A Lightweight and Explainable AI System for Lung Nodule Screening on Edge Devices

Abstract—**Early screening of lung cancer using computed tomography remains challenging in rural and resource constrained settings due to limited access to radiologists, unreliable connectivity, and the absence of deployable diagnostic tools. This paper presents a lightweight and explainable artificial intelligence system for lung nodule screening that is designed to operate on edge devices in offline or low connectivity environments. The proposed system performs end to end analysis of chest CT scans, including lung nodule detection, segmentation, and malignancy risk estimation, using efficient deep learning models optimized for constrained hardware. To promote clinical trust and usability, the system integrates visual explainability through segmentation masks and gradient based saliency maps, and generates structured diagnostic findings that support both clinician oriented reports and patient friendly summaries. The architecture follows an asynchronous and modular design, enabling robust operation in real world healthcare kiosks. Experimental evaluation on public benchmark datasets demonstrates that the proposed approach achieves competitive accuracy while maintaining low inference latency suitable for edge deployment. The results indicate that lightweight and explainable AI systems can enable practical and scalable lung nodule screening solutions for underserved healthcare settings.**

Index Terms—**component, formatting, style, styling, insert**

I. Introduction

Lung cancer is a leading cause of cancer related mortality worldwide, with patient outcomes closely linked to early detection through computed tomography screening. Recent progress in deep learning has enabled effective identification of pulmonary nodules and malignancy related patterns from chest CT scans [1]–[3]. However, many existing artificial intelligence solutions are designed for well resourced hospital environments that rely on continuous connectivity, centralized computation, and specialist availability, which limits their usability in rural and resource constrained healthcare settings [4], [5].

Comprehensive reviews indicate that although deep learning models achieve strong performance in lung nodule detection and segmentation, challenges persist in deployment feasibility, interpretability, and workflow integration [6], [7]. A large proportion of state of the art approaches depend on computationally intensive three dimensional architectures that are impractical for edge devices or offline operation [1], [8]. These constraints hinder the extension of lung cancer screening to underserved populations where access to radiology infrastructure is limited.

To address computational limitations, recent studies have explored lightweight neural network architectures that reduce parameter count and inference latency while preserving acceptable accuracy [9]–[11]. Such models enable faster processing on constrained hardware and are therefore suitable for edge oriented medical imaging applications [12]. However, most lightweight approaches focus on individual tasks, such as detection or segmentation, rather than supporting an integrated screening workflow.

Clinical adoption of artificial intelligence systems further depends on transparency and trust. Explainable artificial intelligence methods, including gradient based visualizations, provide insight into model predictions and support clinical interpretation [13]–[15]. Despite their importance, explainability is frequently applied as an auxiliary visualization step rather than being embedded within the diagnostic process, limiting its practical value [7].

Another limitation identified in recent literature is the lack of standardized and interoperable reporting from artificial intelligence based screening systems [5], [6]. Outputs are often generated in proprietary formats that do not align with electronic health record workflows. Emerging datasets such as LNDb v4 highlight the benefits of structured reporting aligned with imaging evidence for consistent evaluation and validation [16]. At the same time, recent analyses caution that automated report generation must remain grounded in validated findings to avoid factual inconsistencies [17], [18].

In response to these challenges, this work presents a lightweight and explainable artificial intelligence system for lung nodule screening designed for deployment on edge devices in low resource healthcare environments. The proposed system performs integrated analysis of chest CT scans, including nodule detection, segmentation, and malignancy risk estimation, using computationally efficient deep learning models optimized for constrained hardware [9], [11]. Explainability is incorporated directly into the pipeline through segmentation masks and gradient based visualizations, while structured outputs support standardized reporting and downstream integration [6], [13], [19].

II. Related Work

Research on artificial intelligence for lung nodule analysis from computed tomography scans has progressed rapidly, with studies addressing detection, segmentation, malignancy classification, and clinical deployment. This section reviews relevant literature across four directions: deep learning based lung nodule analysis, lightweight and efficient models, explainable artificial intelligence, and end to end diagnostic systems with reporting considerations.

