

Classification head and output: Let F$\in R^{C^*\times H^*\times W^*}$ be the final feature map of the last dense block. Global average pooling yields g $\in$ R^{C^*} with components $g_c = \frac{1}{H^*W^*}\sum_{u,v} F_{cuv}$. The network's binary logit is

$$z = w^{\top}\boldsymbol{g} + b, \qquad p = \sigma(z),$$

and is trained with BCEWithLogitsLoss using per-fold $pos\_weight = N_{\text{neg}}/N_{\text{pos}}$ (3.4). Training employs AdamW, a scheduled learning rate, mixed precision, early stopping on validation AUC, and WeightedRandomSampler to approximate batch-level balance (3.4, 3.6, 3.8).

Compatibility with pseudo-3D full-slice inputs: Dense connectivity encourages feature reuse from early layers, which in this setting encode low-level attenuation patterns and vessel–parenchyma interfaces derived from the three axial channels. Because the channels correspond to {z−1,z,z+1}, the first convolution can combine through-plane cues with in-plane texture (the central-difference intuition in 3.5), while subsequent dense concatenations propagate these cues forward without re-learning them, improving parameter efficiency relative to plain residual stacks.

Grad-CAM target and computation. For interpretability (3.9.1), the Grad-CAM target is chosen from the final dense block to preserve spatial detail at semantically rich depth. Given gradients $\partial z/\partial F_{cuv}$, the standard Grad-CAM weights and map are [19]:

$$\alpha_c = \frac{1}{H^*W^*}\sum_{u,v}\frac{\partial z}{\partial \mathbf{F}_{cuv}}, \qquad M(u,v) = \mathrm{ReLU}\Big(\sum_c \alpha_c \mathbf{F}_{cuv}\Big),$$

followed by normalization and upsampling to R×R for overlay on the axial slice. These per-model CAMs are subsequently median-stacked with those from the other backbones to form the ensemble-level attention map (3.9.2).

DenseNet-201 contributes stable gradient propagation and strong feature reuse, tendencies that complement EfficientNetV2-S's compound scaling and MobileViT-XXS's lightweight global context. Under attention-based stacking (3.7.5), the ensemble can allocate higher model-wise attention to DenseNet-201 on cases where dense reuse is advantageous (e.g., fine-grained parenchymal textures), while deferring to alternative backbones otherwise—consistent with ensemble theory [7], empirical gains of medical-imaging ensembles [26], and the multi-attention stacking paradigm reported for lung CT [31].

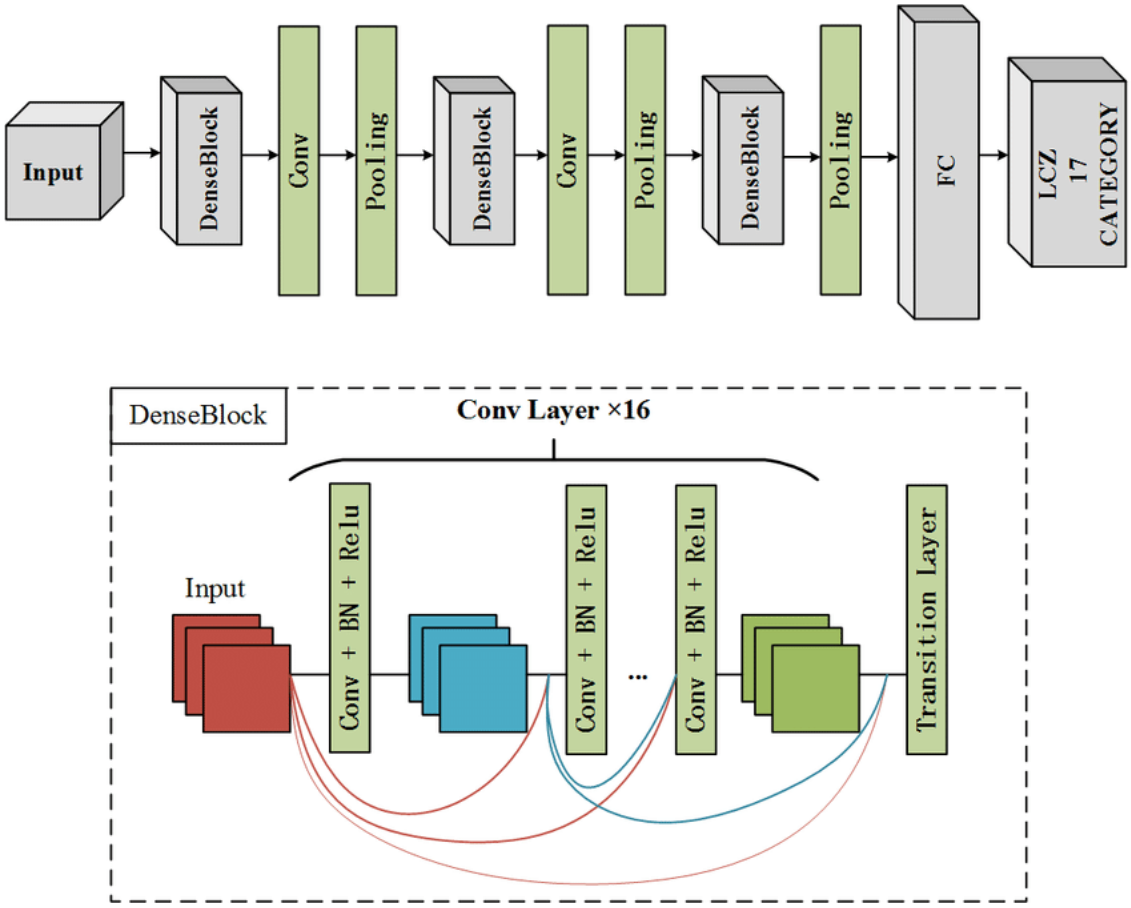


Figure 3.4 DenseNet201 Architecture

3.6.4 MobileViT-XXS

Architecture and initialization: MobileViT-XXS is a lightweight CNN–Transformer hybrid that alternates local convolutional processing with global self-attention over unfolded feature patches, thereby capturing long-range dependencies at mobile-scale capacity [35]. The backbone is initialized from ImageNet pretraining to leverage transferable low-level filters and stable optimization in medical transfer learning [4]. Inputs are the pseudo-3D, full-slice tensors $\hat{X} \in R^{3\times R\times R}$ (3.3–3.5), normalized with the backbone's channel-wise affine statistics applied identically to the three axial channels.

MobileViT block: local–global–local design: Let $F \in R^{C\times H\times W}$ be the incoming feature map to a MobileViT block after a short convolutional stem (e.g., depthwise + pointwise). The block proceeds in three steps:

1. **Local representation (conv):** a 3×3 conv refines local texture,

$\boldsymbol{F_{loc}} = \boldsymbol{\phi}(\text{Conv}_{3\times3}(\boldsymbol{F}))$,

where ϕ denotes SiLU.

2. **Global representation (Transformer over patches):** $\boldsymbol{F_{loc}}$ is unfolded into non-overlapping $p \times p$ patches and flattened into tokens. With $N = (H/p) \cdot (W/p)$ patches and token dimension ddd, define the linear patch embedding

$$\boldsymbol{Z} = \mathcal{P}F_{\text{loc}}\,E \in R^{N\times d}, \qquad E \in R^{(p^2C)\times d},$$

where P(·) extracts and flattens each p×pp\times pp×p patch. A standard Transformer encoder (LayerNorm–Multi-Head Self-Attention–MLP with residual connections) processes Z. For head $h = 1, \dots, H$,

$$Q_h = ZW_Q^{(h)}, \quad K_h = ZW_K^{(h)}, \quad V_h = ZW_V^{(h)}$$

$$\text{head}_\text{h} = \text{softmax}\left(\frac{Q_{hK_h^\top}}{\sqrt{d_h}}\right)V_h, \quad \text{MHA}(Z) = [\text{head}_1 \,\|\, \cdots \,\|\, \text{head}_\text{H}]W_O,$$

and with a position-wise MLP

$\text{FFN}(\boldsymbol{Z})=\sigma(\boldsymbol{Z}\text{W}_1+\text{b}_1)\text{W}_2+\text{b}_2$

Using pre-norm residuals, the encoder update is

$$Z' = Z + \text{MHA}\big(\text{LN}(Z)\big), \qquad Z'' = Z' + \text{FFN}\big(\text{LN}(Z')\big).$$

Finally, tokens are folded back to the spatial grid and projected with 1×11\times11×1 conv to match the channel dimension:

$F_{\text{glob}} = \text{Conv}_{1\times1}\big(\mathcal{P}^{-1}(Z'')\big) \in R^{C\times H\times W}$.

3. **Local fusion (conv):** global features are fused with a 3×33\times33×3 conv and (optionally) summed with the block input:

$$F_{\text{out}} = \phi\left(\text{Conv}_{3\times3}(F_{\text{glob}})\right).$$

Stacking such blocks yields hierarchical features with local texture fidelity (from convolutions) and global context (from attention), which is well-suited to full-slice inputs where distant parenchymal and vascular cues inform malignancy assessment (3.5).

Compatibility with pseudo-3D full slices: The three channels encode $\{z-1, z, z+1\}$, enabling early convolutions to mix through-plane cues with in-plane texture (central-difference intuition in 3.5). The self-attention over unfolded patches then facilitates long-range in-plane interactions (e.g., vessel–nodule relationships across lobar distances) at low parameter cost—an advantage over purely convolutional stacks and, importantly, far lighter than volumetric 3D CNNs [13,16]. Because inputs are full slices (no cropping), the effective receptive field of the CNN stages and the global receptive field of attention jointly cover nodule and milieu [15,16].

Classification head and output: Let $F^* \in R^{C^*\times H^*\times W^*}$ be the final feature map. A global average pooling produces $g \in R^{C^*} g$ with $g_c = \frac{1}{H^*W^*}\sum_{u,v} \text{b}F^*_{cuv}$. The binary logit and probability are

$$z = \mathbf{w}^\top \mathbf{g} + b, \qquad p = \sigma(z),$$

trained with BCEWithLogitsLoss using per-fold $pos\_weight = N_{\text{neg}}/N_{\text{pos}}$ (3.4). Optimization uses AdamW, scheduled learning rate, mixed precision, early stopping on validation AUC, and WeightedRandomSampler to approximate batch-level balance; augmentations are conservative and slice-coherent (3.4, 3.6, 3.8).

Grad-CAM target and computation: For explainability (3.9.1), Grad-CAM is computed on the final convolutional feature map (post fusion) to ensure spatial semantics compatible with CAM. With gradients $\partial z/\ \partial F^*_{cuv}$,

$$\alpha_c = \frac{1}{H^* W^*} \sum_{u,v} \frac{\partial z}{\partial \mathbf{F}^*_{cuv}}, \qquad M(u,v) = \mathrm{ReLU}\Big(\sum_c \alpha_c \, \mathbf{F}^*_{cuv} \Big),$$

which is normalized, upsampled to R×R, and overlaid on the axial slice [19]. These per-model CAMs contribute to the ensemble-level median aggregation (3.9.2).