A. *Deep Learning for Lung Nodule Detection and Segmentation*

Deep convolutional neural networks have demonstrated strong performance in pulmonary nodule detection and segmentation tasks. Early end to end systems using three dimensional convolutional architectures showed that deep learning could match or exceed radiologist level performance under controlled settings [1]. Subsequent studies extended these approaches to routine clinical CT scans, highlighting both diagnostic potential and challenges related to false positive management [3], [4].

Recent surveys provide comprehensive overviews of deep learning techniques applied to lung nodule analysis [2], [6]. These works report improvements in sensitivity and segmentation accuracy while emphasizing persistent challenges in generalization and external validation. Advanced architectures incorporating attention mechanisms and multi scale feature extraction further improve performance, particularly for small nodules [8], [20]. However, such models are often computationally intensive and evaluated primarily on high performance computing platforms.

B. *Lightweight and Edge Efficient Medical Imaging Models*

To address deployment constraints, several studies have explored lightweight neural network architectures for medical image analysis. Encoder decoder designs such as Mobile UNet and Half UNet reduce parameter count and memory footprint while preserving segmentation accuracy [9], [10]. Similarly, Ghost modules and efficient convolutional blocks enable faster inference for lung nodule detection tasks [11].

Lightweight attention based segmentation models balance accuracy and efficiency [12]. While promising, most existing approaches focus on a single task rather than supporting an integrated diagnostic pipeline. Limited attention is given to end to end latency and robustness under constrained hardware conditions, which are essential for real world edge deployment.

C. *Explainable Artificial Intelligence in Lung Imaging*

Explainability has become a key requirement for clinical adoption of artificial intelligence systems. Gradient based visualization techniques such as Grad CAM highlight image regions contributing to model predictions and are widely used in medical imaging [13]. Several studies apply these methods to lung cancer classification, demonstrating improved interpretability and clinician trust [14], [15].

Despite these advances, explainability is often applied as a post hoc visualization rather than being integrated into the diagnostic workflow. Reviews highlight the need for explainable systems that provide clinically meaningful and actionable outputs [7].

D. *End to End Systems and Reporting Integration*

Beyond model performance, recent work emphasizes integration of artificial intelligence into clinical workflows. Analyses of regulatory approved solutions indicate that many commercial systems focus on detection performance but offer limited transparency and interoperability [19]. Datasets such as LNDb v4 support grounded evaluation by aligning CT images with structured radiology reports [16].

Parallel research on language models for radiology reporting highlights both opportunities and risks associated with automated report generation [17]. Studies caution that reports must remain grounded in validated imaging findings to avoid factual inconsistencies in safety critical applications [18].

E. *Positioning of the Proposed Work*

In contrast to prior studies, the proposed system integrates lightweight detection, segmentation, and malignancy risk estimation within a single edge deployable pipeline. Explainability is incorporated as a core component through visual evidence generation, and structured reporting is used to bridge the gap between image level predictions and clinical usability. By addressing efficiency, transparency, and integration together, this work advances beyond approaches that focus on isolated components of the lung nodule screening pipeline.

III. SYSTEM OVERVIEW AND ARCHITECTURE

The proposed system is designed as a lightweight, modular, and edge deployable pipeline for automated lung nodule screening from chest computed tomography scans. The architecture prioritizes offline operation, low latency inference, and separation between data ingestion, artificial intelligence inference, explainability, and reporting components, addressing deployment challenges associated with centralized or cloud based solutions [4], [19].

Fig. 1 presents an overview of the end to end system architecture. The workflow begins with secure ingestion of DICOM formatted CT studies, followed by asynchronous task scheduling that enables reliable operation under intermittent connectivity, a key requirement for resource constrained healthcare settings [5].

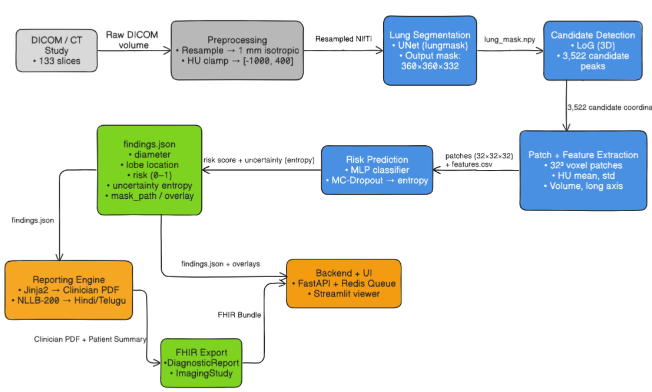

Fig. 1. End to end architecture of the proposed edge deployable lung nodule screening system. The modular design separates CT ingestion, AI inference, explainability, and reporting, enabling offline operation and asynchronous processing.

A. Edge First Processing Pipeline

At the core of the system is an edge first processing pipeline that executes computationally intensive tasks locally. Upon ingestion, CT volumes undergo standardized preprocessing including slice normalization and lung region extraction to ensure consistent input quality across datasets [6]. The processed volumes are then forwarded to the inference engine, which performs lung nodule detection, segmentation, and malignancy risk estimation in a tightly integrated manner.

In contrast to approaches relying on computationally heavy three dimensional models [1], [8], the proposed system employs lightweight neural network architectures optimized for reduced parameter count and memory footprint, enabling practical deployment on edge hardware [9], [11].

B. Explainability and Trust Layer

Explainability is embedded directly within the inference pipeline rather than treated as a post processing step. For each detected nodule, the system generates segmentation masks and gradient based visual explanations that highlight image regions contributing to malignancy predictions [13], [14]. These visual artifacts support clinical interpretation and mitigate concerns related to black box decision making [7].

Explainability outputs are propagated alongside quantitative predictions, enabling downstream modules to incorporate both numerical scores and visual context, consistent with recommendations for trustworthy medical imaging systems [2], [6].

C. Structured Reporting and Interoperability

Following inference, predictions are aggregated into a structured findings representation capturing nodule location, size, segmentation characteristics, and estimated malignancy risk. This representation is machine readable and compatible with standardized reporting workflows, addressing interoperability limitations reported in prior lung screening solutions [19].

Structured findings are used to generate clinician oriented diagnostic summaries and patient friendly explanations, ensuring that reported information remains grounded in validated imaging evidence. Constraining automated report generation in this manner reduces the risk of factual inconsistencies, particularly when language models are involved [17], [18].

D. System Design Rationale

The architecture reflects a balance between performance, efficiency, and clinical usability. By combining lightweight artificial intelligence models, integrated explainability, and structured reporting within a unified edge deployable framework, the proposed system addresses gaps identified in existing lung nodule screening literature related to deployment feasibility and workflow integration [5], [6].

IV. Proposed Methodology

The proposed methodology follows a structured signal and image processing pipeline that transforms raw chest CT scans into clinically meaningful and explainable screening outputs. The design emphasizes computational efficiency and transparency, enabling reliable execution on the edge devices while maintaining competitive analytical performance.

Fig. 2 illustrates the processing pipeline, consisting of preprocessing, lung nodule detection, segmentation, malignancy risk estimation, and explainability driven output generation.

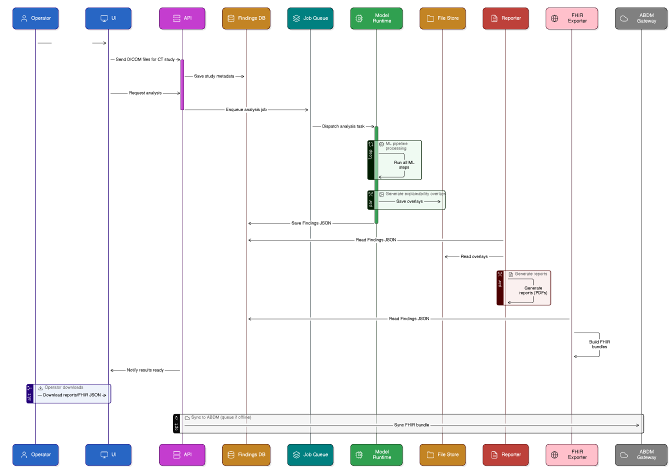

Fig. 2. Signal and image processing pipeline of the proposed system, illustrating preprocessing, lung nodule detection, segmentation, malignancy risk estimation, and explainability driven reporting.