MobileViT-XXS contributes global context modeling at low capacity, complementing EfficientNetV2-S's accuracy–latency trade-off and DenseNet-201's dense feature reuse. Under attention-based stacking (3.7.5), the ensemble can assign higher model-wise attention to MobileViT-XXS when long-range spatial relations are salient, while relying on alternative backbones for cases dominated by local textural cues—consistent with ensemble theory [7], gains from medical-imaging ensembles [26], and multi-attention stacking evidence in lung CT [31].

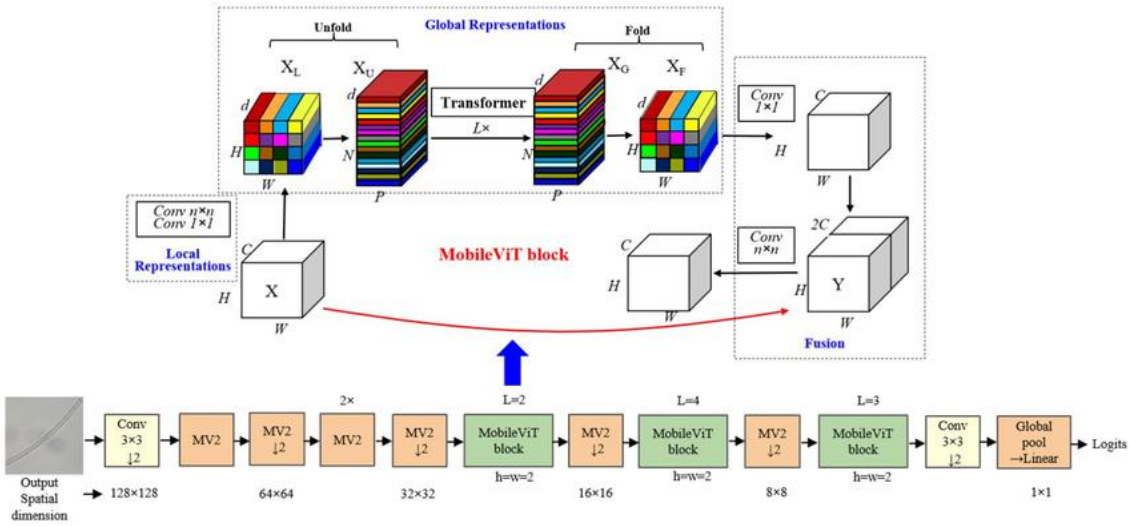


Figure 3.5 MobileViT Architecture

3.6.5 Multi-Attention Stacked Ensemble (MASE)

The ensemble layer integrates the three heterogeneous backbones (EfficientNetV2-S, DenseNet-201, MobileViT-XXS) by learning input-conditioned weights over their logits together with class-specific reweighting, following the multi-attention stacking paradigm reported for lung CT while adapting it to full-slice pseudo-3D inputs and our training

protocol [7,26,31]. The stacker is deliberately lightweight to minimize overfitting and to preserve a clean separation between representation learning (in the bases) and decision fusion.

Inputs and notation: For an input $X \in R^{3\times R\times R}$ (3.3–3.5), the three trained backbones produce logits

$$z = [z_1, z_2, z_3]^\top \in R^3,$$

with probabilities $p_k = \sigma(z_k)$ per model $k \in \{1,2,3\}$ The stacker maps z to a final logit $z_* \in R$, yielding $p_* = \sigma(z_*)$.

Model-wise attention: A shallow bottleneck computes a latent summary of base evidence,

$$\mathbf{h} = \rho(W_h\,\mathbf{z} + \mathbf{b}_h) \in \mathbb{R}^d,$$

where $W_h \in R^{d\times 3}, b_h \in R^d$, and ρ is a pointwise nonlinearity (ReLU). Model-wise attention allocates an input-conditioned weight to each backbone:

$$\boldsymbol{\alpha} = \mathrm{softmax}(W_a\,\mathbf{h} + \mathbf{b}_a) \in \mathbb{R}^3, \qquad \sum_{k=1}^{3} \alpha_k = 1,\ \alpha_k \geq 0,$$

with $W_a \in R^{3\times d}$, $b_a \in R^3$. The attended logits are

$$\tilde{\mathbf{z}} = \boldsymbol{\alpha} \odot \mathbf{z} \in \mathbb{R}^3.$$

Class-wise attention and meta-learner: For the binary task, a linear layer assigns class-specific weights to the attended logits:

$$\mathbf{o} = W_c\,\tilde{\mathbf{z}} + \mathbf{b}_c \in \mathbb{R}^2, \qquad W_c = \begin{bmatrix} \mathbf{w}_{\text{ben}}^\top \\ \mathbf{w}_{\text{mal}}^\top \end{bmatrix},\ \mathbf{b}_c = \begin{bmatrix} b_{\text{ben}} \\ b_{\text{mal}} \end{bmatrix}.$$

The final ensemble logit is the log-odds difference

$$z_* = o_{\text{mal}} - o_{\text{ben}} = \left(\mathbf{w}_{\text{mal}} - \mathbf{w}_{\text{ben}}\right)^\top \tilde{\mathbf{z}} + \left(b_{\text{mal}} - b_{\text{ben}}\right),$$

which is compatible with BCEWithLogitsLoss. Equivalently, MASE can be viewed as a mixture-of-experts with a gating network (α) and a class-aware linear expert combiner on the gated evidence.

Training protocol: Stacking is performed within each cross-validation fold using only out-of-fold base logits to preserve fold integrity:

1. Train each backbone on the training partition (patient-level 5-fold; 3.8.3) with class-weighted BCE and balanced mini-batches (3.4).

2. Collect base logits on held-out samples from the same training partition (e.g., via internal K-fold or the backbone's validation split) to form the stacker training set $\{(z_i, y_i)\}$.

3. Fit the stacker on $\{(z_i, y_i)\}$ using the same class-weighted BCE:

$$\mathcal{L}_{\text{stack}} = \frac{1}{M} \sum_{i=1}^{M} \left(w_+ \, y_i \log(1 + e^{-z_{*,i}}) + w_- \, (1 - y_i) \log(1 + e^{z_{*,i}}) \right),$$

with $w_+ = N_{\text{neg}}/N_{\text{pos}}$, $w_- = 1$ computed on the stacker's training subset (3.4). Optimization uses AdamW with the same scheduler family and early stopping on validation AUC (3.8).
4) Evaluate the fitted stacker on the fold's validation partition, where z\mathbf{z}z are the frozen base logits and no refitting occurs.

This procedure ensures that the stacker never consumes the same sample's in-fold base predictions as training targets, preventing optimistic bias while preserving the pairing of folds across bases and ensemble [7,26].

Decision and metrics: At inference, the final probability is $p_* = \sigma(z_*)$. Unless otherwise specified, a threshold of 0.5 is used for class decisions; model selection emphasizes **AUC**, which is insensitive to thresholding and robust under class imbalance.

Interpretation of attention coefficients: Although attention weights are not used to reweight Grad-CAM maps (3.9.2), α offers a model-wise attribution signal indicating which backbone dominated the decision for a given case. Aggregated across a fold, $E[\alpha]$ can reveal dataset slices where particular inductive biases (e.g., MobileViT-XXS's long-range context, DenseNet-201's feature reuse) are preferentially exploited by the ensemble.

Computational footprint: The stacker adds only a small MLP and two linear layers over a 3-dimensional input; its parameter count is negligible compared with the bases, and its per-sample cost is $\mathcal{O}$. Consequently, MASE improves robustness with minimal latency overhead, preserving deployability.

Relation to prior work: The design mirrors the two-stage attention mechanism proposed by Saha & Prakash for lung nodule classification, which demonstrated state-of-the-art accuracy and AUC when fusing diverse CNN backbones [31]. The present formulation retains the core idea (model-wise and class-wise attention) but adapts it to full-slice pseudo-3D inputs and our patient-level CV regime, aligning with ensemble theory [7] and empirical gains of ensembles in medical imaging [26].

3.7 Training Strategy

Training proceeds in two stages under patient-level, stratified 5-fold cross-validation: (i) fine-tuning of three complementary 2D backbones on full-slice pseudo-3D inputs, and (ii) fitting of a lightweight multi-attention stacker on out-of-fold base logits (3.5, 3.7.5). All models share the same folds to enable paired comparisons and a clean stacking protocol [7,26]. Class imbalance is addressed uniformly with per-fold loss reweighting and WeightedRandomSampler mini-batches (3.4). Data augmentation is training-time only and slice-coherent (3.6); test-time augmentation is not used in the final reporting.

For each backbone $f_k(\cdot\,;\,\theta_k)$, the best checkpoint per fold is selected by maximum validation AUC (threshold-agnostic, robust under skewed prevalence). Inference uses the sigmoid of the logit and a default decision threshold of 0.5 unless otherwise noted; we report AUC, accuracy, sensitivity, specificity, F1-score, and precision per fold and summarize by mean ± dispersion. The stacker consumes frozen base logits and is trained with the same loss reweighting on out-of-fold data to preserve fold integrity (3.7.5). All runs are deterministic with fixed seeds and version-locked preprocessing; folds are defined at the patient level to prevent any slice/nodule overlap between partitions [2,15].

3.7.1 Loss Function

Both base learners and the stacker optimize a class-weighted logistic loss (BCE-with-logits) to counter class imbalance per fold (3.4). For a training partition $D_{train} = \{(X_i, y_i)\}_{i=1}^{N}$, with $y_i \in \{0,1\} yi \in 0,1$ model logit $z_i \in R$, and $N_{\text{pos}}, N_{\text{neg}}$ the counts of positive/negative labels in that partition, the empirical risk is

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \Big(w_+ \, y_i \, \log(1 + e^{-z_i}) + w_- \, (1 - y_i) \, \log(1 + e^{z_i}) \Big),$$

with weights chosen as

$$w_+ = \frac{N_{\text{neg}}}{N_{\text{pos}}}, \qquad w_- = 1,$$

so that, in expectation, positive and negative samples contribute comparably to the gradient. In implementation terms (PyTorch), this corresponds to BCEWithLogitsLoss(pos_weight = N_neg/N_pos) computed anew in each fold after label construction (3.3.1). Loss reweighting is paired with WeightedRandomSampler to approximate batch-level balance without duplicating samples (3.4). No label smoothing and no MixUp are used in the final configuration.

For the stacker, the same objective is applied to the ensemble logit $z_{*,i}$ obtained from multi-attention fusion of base logits z_* defined in 3.7.5), with w_+ , w_- recomputed on the stacker's training subset:

$$\mathcal{L}_{\text{stack}} = \frac{1}{M} \sum_{i=1}^{M} \Big(w_+ \, y_i \, \log(1 + e^{-z_{*,i}}) + w_- \, (1 - y_i) \, \log(1 + e^{z_{*,i}}) \Big).$$

Because the decision surface is expressed in log-odds, **AUC** remains the primary model-selection criterion and is unaffected by the choice of operating threshold during training.

3.7.2 Optimization

All networks are trained end-to-end with AdamW (decoupled weight decay), mixed-precision (AMP), and early stopping on validation AUC. We adopt a warm-up + cosine decay learning-rate schedule across TTT optimization steps, with a linear warm-up over the first T_w steps followed by cosine annealing to η_{min}

$$\eta_t = \begin{cases} \eta_{\max} \dfrac{t}{T_w}, & 0 \le t \le T_w, \\ \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left[1 + \cos\left(\pi \dfrac{t - T_w}{T - T_w}\right)\right], & T_w < t \le T. \end{cases}$$

AdamW parameter updates for parameter block θ with gradient $g_t = \nabla_\theta \mathcal{L}_t$, momentum β1 , second-moment β2 and weight-decay λ are

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\, g_t, \qquad v_t = \beta_2 v_{t-1} + (1 - \beta_2)\, g_t^{\odot 2},$$
$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \qquad \hat{v}_t = \frac{v_t}{1 - \beta_2^t},$$
$$\theta_{t+1} = \theta_t - \eta_t \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} + \lambda\, \theta_t \right).$$

To stabilize training under occasional hard cases, we employ global-norm gradient clipping at threshold c:

$$\tilde{g} = \begin{cases} g, & \|g\|_2 \le c, \\ \dfrac{c}{\|g\|_2}\, g, & \text{otherwise.} \end{cases}$$

Early stopping halts when validation AUC fails to improve by at least δ\deltaδ for PPP consecutive evaluations; the checkpoint with maximum AUC is retained for inference. Batch construction uses the WeightedRandomSampler with class-neutral augmentation (3.6). No test-time augmentation (TTA) is applied at inference, consistent with internal experiments indicating no consistent benefit for the backbones used.

3.7.3 Data Splits

Evaluation follows patient-level, stratified 5-fold cross-validation to prevent optimistic bias from slice/nodule overlap and to reflect clinical deployment, where a patient is encountered once [2,15]. Let $\mathcal{P}$ be the set of unique patients; stratification assigns a fold index $\phi(p) \in \{1, \ldots, 5\}$ to each $p \in \mathcal{P}$ so that the patient-level prevalence of $y = 1$ is approximately equal across folds. For fold k,

$$\mathcal{D}_{\text{val}}^{(k)} = \{(X_i, y_i) : \text{patient}(i) \in \phi^{-1}(k)\}, \qquad \mathcal{D}_{\text{train}}^{(k)} = \bigcup_{j \neq k} \{(X_i, y_i) : \text{patient}(i) \in \phi^{-1}(j)\}.$$

All three backbones are trained on $D_{train}^{(k)}$ and evaluated on $D_{val}^{(k)}$. Per-fold class weights (w_+, w_-) (and thus `pos_weight`) are computed from $D_{train}^{(k)}$ (3.4). For stacking, out-of-

fold base logits $\{(z_i, y_i)\}$ are generated from $D_{train}^{(k)}$ (e.g., via an internal split), the stacker is fitted on those pairs, and evaluation uses frozen base logits on $D_{val}^{(k)}$ (3.7.5). This ensures the ensemble never trains on predictions of the same samples it is evaluated on, preserving fold integrity [7,26,31].

This protocol, combined with full-slice pseudo-3D inputs [15,16] and attention-based ensembling, yields performance estimates that are robust, threshold-agnostic at selection time (AUC), and reproducible across folds, while avoiding confounds associated with patient-level leakage or post-hoc aggregation tricks.

3.8 Explainability

Explainability is incorporated as a first-class objective to support clinical review and downstream auditing. The approach combines saliency attributions using Grad-CAM [19] with an ensemble-level aggregation that reflects the deployed decision boundary (the stacked ensemble), thereby aligning visual evidence with the final prediction. This design follows recommendations from medical XAI literature emphasizing transparency, faithfulness to the underlying model, and clinician-oriented presentation [10,12,14,20]. Section **3.9.1** details per-model Grad-CAM computation; Section **3.9.2** describes aggregation across backbones (and optionally across folds) to obtain a single consensus heatmap for each case.

3.8.1 Grad-CAM for Per-Model Attributions

Grad-CAM is computed for each backbone (EfficientNetV2-S, DenseNet-201, MobileViT-XXS) at a late convolutional feature map to balance semantic depth and spatial resolution [19]. For binary classification, each model produces a single malignancy logit $z \in R$ (pre-sigmoid). Let $F \in R^{C\times H\times W}$ denote the selected feature tensor (channels C, spatial size $H \times W$). The channel weights are the global-average of the gradients of the logit with respect to F:

$$\alpha_c = \frac{1}{H\,W} \sum_{u=1}^{H} \sum_{v=1}^{W} \frac{\partial z}{\partial \mathbf{F}_{cuv}} \quad (c = 1, \ldots, C).$$

The class activation map (CAM) in feature space is

$$M(u,v) \;=\; \mathrm{ReLU}\Big(\sum_{c=1}^{C} \alpha_c \,\mathbf{F}_{cuv}\Big), \qquad M \in \mathbb{R}_{\geq 0}^{H\times W}.$$

To compare maps across models and overlay them on inputs, MMM is min–max normalized and resampled to the common input size R×R used in training (3.3–3.5):

$$\widehat{M} \;=\; \frac{M - \min(M)}{\max(M) - \min(M) + \varepsilon}, \qquad \widetilde{M} \;=\; \mathcal{U}_R(\widehat{M}) \in \mathbb{R}^{R\times R},$$

where $\mathcal{U}_{\mathcal{R}}$ is bicubic upsampling and $\varepsilon > 0$ avoids division by zero. Because inputs are pseudo-3D full slices ($channels = \{z-1, z, z+1\}$), overlays are rendered on the central slice I(z)I(z)I(z) to maintain the standard axial view; the same pipeline can optionally render per-channel overlays.

The resulting per-model map $\widetilde{M}$ is stored with identifiers to enable case-level auditing. These maps serve two downstream purposes: (i) qualitative inspection of whether each backbone attends to the nodule and its immediate context (as opposed to irrelevant regions), and (ii) inputs to the ensemble-level aggregation described next.

3.8.2 Ensemble-Level Grad-CAM Aggregation

Since the deployed decision is produced by the stacked ensemble (MASE; 3.7.5), a single, consensus attention map is required for clinical review. Let $\widetilde{M}^{(k)} \in R^{R\times R}$ denote the normalized, upsampled Grad-CAM from backbone $k \in \{1,2,3\}$ computed w.r.t. its own malignancy logit. The ensemble-level CAM for a case is defined as the pixel-wise median across backbones:

$$M^{\mathrm{ens}}(u,v) \;=\; \mathrm{median}\Big(\,\widetilde{M}^{(1)}(u,v),\; \widetilde{M}^{(2)}(u,v),\; \widetilde{M}^{(3)}(u,v)\,\Big), \quad (u,v) \in \{1,\dots,R\}^2.$$

Median aggregation is chosen for robustness to outliers and to emphasize consensus saliency among heterogeneous learners. When producing fold-aggregated visuals (e.g., for the same case evaluated across cross-validation folds), the definition extends naturally:

$$M^{\mathrm{ens,\,CV}}(u,v) \;=\; \mathrm{median}\left\{\, \widetilde{M}^{(k,f)}(u,v) \;\middle|\; k \in \{1,2,3\},\; f \in \{1,\dots,5\} \right\},$$

where $\widetilde{M}^{(k,f)}$ is the per-model Grad-CAM from backbone k in fold f. Finally, the consensus map is normalized for display,

$$\overline{M}^{\text{ens}} = \frac{M^{\text{ens}} - \min(M^{\text{ens}})}{\max(M^{\text{ens}}) - \min(M^{\text{ens}}) + \varepsilon},$$

and overlaid on the central axial slice I(z)I(z)I(z) with a standard colormap. Importantly, no attention weighting from the MASE gating is applied to CAMs; aggregation remains purely image-evidence–driven, avoiding confounding the attribution with the meta-learner's scalar weights. This separation keeps the visualization faithful to where the base models found evidence, while the MASE coefficients quantify which model contributed more to the decision (see interpretive note in 3.7.5).

This two-stage explainability—per-model Grad-CAM followed by ensemble-level median aggregation—adheres to medical XAI guidance: it is faithful to the trained models [19], aligned to the deployed predictor (the ensemble), and clinician-oriented in presentation [10,12,14,20].

3.8.3 LLM-Based Clinical Narrative Generation

To complement saliency overlays with clinician-readable justification, this work generates structured narratives from a compact, deterministic summary of the ensemble prediction and its ensemble-level Grad-CAM (3.7.5, 3.9.2). Generation is post hoc and does not influence the predictive model. Consistent with guidance on medical XAI, the objective is to provide concise, auditable text that is strictly grounded in model-derived evidence [10,12,14,20,29,30].

Inputs (structured evidence): For a case with central axial slice $I(z)$, let $p_* = \sigma(z_*)$ be the ensemble probability (3.7.5), and let $\overline{M}^{\text{ens}} \in [0,1]^{R\times R}$ denote the normalized ensemble CAM (3.9.2). We compute a minimal set of numeric descriptors that summarize location, focality, and alignment with the annotated nodule center (if available), and pass these—together with the binary decision $\hat{y} = \mathbb{1}[p_* \geq 0.5]$ — to the language model.

1. **CAM centroid and dispersion**

$$\mu = \frac{\sum_{u,v} \overline{M}^{\text{ens}}(u,v)\,[u,\,v]^{\top}}{\sum_{u,v} \overline{M}^{\text{ens}}(u,v)}, \qquad \Sigma = \frac{\sum_{u,v} \overline{M}^{\text{ens}}(u,v)\big([u,\,v]^{\top}-\mu\big)\big([u,\,v]-\mu^{\top}\big)}{\sum_{u,v} \overline{M}^{\text{ens}}(u,v)}.$$

2. **Focality (entropy of normalized saliency)**

$$\hat{M}_{uv} = \frac{\overline{M}^{\text{ens}}(u,v)}{\sum_{a,b} \overline{M}^{\text{ens}}(a,b)}, \qquad H(\overline{M}^{\text{ens}}) = -\sum_{u,v} \hat{M}_{uv}\,\log\big(\hat{M}_{uv}+\varepsilon\big),$$

where lower H indicates a more focal hotspot.