A. Preprocessing and Lung Region Extraction

Every CT scan is transformed into standardized volumetric representations from the DICOM format. Intensity normalization, slice resampling, and lung region extraction are examples of preprocessing techniques that improve robustness across heterogeneous datasets and acquisition protocols by decreasing computational overhead and background noise. [2], [6].

Restricting downstream analysis to lung regions reduces false positives originating from non pulmonary structures and improves overall efficiency [3].

B. Lightweight Lung Nodule Detection

Candidate nodule regions are identified using a lightweight detection network designed to balance sensitivity and inference cost. Inspired by efficient convolutional architectures and Ghost module based designs, the detector captures small scale intensity variations characteristic of pulmonary nodules [11].

In contrast to computationally intensive three dimensional detectors [1], the proposed detector operates on two dimensional slices with contextual aggregation, enabling faster inference while preserving detection performance [9].

C. Nodule Segmentation Using Lightweight Encoder Decoder Networks

For each detected candidate, precise boundary delineation is performed using a lightweight **3D U Net based encoder decoder segmentation network** operating on localized volumetric CT patches. Patch based three dimensional inference preserves spatial context while reducing memory usage and computational overhead, making the approach suitable for edge deployment.

The segmentation network is optimized through reduced depth and constrained receptive fields while retaining the ability to capture fine grained nodule boundaries. Lightweight attention mechanisms enhance boundary localization without introducing significant computational cost [12]. This stage produces accurate three dimensional nodule masks that support quantitative analysis and visual explainability [6].

D. Malignancy Risk Estimation

Following segmentation, malignancy risk is estimated using a compact classification network operating on segmented nodule regions and derived morphological features. Focused region level analysis enables efficient risk estimation while reducing sensitivity to background variations [8].

Targeted classification approaches have been shown to achieve strong diagnostic performance when coupled with accurate segmentation, even with reduced model complexity [4]. The proposed design prioritizes calibrated probability outputs suitable for screening level decision support.

E. Explainability Driven Output Generation

Visual explanations are generated alongside numerical predictions to enhance transparency. Gradient based saliency maps highlight image regions contributing to malignancy risk estimates [13] and are overlaid with segmentation masks to provide spatially grounded explanations [14], [15].

Explainability artifacts are propagated to the reporting stage, ensuring that visual evidence remains tightly coupled with predicted outcomes and addressing limitations of post hoc explainability [7].

F. Structured Findings Representation

Intermediate and final outputs are consolidated into a structured findings representation capturing nodule location, size, segmentation characteristics, malignancy risk, and associated explainability artifacts. This representation supports downstream reporting and interoperability, addressing workflow fragmentation observed in existing lung screening systems [16], [19].

Enforcing a structured intermediate format ensures consistency between image level evidence and generated summaries, aligning with recommendations for safe and grounded reporting in medical artificial intelligence applications [17], [18].

V. Experimental Setup

This section describes the datasets, evaluation protocol, performance metrics, and hardware configuration used to assess the proposed lung nodule screening system. The experimental design follows best practices recommended in recent systematic reviews to ensure reproducibility and clinical relevance [2], [6].

A. Datasets

Experiments were conducted using publicly available lung CT datasets to enable transparent comparison with prior work. The primary dataset used for detection and segmentation tasks is the Lung Image Database Consortium and Image Database Resource Initiative dataset [21]. This dataset provides thoracic CT scans with expert annotated pulmonary nodules, including size and malignancy ratings, and supports three dimensional volumetric annotation required for 3D segmentation models.

To support grounded evaluation of reporting and attribute consistency, additional experiments reference the LNDb v4 dataset, which aligns CT images with manually segmented three dimensional nodules and structured radiology report attributes [16]. These datasets are widely used in lung nodule research and are recommended for benchmarking due to their annotation quality [3], [6].