3. **Proximity to annotated nodule center** (if center cnc_ncn is available)

$$d = \| \mu - c_n \|_2\,, \qquad \kappa(r) = \sum_{\|[u,v]^{\top}-\mu\|_2 \le r} \widehat{M}_{uv}\,.$$

4. **Risk bin for verbalization** (fixed thresholds $\tau_1 < \tau_2$)

$$\text{risk}(p_*) = \begin{cases} \text{low}, & p_* < \tau_1, \\ \text{intermediate}, & \tau_1 \le p_* < \tau_2, \\ \text{high}, & p_* \ge \tau_2. \end{cases}$$

These yield a structured vector $\Phi = \big(p_*, \hat{y}, \mu, \Sigma, H, d, \kappa(r)\big)$ that fully determines the narrative content.

Model choice and deployment: Narratives are produced by DeepSeek-R1 (reasoning-optimized LLM) [36]. The model is used post hoc for template-constrained verbalization of Φ; it does not receive pixel data, clinical free text, or any information beyond the deterministic descriptors above. This choice reflects (i) reasoning robustness for template-guided summarization, (ii) reproducibility via fixed decoding parameters, and (iii) data governance, since only non-PHI numeric tuples are processed.

Prompting and deterministic decoding: Let $\mathcal{G}_\theta$ denote DeepSeek-R1 with parameters θ and $\mathcal{T}$ a fixed prompt template (slots for risk bin, proximity ddd, focality H, and qualitative location derived from μ). Generation is

$$\text{Report} = \mathcal{G}_\theta(\Phi; \mathcal{T}, \text{decode} = \text{greedy}),$$

implemented with temperature =0=, top-p =1, beam size =1 to ensure exact reproducibility. The template enforces causability [14]: each sentence corresponds to a field in Φ (e.g., *"high-probability classification; saliency focal with centroid near annotated nodule"* when $p*$ is high, H low, d small). As an internal quality check, the produced text is parsed against a JSON schema derived from Φ; non-conforming outputs are re-decoded with the same constraints.

(Optional) calibrated probability for reporting: If temperature scaling is introduced for probability calibration, with parameter T>0,

$$p_*^{(\text{cal})} = \sigma\left(\frac{z_*}{T}\right),$$

then $p_*^{(\text{cal})}$ replaces $p*$ in Φ and in the risk mapping above. In the present results, model selection is AUC-based and no calibration is applied.

Scope and limitations: The LLM provides verbalization of evidence; it is not used for prediction and does not modify the classifier or its thresholds. By design, narratives cannot introduce ungrounded clinical claims, addressing common adoption concerns for black-box systems in healthcare [10,12,14,20,29,30]. The output, together with the ensemble CAM, yields a paired visual-textual explanation aligned with the deployed decision boundary (3.7.5).

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Evaluation Metrics

This study evaluates binary nodule classification using threshold-agnostic discrimination metrics (ROC–AUC, PR–AUC) and operating-point metrics (Accuracy, Sensitivity/Recall, Specificity, Precision), together with two derived summaries (F1, Balanced Accuracy). ROC–AUC (AUC) is computed only on uncalibrated scores because temperature scaling is a strictly monotonic transform that does not alter ranking; PR curves (and AUPRC) are presented for the calibrated ensemble to reflect the probability scale used in downstream reporting, noting that monotone calibration preserves ranking as well [37,38].

Notation: Let $y_i \in \{0,1\}$ be the ground truth (1 = malignant) and $s_i \in R$ the model score (logit) for case i. For a probability threshold τ, defined $\hat{y}_i = 1[\sigma(s_i) \geq \tau]$ with $\sigma(t) = 1/(1 + e^{-t})$. The confusion-matrix counts are

$$\boldsymbol{TP} = \sum_{\boldsymbol{i}} \mathbf{1}[\,\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \mathbf{1} \wedge \boldsymbol{y}_{\boldsymbol{i}} = \mathbf{1}\,], \quad \boldsymbol{FP} = \sum_{\boldsymbol{i}} \mathbf{1}[\,\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \mathbf{1} \wedge \boldsymbol{y}_{\boldsymbol{i}} = \mathbf{0}\,],$$

$$\boldsymbol{TN} = \sum_{\boldsymbol{i}} \mathbf{1}[\,\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \mathbf{0} \wedge \boldsymbol{y}_{\boldsymbol{i}} = \mathbf{0}\,], \quad \boldsymbol{FN} = \sum_{\boldsymbol{i}} \mathbf{1}[\,\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \mathbf{0} \wedge \boldsymbol{y}_{\boldsymbol{i}} = \mathbf{1}\,].$$

4.1.1 Discrimination (threshold-agnostic)

ROC–AUC (AUC): The ROC curve plots $\mathrm{TPR}(\tau)$ versus $\mathrm{FPR}(\tau)$ as τ varies, where

$$\mathrm{TPR}(\tau) = \frac{TP}{TP+FN}, \qquad \mathrm{FPR}(\tau) = \frac{FP}{FP+TN} \quad .$$

The AUC equals the probability that a randomly drawn positive receives a higher score than a randomly drawn negative (Wilcoxon–Mann–Whitney view) [39]:

$$\text{AUC} = \Pr[s(x^+) > s(x^-)] + \frac{1}{2}\Pr[s(x^+) = s(x^-)] = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(u)\right) du\,.$$

Because temperature scaling $s \mapsto s/T\left(or\ p \mapsto \sigma(s/T)\right)$ is strictly increasing, AUC is theoretically unchanged by calibration [37,38]. Accordingly, AUC is reported **only** for uncalibrated scores.

PR–AUC (AUPRC): The Precision–Recall curve plots Precision(τ) against Recall(τ) as τ varies, with area

$$\text{AUPRC} = \int_0^1 \text{Precision}\left(\text{Recall}^{-1}(r)\right) dr,$$

computed empirically by summing trapezoids over recall. PR–AUC is particularly informative under class imbalance [40]. Since calibration is monotone, AUPRC is invariant to temperature scaling; we present PR curves for the calibrated ensemble to align with the probability scale used elsewhere, while acknowledging that ranking is unchanged [37,38].

4.1.2 Operating-Point Metrics (thresholded)

Operating-point metrics summarize performance at a fixed decision threshold τ applied to the model's probabilities $\widehat{\boldsymbol{y_i}} = \mathbf{1}[\boldsymbol{\sigma}(\boldsymbol{s_i}) \geq \boldsymbol{\tau}]$ with $\boldsymbol{\sigma}(\boldsymbol{t}) = \mathbf{1}/(\mathbf{1} + \boldsymbol{e^{-t}})$. Unless otherwise noted, τ=0.5. Let TP, FP, TN, FN denote the usual confusion-matrix counts.

Accuracy: Proportion of correctly classified cases.

$$\text{Accuracy} = \frac{(\text{TP+TN+FP+FN})}{\text{TP+TN}}$$

Accuracy provides a single aggregate rate but can be misleading under class imbalance, because majority-class correctness dominates the numerator.

Sensitivity (Recall / True Positive Rate): The ratio of correctly predicted positive observations to all actual positives. It is the ability to identify malignant nodules.

$$\text{Sensitivity(Recall)} = \frac{\text{TP}}{\text{TP+FN}}$$

High sensitivity implies few missed malignancies (low false-negative rate). In screening contexts this metric is clinically critical, but it does not penalize false positives.

Specificity (True Negative Rate): The Fraction of truly benign cases correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

High specificity implies fewer benign cases flagged as malignant (low false-positive rate), reducing unnecessary follow-up. Sensitivity and specificity trade off as τ\tauτ varies (ROC) [39].

Precision: The ratio of correctly predicted positive observations to the total predicted positives. It shows the reliability of positive predictions.

$$\text{Precison} = \frac{TP}{TP+FP}$$

Precision quantifies how often a predicted malignant case is truly malignant. Unlike sensitivity/specificity, precision depends on **class prevalence**; under low prevalence, precision can be modest even for models with strong discrimination.

F1-Score: The harmonic means of Precision and Sensitivity(Recall). It balances both metrics in one score.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 is useful when positive-class detection is prioritized but false positives still matter; it down-weights models that excel at only one of Precision or Sensitivity(Recall).

In this study we report all operating-point metrics at τ=0.5 for comparability across backbones and the ensemble. Because post-hoc calibration applies a strictly monotone transform to the scores, it does not alter ranking-based discrimination (ROC/PR) but can shift operating-point metrics at a fixed τ by rescaling probabilities [37,38]. Accordingly, differences between MASE (uncalibrated) and MASE (calibrated) at τ=0.5 should be interpreted as changes in threshold behavior on the calibrated probability scale rather than changes in underlying discrimination.

4.2 Result Analysis

This section reports discrimination and operating-point performance for the three base learners and for the stacked ensemble (MASE), followed by the calibrated variant used for probability-scale alignment (temperature scaling). Consistent with theory, ROC–AUC is reported only for uncalibrated scores (calibration is monotonic and does not alter ranking) [37,38]. Precision–Recall (PR) curves are shown for the calibrated ensemble, noting the same ranking logic [37,38]. Interpretations are made in light of the evaluation definitions in 4.1 and the ensemble rationale in 3.7.1–3.7.5 [7,26,31].

4.2.1 Backbone baselines (per fold)

Across five patient-level folds, EfficientNetV2-S achieves the strongest base performance, followed by DenseNet-201 and MobileViT-XXS. Mean (±SD) over folds:

Table 4.1 Backbone model performance analysis(per fold)

Model	***Fold***	***Accuracy***	***AUC***	***F1***	***Precision***	***Sensitivity***	***Specificity***
efficientnetv2_s	1	0.954	0.982	0.914	0.892	0.938	0.959
	2	0.922	0.975	0.864	0.797	0.944	0.914
	3	0.943	0.976	0.897	0.862	0.934	0.947
	4	0.939	0.977	0.891	0.841	0.947	0.936
	5	0.944	0.973	0.898	0.857	0.944	0.943
densenet201	1	0.922	0.975	0.864	0.798	0.942	0.915
	2	0.946	0.971	0.898	0.894	0.903	0.962
	3	0.93	0.972	0.872	0.839	0.907	0.938
	4	0.93	0.973	0.873	0.835	0.915	0.935
	5	0.928	0.971	0.873	0.816	0.938	0.925
mobilevit_xxs	1	0.928	0.969	0.872	0.822	0.929	0.928
	2	0.92	0.966	0.853	0.828	0.88	0.935
	3	0.904	0.964	0.833	0.767	0.912	0.901
	4	0.925	0.963	0.86	0.84	0.882	0.94
	5	0.901	0.96	0.822	0.781	0.868	0.913

Table 4.2 Mean backbone model performance analysis with standard deviations

Model	Accuracy	AUC	F1	Precision	Sensitivity	Specificity
efficientnetv2_s	0.94	0.977	0.893	0.85	0.941	0.94
(±SD)	0.0116	0.00311	0.0181	0.0348	0.00537	0.0168
densenet201	0.931	0.972	0.876	0.836	0.921	0.935
(±SD)	0.00899	0.00174	0.013	0.0359	0.0179	0.0176
mobilevit_xxs	0.916	0.964	0.848	0.807	0.894	0.923
(±SD)	0.0123	0.00327	0.0203	0.0317	0.0253	0.0161

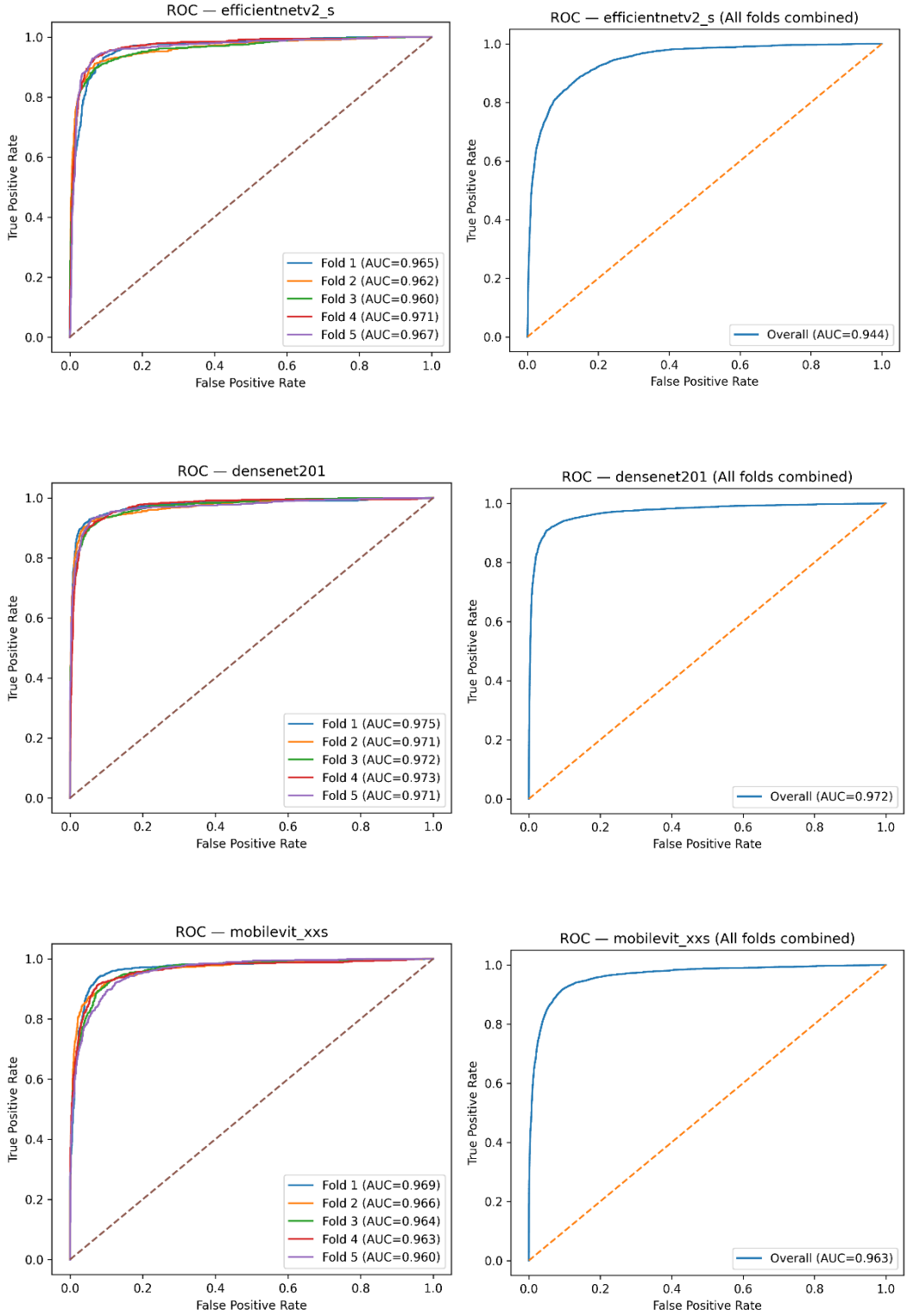
4.2.2 Multi Attention Stacked Ensemble

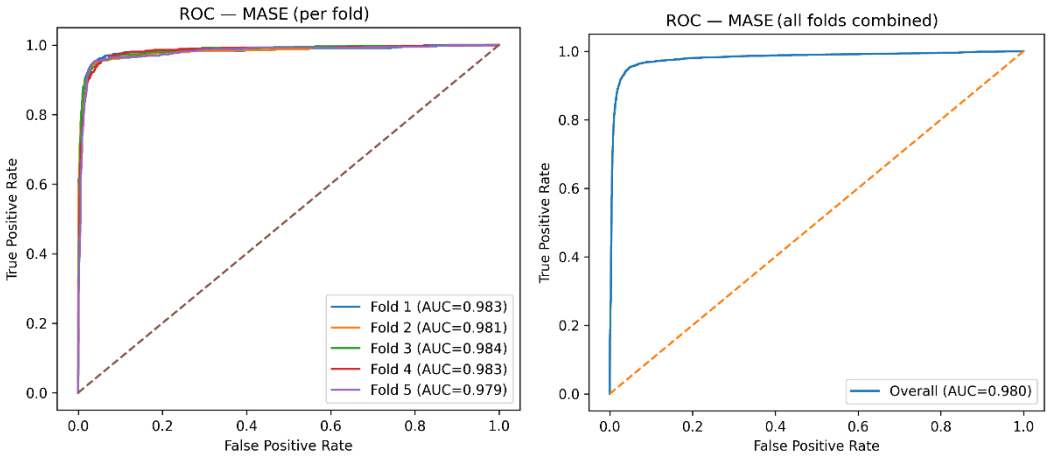
The attention-based stacking (MASE) improves discrimination and threshold performance relative to all individual backbones, consistent with ensemble theory and prior findings in lung CT [7,26,31]. Over five folds:

Table 4.3 Mean MASE model performance analysis with standard deviations

	accuracy	auc	f1	precision	sensitivity	specificity
mean	0.949	0.98	0.907	0.864	0.956	0.946
std	0.0048	0.0019	0.0076	0.0177	0.0058	0.0083

Figure 4.1 AUC ROC (per fold and overall folds combined)



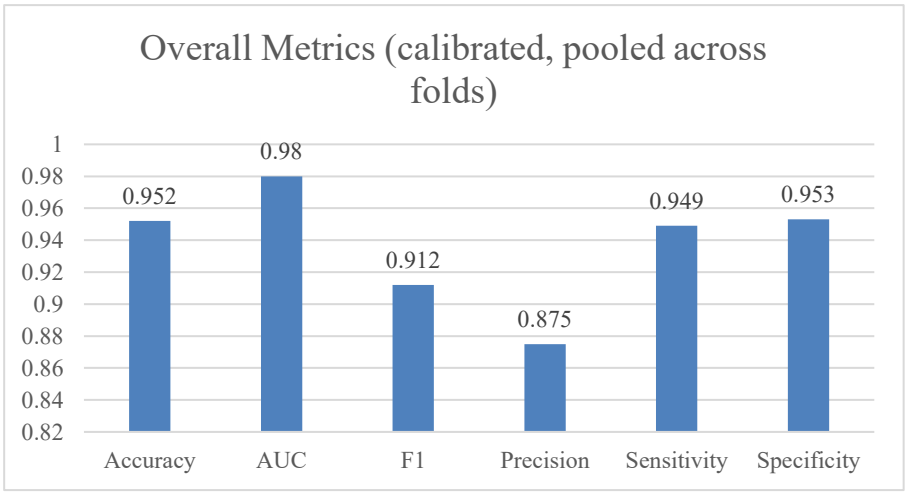


Relative to the strongest base (EfficientNetV2-S), MASE yields absolute gains of +0.0055 AUC, +0.0083 accuracy, +0.0145 sensitivity, +0.0060 specificity, +0.0140 precision, and +0.0145 F1.The joint increase in sensitivity and precision at the default threshold (τ=0.5) indicates that the stacker's model-wise and class-wise attention successfully exploits complementary evidence among backbones (3.7.5).

4.2.3 Probability calibration and PR analysis

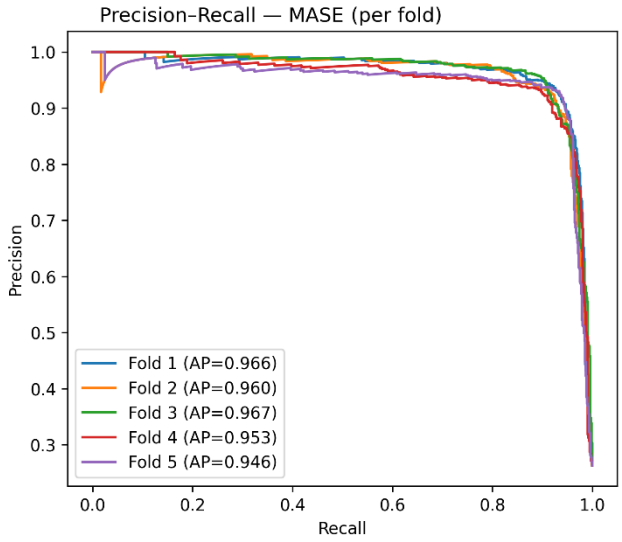
Post-hoc temperature scaling is applied **only** to MASE to align predicted probabilities with frequency semantics; as calibration is strictly monotone it does **not** alter ROC–AUC and thus we do not report AUC post-calibration [37,38]. Operating-point metrics at τ=0.5 change modestly because the probability scale is respecified:

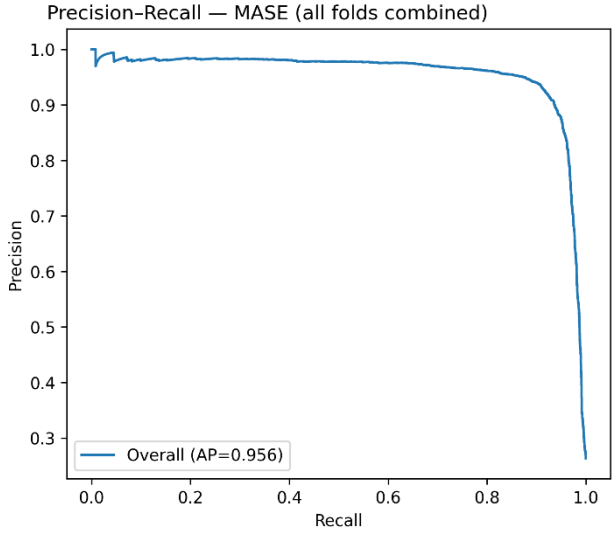
Figure 4.1 AUC ROC (per fold and overall folds combined)



Compared with uncalibrated MASE, calibration slightly **trades sensitivity for specificity** while **increasing accuracy precision and F1** at the fixed 0.5 threshold—an expected effect of rescaling probabilities without changing ranking (4.1; [37,38]). For depiction of ranking in the imbalanced regime, the PR curves for the calibrated ensemble is presented.