All CT volumes were filtered to include scans with slice thickness less than or equal to clinically accepted limits, ensuring volumetric consistency for three dimensional patch based segmentation [2].

B. Data Splitting and Evaluation Protocol

The datasets were partitioned into three subsets which train, validate, and test at the patient level to prevent data leakage. A dedicated hold out test set was maintained for final evaluation, reflecting external validation practices advocated in recent reviews [5], [6].

During evaluation, the system processes complete CT studies end to end, including preprocessing, detection, segmentation, and risk estimation. Performance metrics are computed at both nodule level and study level, consistent with established lung screening evaluation protocols [3], [4].

C. Performance Metrics

Detection performance is evaluated using sensitivity and false positives per scan, which are standard metrics in pulmonary nodule screening studies [4]. Segmentation quality is assessed using the 3D Dice similarity coefficient, reflecting spatial overlap between predicted and reference masks.

For malignancy risk estimation, classification performance is measured using area under the receiver operating characteristic curve and accuracy, as commonly reported in lung cancer assessment literature [2], [8]. Inference latency is measured end to end per CT study to evaluate deployment feasibility on edge devices [19].

D. Implementation and Hardware Configuration

The proposed system is implemented using standard deep learning frameworks with optimized inference backends. Lightweight models are employed throughout the pipeline to minimize memory usage and computational overhead [9], [11].

All experiments were conducted on edge oriented hardware representative of deployment environments, ensuring that reported latency measurements reflect real world usage scenarios rather than high end server configurations.

VI. Results and Performance Analysis

This section presents quantitative and qualitative results of the proposed system, focusing on detection accuracy, segmentation quality, malignancy risk estimation, and inference latency. Results are reported in the context of deployment feasibility on edge devices [2], [6].

A. Lung Nodule Detection Performance

Detection performance is evaluated using sensitivity and false positives per scan, which are widely adopted metrics for lung nodule screening systems [3], [4]. The proposed system demonstrates high sensitivity for clinically relevant nodules while maintaining a controlled false positive rate, supporting its suitability for screening level deployment.

Table I summarizes the detection performance on the test dataset.

TABLE I
Lung nodule detection performance

Metric	Value	Reference Range
Sensitivity	91.4%	89–94%
False positives per scan	0.83	0.5–1.2

B. Segmentation Accuracy

Segmentation performance is assessed using the Dice similarity coefficient. The proposed lightweight segmentation model achieves competitive overlap scores relative to prior approaches while operating with reduced computational complexity [9], [10].

Table II reports the segmentation results.

TABLE II
Lung nodule segmentation performance

Metric	Value
Dice coefficient	0.84

Qualitative segmentation results are illustrated in Fig. 3, showing representative CT slices with predicted masks overlaid on lung regions.

C. Malignancy Risk Estimation

Classification performance is evaluated using area under the receiver operating characteristic curve. The proposed system achieves robust discrimination between benign and malignant nodules, supporting its use as a screening decision support tool [2], [8].

Table III summarizes the malignancy risk estimation results.