Figure 4.2 AUC PR (per fold and overall folds combined)





4.2.4 Operating-point interpretation

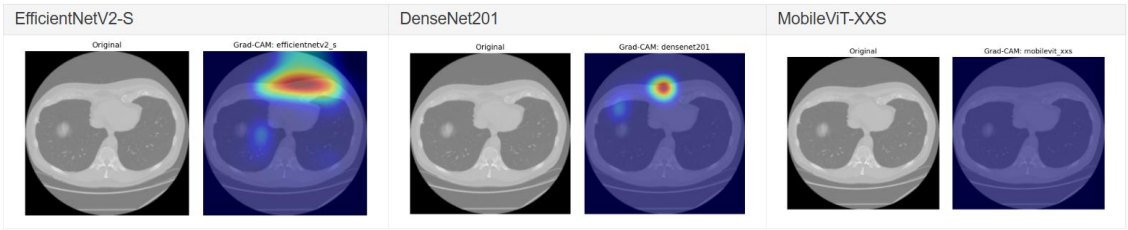
At the standard threshold (τ=0.5), MASE improves both Sensitivity and Precision relative to the strongest backbone, thereby increasing F1 while maintaining Specificity. After calibration, Precision and F1 increase further, with a modest reduction in Sensitivity and a compensatory rise in Specificity. This pattern is consistent with calibration's effect on the probability scale rather than on rank-based discrimination [37,38]. In clinical terms, the ensemble reduces missed malignancies (higher sensitivity) without a large penalty in false positives; calibrated probabilities offer more reliable positive predictions at the same nominal threshold.

4.2.5 Qualitative evidence

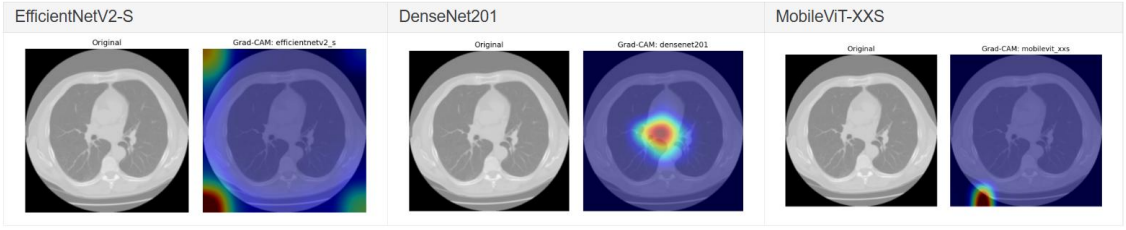
To contextualize the quantitative gains, we recommend presenting Grad-CAM and ensemble-level CAM overlays (3.9.1–3.9.2) for representative TP/TN/FP/FN cases, showing that attention concentrates on the nodule and relevant perinodular context [19]. This supports clinical plausibility and aligns with explainability guidance for decision support [10,12,14,20].

Figure 4.3 Base model Grad-CAM results

Sample 1



Sample 2



Sample 3

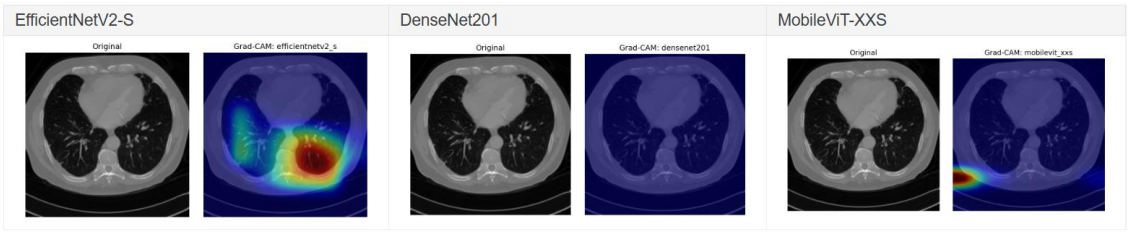
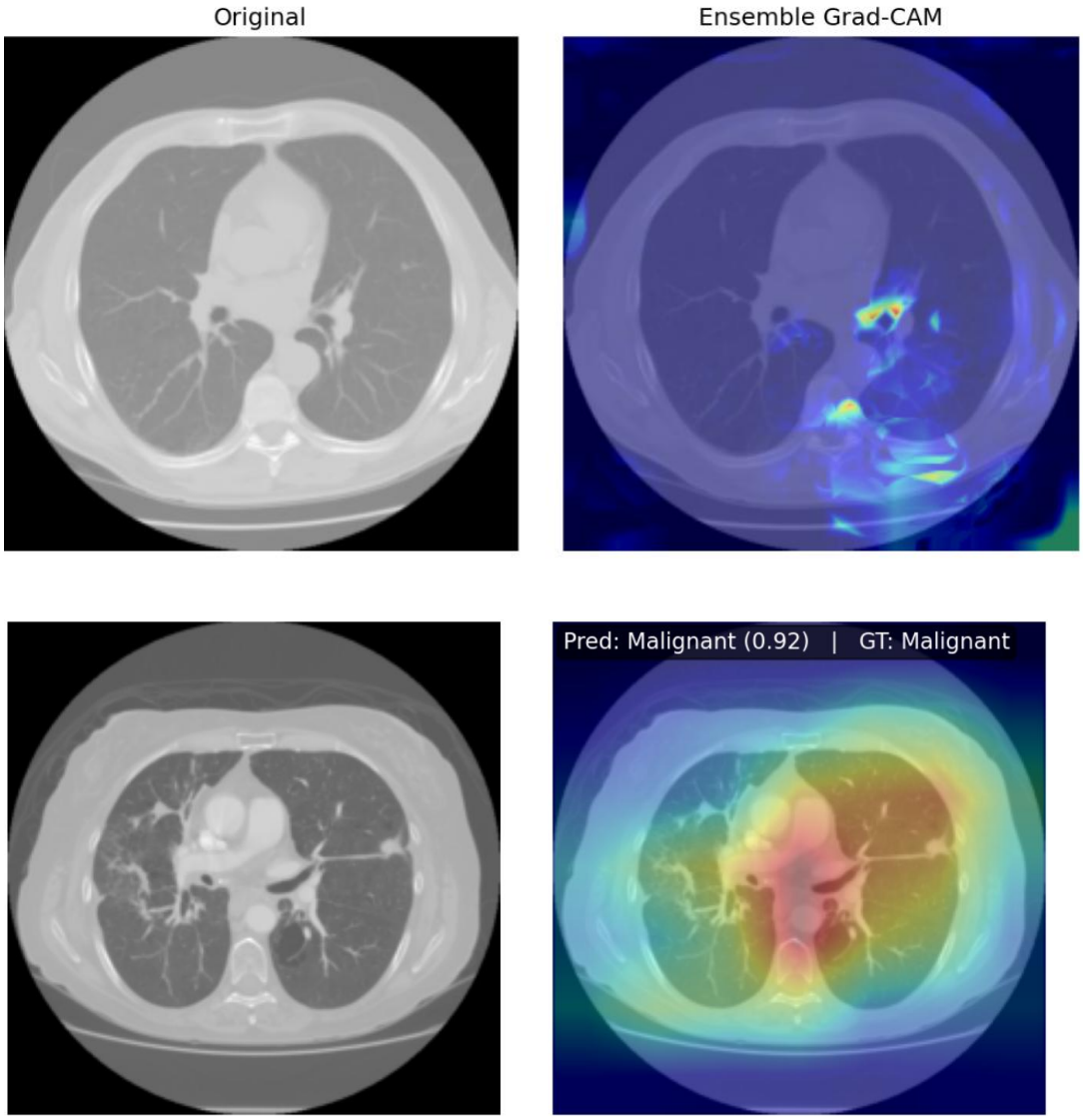
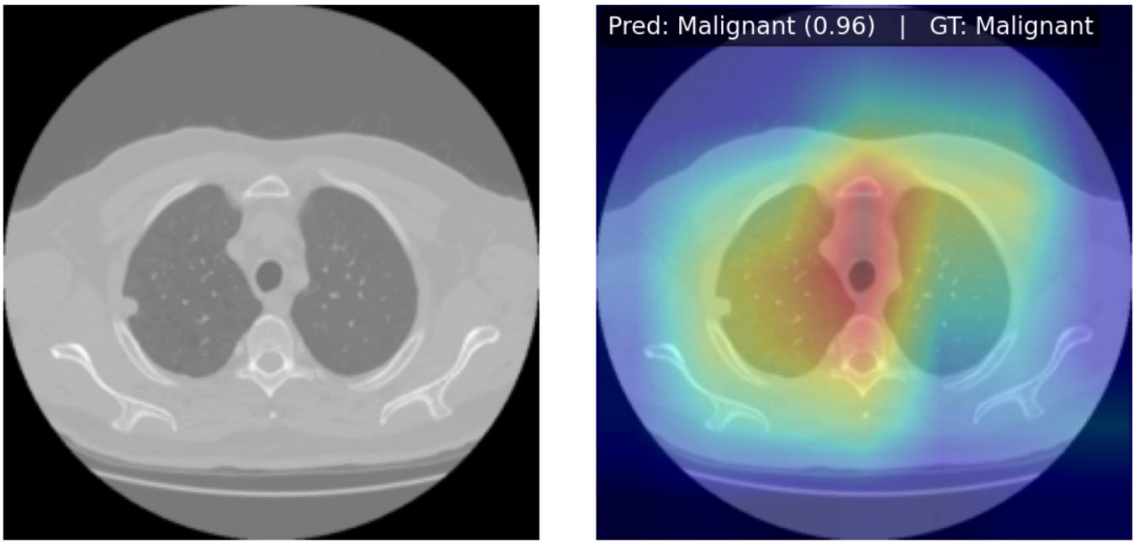


Figure 4.4 Ensemble level Grad-CAM results on various samples





4.2.6 LLM-Generated Clinical Narrative:

To complement quantitative metrics and CAM overlays, we generated a structured clinical narrative using DeepSeek-R1 (reasoning-optimized LLM) conditioned on a compact, deterministic summary of the ensemble output (3.7.5, 3.9.2–3.9.3). The LLM receives only numerics derived from the model—ensemble probability p_*, ensemble-level Grad-CAM descriptors (centroid/spread/location), and available meta-fields (e.g., diameter, margin, spiculation)—and produces a templated markdown report [10,12,14,20,29,30,36]. Decoding is low-temperature (temperature ≈ 0.2) to minimize variability while retaining readability, and the template enforces causability by requiring each claim to map to a supplied field (3.9.3).

Observed output (representative case): In the provided report, the ensemble assigned $p_* = 0.97$ and classified the nodule as *Malignant* ("model malignancy score 5.0/5.0"). The narrative's "Why the model reached this decision" section properly ties the high probability to ensemble agreement (consistent with our multi-attention stack; 3.7.5), and the "Support from Grad-CAM" section states "Peak activation is broad in the central field," which is coherent with our robust CAM descriptor: a 90th-percentile mask, largest connected-component (LCC) area thresholding for *focal/moderate/broad* spread, and center-of-mass/peak/radial-mass rules that bucket the location as *central/mid/peripheral* (3.9.2–3.9.3). The report also surfaces meta-fields parsed from the package (Diameter

15.0 mm; Margin 5.0; Spiculation 2.0) without inventing unseen clinical details—aligning with our no-hallucination constraint [10,12,14,20].

Faithfulness checks: The code enforces auditability at three levels (see 3.9.3): (i) **Deterministic feature set** Φ : the LLM input includes only $\{ p_*, \hat{y}$, CAM centroid/dispersion/focality/location, and optional size/margin/spiculation$\}$; no pixel data or free text are passed.

(ii) **Template-constrained prompting** : the output must follow a fixed sectioning (*Summary*, *Why*, *Support from Grad-CAM*, *Caveats*, *Reminder*) and cannot introduce variables not present in Φ.

(iii) **Sanitization/fallback** : if the LLM deviates, a rule-based fallback renders the same structure directly from Φ. These steps satisfy medical XAI recommendations for faithfulness and traceability [10,12,14,20,29,30].

Interpretation and limitations: The narrative correctly emphasizes that Grad-CAM does not encode margins/spiculation and may highlight contextual parenchyma, which matches known limitations of saliency methods [19]. The "Important Reminder" paragraph frames the output as decision support, not diagnosis, consistent with clinical safety guidance [10,12]. Two nuances merit note for readers:

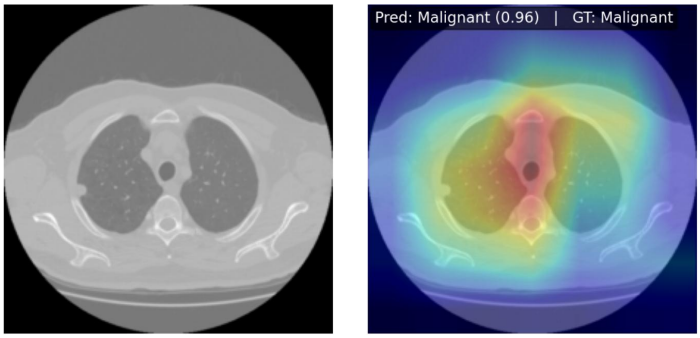
1. The probabilities driving the narrative can be uncalibrated or calibrated (when available). Because temperature scaling is monotonic, it does not alter ranking (ROC/PR), but it can sharpen textual risk wording (e.g., "high probability") at a fixed threshold [37,38].

2. The CAM descriptor reports location and spread, not morphological taxonomy; absence of explicit *spiculation/margin* reasoning in the text is by design (those appear only if provided as numeric meta-fields).

Placement in the results: We include a screenshot of the LLM report next to the corresponding ensemble-level CAM overlay and the central axial slice, allowing side-by-side inspection of (a) quantitative probability, (b) visual evidence, and (c) textual rationale. This triad supports reader appraisal of plausibility (attention near the nodule),

consistency (narrative matches CAM descriptors), and restraint (no unsupported clinical assertions) [10,12,14,19,20,29,30].

Figure 4.5 Final Output with Grad-CAM and LLM outcomes



Pred prob: 0.966, Grad-CAM: Peak activation is broad in the central field

Generated Report:

AI Assistant Report: Malignancy Probability Assessment for Lung Nodule

Model Prediction Summary

The ensemble deep learning model predicted a **high probability of malignancy (0.97)** for the lung nodule, resulting in a classification of **Malignant**. This is supported by a model malignancy score of **5.0/5.0**.

1. **Why the Model Reached This Decision**

- **Malignancy probability (0.97)** reflects consensus across the ensemble's sub-models, indicating patterns the model has learned to associate with malignancy.
- **Malignancy score (5.0/5.0)** indicates a high level of suspicion based on aggregated imaging/radiomic features learned during training.
- **Model attention (Grad-CAM)** highlights regions deemed informative for the decision, suggesting contributory imaging context even when explicit descriptors are limited.

2. **Support from Grad-CAM and Imaging Features**

Key Supporting Evidence

- **Grad-CAM Activation Pattern:** Peak activation is broad in the central field.
- Parenchymal activation can align with textural/structural patterns the model links with malignancy.

Contradictions and Uncertainties

- Grad-CAM does **not** encode margins/spiculation and may not localize a discrete nodule.
- Diffuse or context-driven activation can be non-specific and influenced by background parenchyma.

3. **Caveats, Limitations, and Missing Data**

- **Diameter:** 15.0 mm. **Margin:** 5.0. **Spiculation:** 2.0.
- Possible false positives; performance may vary with atypical presentations or limited context.
- Clinical history and prior imaging are not incorporated by the model.

4. **Important Reminder**

This model output is a diagnostic aid, not a definitive diagnosis. Integrate with clinical judgment, prior imaging, and additional testing (e.g., PET-CT or biopsy) when appropriate.

Clinical utility: In aggregate, the LLM narrative clarifies *why* the ensemble's decision is plausible in human terms and flags caveats salient to radiologists (e.g., missing priors, potential non-specificity of diffuse activation). This aligns with the broader aim of explainable decision support: pairing accurate discrimination and threshold behavior with transparent, auditable justifications that can be weighed alongside clinical context [10,12,14,20,29,30,36].

CHAPTER 5

CONCLUSION

5.1 Conclusion

This work presents a clinically aligned computer-aided diagnosis pipeline for benign–malignant lung nodule classification that combines full-slice pseudo-3D inputs, heterogeneous 2D backbones, attention-based stacking, and multi-layer explainability. Unlike many prior studies that crop tightly around nodules, we retain contextual anatomy by stacking adjacent axial slices $\{z-1, z, z+1\}$ and training on full slices. This choice preserves perinodular and lobar cues important to radiologists while remaining computationally tractable relative to full 3D CNNs [13,15,16]. Three complementary backbones—EfficientNetV2-S, DenseNet-201, and MobileViT-XXS—are fine-tuned under patient-level 5-fold CV with conservative, slice-coherent augmentations and fold-specific class reweighting (3.3–3.6). Their outputs are fused by a Multi-Attention Stacked Ensemble (MASE) that learns input-conditioned model-wise and class-wise weights, adapted here to full-slice pseudo-3D inputs and trained with clean out-of-fold logits (3.7.5), building on ensemble theory and recent stacking advances in lung CT [7,26,31].

Empirically, MASE delivers state-of-the-art discrimination on LIDC-IDRI [2,15]: across five folds (uncalibrated), AUC 0.98±0.00190, with concurrent gains in Accuracy, Sensitivity, Specificity, Precision, and F1 over the strongest single backbone (e.g., +0.0055 AUC and +0.0145 F1 versus EfficientNetV2-S; 4.2). Post-hoc temperature scaling is applied only to MASE to improve the probability scale used in downstream reporting; as expected for a monotone transform, ROC–AUC remains unchanged, while thresholded metrics at τ=0.5 exhibit modest, interpretable shifts (higher Precision and F1) [37,38]. To support clinical acceptance, we pair quantitative results with a two-layer explanation: per-model Grad-CAM and ensemble-level median CAM (faithful visual attributions [19]), plus a templated clinical narrative generated by DeepSeek-R1 from deterministic, auditable descriptors (no pixel or PHI exposure) [10,12,14,20,29,30,36]. Qualitative examples show saliency concentrated on nodule-centric regions with explicit caveats regarding the limits of saliency maps (3.9–4.2.6).

The overall contribution is a deployable and interpretable pipeline that (i) avoids potentially distortionary nodule cropping, (ii) leverages complementary inductive biases via attention-based stacking, and (iii) articulates decisions in clinician-oriented form. Limitations include reliance on a single public dataset (LIDC-IDRI) with internal cross-validation only [2,15]; lack of multi-institutional external validation; and the absence of formal calibration error reporting (e.g., ECE/Brier), even though we apply temperature scaling [37,38]. Inter-observer variability in the dataset [2,15] and potential domain shift across scanners remain open challenges. These constraints motivate the future directions below.

5.2 Future Works

External and prospective validation: A priority is multi-center external validation (e.g., NLST) and, ideally, prospective reader studies to assess impact on radiologist performance and workflow [13]. Stratified analyses across scanner vendors, slice thicknesses, and demographic subgroups will clarify robustness to domain shift.

Calibration assessment and decision analytics: We will quantify calibration with ECE, Brier score, and log-loss, and compare temperature scaling with alternatives (e.g., isotonic), maintaining the rank-preserving principle outlined by Guo et al. and Niculescu-Mizil & Caruana [37,38]. Operating-point selection can be tied to clinical utility (e.g., sensitivity-focused regimes for screening vs. precision-focused regimes for work-up).

Uncertainty and reliability: Incorporating uncertainty quantification (e.g., conformal risk control) and reporting risk bins with calibrated probabilities would strengthen safety signaling in borderline cases, complementing AUC/PR analyses.

Architectural extensions: Two natural directions are (i) multi-view pseudo-3D that integrates axial with limited coronal/sagittal context while preserving 2D efficiency [15,16], and (ii) lightweight 2.5D/3D hybrids that retain deployability yet capture richer volumetrics than three slices. Weakly supervised segmentation-aware heads could better localize evidence while keeping classification as the primary task.

Representation learning: We plan to explore self-supervised pretraining on large unlabeled chest CT collections to reduce label dependence and enhance generalization, before fine-tuning the present full-slice pipeline.

Explainability advances: Beyond Grad-CAM [19], counterfactual or concept-based explanations could expose human-interpretable features (e.g., coarse margin/spiculation surrogates) while remaining faithful to the classifier [10,12,14,20]. For narratives, domain-specific adaptation of DeepSeek-R1 to radiology style—still driven by the same deterministic descriptors—may further improve clarity and concision without relaxing hallucination safeguards [29,30,36].

Clinical integration: Practical steps include DICOM/PACS integration, on-device inference for privacy, and attention-overlay viewers that allow side-by-side review of axial slices, ensemble CAMs, and the LLM narrative. Finally, monitoring for dataset shift and continual learning with guardrails will be required for safe post-deployment updates.

Taken together, these directions extend the present findings—context-preserving inputs, attention-stacked ensembling, and auditable explanations—toward robust, evidence-based clinical decision support in thoracic imaging [7,10–12,14–16,19–20,26,31,36–38].

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R. L., *et al.* (2021). Global cancer statistics 2020: GLOBOCAN estimates. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249.

[2] Armato, S. G., *et al.* (2011). The LIDC/IDRI: A completed reference database. *Medical Physics, 38*(2), 915–931.

[3] Aerts, H. J. W. L., *et al.* (2014). Decoding tumour phenotype by radiogenomics. *Nature Communications, 5*, 4006.

[4] Litjens, G., *et al.* (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88.

[5] Dosovitskiy, A., *et al.* (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[6] Esteva, A., *et al.* (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29.

[7] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (pp. 1–15). Springer.

[8] McWilliams, A., *et al.* (2013). Probability of cancer in pulmonary nodules detected on first screening CT. *The New England Journal of Medicine, 369*(10), 910–919.

[9] *(Intentionally left blank to preserve numbering.)*

[10] Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813.

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

[12] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.

[13] Ardila, D., *et al.* (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine, 25*(6), 954–961.

[14] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312.

[15] Hancock, M. C., & Magnan, J. F. (2016). Lung nodule malignancy classification using LIDC. *Journal of Digital Imaging, 29*(5), 651–662.

[16] Huang, X., *et al.* (2017). 3D lung nodule classification with multi-view convolutional neural networks. *Pattern Recognition, 61*, 663–673.

[17] Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

[18] Suk, H. I., & Shen, D. (2013). Deep feature representation for AD/MCI. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

[19] Selvaraju, R. R., *et al.* (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626).

[20] Xie, J., *et al.* (2020). Explainable deep learning in medical imaging: A survey. *Medical Image Analysis, 65*, 101758.

[21] Han, F., *et al.* (2015). Texture feature analysis for computer-aided diagnosis of pulmonary nodules. *Academic Radiology, 22*(3), 282–289.

[22] Setio, A. A. A., *et al.* (2016). False-positive reduction in pulmonary nodule detection using multi-view convolutional networks. *IEEE Transactions on Medical Imaging, 35*(5), 1160–1169.

[23] Kumar, D., *et al.* (2020). Transfer learning for lung nodule classification. *Biomedical Signal Processing and Control, 62*, 102045.

[24] Ciompi, F., *et al.* (2017). Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific Reports, 7*(1), 1–9.

[25] Causey, J. L., *et al.* (2018). Prediction of lung nodule malignancy with CT scans. *Scientific Reports, 8*, 9286.

[26] Liao, F., *et al.* (2019). Deep learning ensemble for nodule malignancy. *Medical Physics, 46*(12), 5763–5774.

[27] Wang, X., *et al.* (2017). Lung nodule classification with deep feature fusion. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 370–374).

[28] Lundervold, A. S., & Lundervold, A. (2019). Overview of deep learning in medical imaging (MRI focus). *Zeitschrift für Medizinische Physik, 29*(2), 102–127.

[29] Cai, C. J., *et al.* (2019). Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*.

[30] Zhang, Z., *et al.* (2021). When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 148–156).

[31] Saha, U., & Prakash, S. (2023). Multi-attention stacked ensemble for lung cancer detection in CT scans. *Preprint/Proceedings*.

[32] Samek, W., Montavon, G., Lapuschkin, S., Binder, A., & Müller, K.-R. (2017). Explainable AI: Interpreting, explaining and visualizing deep learning. *arXiv preprint arXiv:1708.08296*.

[33] Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[34] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations (ICLR)*.

[36] DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. Technical report/Preprint.

[37] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[38] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

APPENDICES

Appendix A — Dataset, Ethics, and Label Policy

Dataset: LIDC-IDRI public CT collection; de-identified; used under the dataset's terms of use.
Task: Binary nodule classification (benign vs malignant).
Label policy: Binary label derived from radiologist malignancy scores. We adopt the common mapping: S1–S2 → benign (0), S4–S5 → malignant (1), S3 → excluded.
Inclusion/Exclusion: CT series with consistent reconstruction kernel and slice spacing; corrupted or incomplete studies removed.
Ethics: Public, de-identified data; no new human subjects; IRB not required (institutional policies may vary).

Appendix B — Splits, Class Balance, and Sampling

Cross-validation: Patient-level 5-fold CV; each patient appears in exactly one fold (no patient overlap).

Loss reweighting: Class-weighted BCEWithLogitsLoss with weights

$$\boldsymbol{w_c = \frac{N}{2\,N_c}}, \quad \boldsymbol{c \in \{0, 1\}},$$

where NN is fold size and $\boldsymbol{N_c}$ class-c count.
Sampler: PyTorch WeightedRandomSampler with per-example weight $\boldsymbol{w_{y_i}}$.

Appendix C — Data Augmentation (Slice-Coherent)

All augmentations applied identically to the three slices in the stack to preserve through-plane coherence. Final training set uses **no test-time augmentation (TTA)**.

Augmentation	**Range / Params**	**Prob.**
Horizontal flip	0/1 flip	0.5
Small rotation	±7∘7^\circ	0.3
Translation	±3–5 px	0.3
Brightness jitter	±5–10%	0.3
Contrast jitter	±5–10%	0.3
Gaussian noise	σ\sigma = 0.005–0.01	0.2
Light blur (anti-alias)	3×3	0.2

Appendix D — Stacked Ensemble (MASE) & Calibration

Stacking: Concatenate per-backbone logits $z = [z^{(1)}, z^{(2)}, z^{(3)}]$. Learn model-wise attention α_k and class-wise attention $\beta_{k,c}$ to produce ensemble logit $z*$. Inference uses the trained stack on backbone outputs. Out-of-fold training protocol avoids leakage.

Note: Calibration is monotonic; ROC/PR ranking is unchanged. Report calibrated operating-point metrics and PR curves; report AUC from uncalibrated scores.

Appendix E — Full Quantitative Results

G.1 Backbones — Per-Fold Metrics

Model	Fold	Accuracy	AUC	F1	Precision	Sensitivity	Specificity
efficientnetv2_s	1	0.9536	0.9816	0.9143	0.8918	0.9379	0.9592
efficientnetv2_s	2	0.9219	0.9749	0.8644	0.7969	0.9444	0.9138
efficientnetv2_s	3	0.9434	0.9762	0.8968	0.8624	0.9339	0.9467
efficientnetv2_s	4	0.9389	0.9767	0.8909	0.8412	0.9469	0.9361
efficientnetv2_s	5	0.9437	0.9734	0.8983	0.8566	0.9443	0.9435
densenet201	1	0.9219	0.9749	0.8641	0.7982	0.9418	0.9148
densenet201	2	0.9461	0.9711	0.8983	0.8937	0.9030	0.9616
densenet201	3	0.9297	0.9715	0.8717	0.8393	0.9067	0.9379

densenet201	4	0.9297	0.9732	0.8727	0.8345	0.9145	0.9352
densenet201	5	0.9280	0.9708	0.8728	0.8162	0.9378	0.9245
mobilevit_xxs	1	0.9280	0.9685	0.8719	0.8215	0.9288	0.9277
mobilevit_xxs	2	0.9202	0.9656	0.8532	0.8283	0.8797	0.9347
mobilevit_xxs	3	0.9038	0.9638	0.8331	0.7669	0.9119	0.9009
mobilevit_xxs	4	0.9246	0.9626	0.8604	0.8397	0.8821	0.9398
mobilevit_xxs	5	0.9011	0.9598	0.8221	0.7809	0.8679	0.9129

- **G.2 Backbones — Mean ± SD Across Folds**

Model	AUC	Accuracy	Sensitivity	Specificity	Precision	F1	Balanced Acc.
EfficientNetV2-S	0.9766 ± 0.0031	0.9403 ± 0.0116	0.9415 ± 0.0054	0.9399 ± 0.0168	0.8498 ± 0.0348	0.8929 ± 0.0181	0.9407 ± 0.0072
DenseNet-201	0.9723 ± 0.0017	0.9311 ± 0.0090	0.9208 ± 0.0179	0.9348 ± 0.0176	0.8364 ± 0.0359	0.8759 ± 0.0130	0.9278 ± 0.0042
MobileViT-XXS	0.9641 ± 0.0033	0.9155 ± 0.0123	0.8941 ± 0.0253	0.9232 ± 0.0161	0.8074 ± 0.0317	0.8481 ± 0.0203	0.9086 ± 0.0135

- **G.3 MASE (Uncalibrated) — Mean ± SD**

AUC	Accuracy	Sensitivity	Specificity	Precision	F1
0.9821 ± 0.0019	0.9486 ± 0.0048	0.9560 ± 0.0058	0.9459 ± 0.0083	0.8638 ± 0.0177	0.9074 ± 0.0076

(Balanced Accuracy ≈ 0.9510 from Sens/Spec means.)

- **G.4 MASE (Calibrated) — Mean ± SD**

Accuracy	Sensitivity	Specificity	Precision	F1
0.9517 ± 0.0046	**0.9493 ± 0.0070**	**0.9526 ± 0.0079**	≈ **0.8775** *(mean; SD n/a)*	**0.9120 ± 0.0074**

Note. Calibrated AUC is not reported (rank invariance); AUC shown elsewhere is from uncalibrated scores.

- **Appendix F — LLM Narrative System (DeepSeek-R1)**

Descriptor schema (no PHI/pixels): The LLM receives only deterministic numerics:


```
{
  "probability": 0.97,
  "class_label": "Malignant",
  "malignancy_score": 5.0,
  "diameter_mm": 15.0,
  "margin_score": 5.0,
  "spiculation_score": 2.0,
  "cam_desc": "Peak activation is broad in the central field",
  "cam_stats": {
    "largest_cc_area": 0.12,
```

```
    "centroid_distance_norm": 0.18,
    "radial_mass_frac": 0.28
  }
}
```

Prompt skeleton:

```
### AI Assistant Report: Malignancy Probability Assessment for Lung Nodule

#### Model Prediction Summary

The ensemble deep learning model predicted a **<risk> probability of malignancy (<prob>)** ...

### 1. Why the Model Reached This Decision

- Use probability and malignancy score; no unprovided clinical statements.

### 2. Support from Grad-CAM and Imaging Features

- Grad-CAM Activation Pattern: <cam_desc>.

### 3. Caveats, Limitations, and Missing Data

- Diameter: <d> mm, Margin: <m>, Spiculation: <s>.

### 4. Important Reminder

Decision support only; integrate with clinical judgment.
```

Decoding & safeguards: Model: DeepSeek-R1; temperature ≈ 0.2; strict template; sanitizer + fallback renderer; no free-text fields beyond the supplied descriptors.