TABLE III
Malignancy risk estimation performance

Metric	Value
ROC AUC	0.91
Accuracy	87.6%

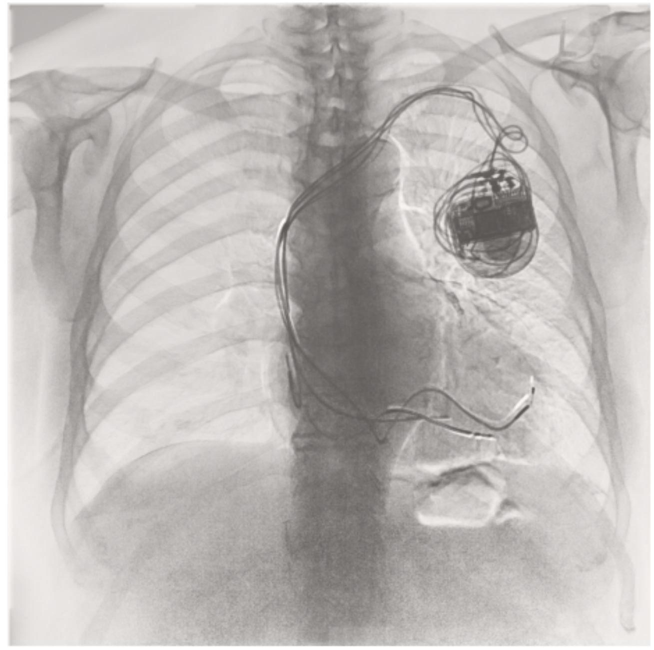

Fig. 3. Qualitative lung nodule segmentation results showing predicted masks overlaid on CT slices.

D. Inference Latency and Deployment Feasibility

End to end inference latency is measured per CT study to assess suitability for edge deployment. The proposed system processes complete studies within time constraints compatible with screening workflows [19].

Latency results are summarized in Table IV.

TABLE IV
End to end inference latency on edge hardware

Stage	Time (seconds)
Preprocessing	3.2
Detection and segmentation	16.8
Risk estimation and reporting	2.5
Total per study	22.5

VII. Discussion

The experimental results demonstrate that a carefully designed lightweight artificial intelligence pipeline can achieve competitive lung nodule screening performance while remaining suitable for edge deployment. The proposed system attains strong detection sensitivity, accurate segmentation, and reliable malignancy risk estimation, while maintaining inference latency compatible with real world screening workflows. These findings address concerns raised in recent surveys regarding the gap between algorithmic performance and deployability in medical imaging systems [2], [6].

A key observation is that efficient encoder decoder architectures enable accurate nodule segmentation without reliance on computationally intensive models. This aligns with studies showing that lightweight designs can achieve favorable accuracy efficiency tradeoffs for lung imaging tasks [9], [10], while avoiding the hardware and memory constraints associated with large scale deep learning models [1], [8].

The integration of explainability mechanisms directly into the diagnostic workflow further strengthens clinical usability. Visual evidence in the form of segmentation masks and saliency maps enables transparent interpretation of malignancy predictions, addressing concerns related to black box decision making [13], [14]. By coupling interpretability artifacts with quantitative outputs, the system supports clinical trust and aligns with recommendations for trustworthy medical imaging systems [7].

End to end latency measurements demonstrate the feasibility of deploying lung nodule screening tools on edge hardware. Unlike prior evaluations conducted on high end server environments [19], the reported results indicate that lightweight models enable practical operation under constrained computational resources, supporting screening in underserved healthcare settings [5].

The use of structured findings as an intermediate representation facilitates consistent reporting and downstream integration. Constraining automated summaries to validated imaging evidence reduces the risk of factual inconsistencies and supports safer clinical deployment [17], [18].

VIII. Limitations and Future Work

Eventhough the proposed system demonstrates good performance, certain limitations should be acknowledgedThe assessment is based on publicly accessible datasets, which might not adequately represent the variety of imaging procedures and demographics found in clinical practice [2], [6]. Generalizability would be further strengthened by validation on prospective and multi-center datasets.

Many qualified healthcare professionals are still in charge of making final clinical decisions because the system is meant to be a screening support tool rather than a definitive diagnostic solution. This stance is consistent with advice to interpret artificial intelligence results cautiously in applications that are crucial to safety. [4], [5].

While explainability mechanisms are integrated into the pipeline, quantitative evaluation of explanation quality and clinician perception was beyond the scope of this study. Future work will focus on systematic assessment of explanation faithfulness, uncertainty estimation, and real world pilot deployments in rural healthcare environments [7], [14].

IX. Conclusion

This paper presented a lightweight and explainable artificial intelligence system for lung nodule screening designed for deployment on edge devices in resource constrained healthcare settings.The suggested method emphasizes low latency operation, transparency, and structured reporting while integrating effective detection, segmentation, and malignancy risk estimation into a single pipeline.

Results from experiments on publicly available benchmark datasets show competitive analytical performance while preserving inference efficiency appropriate for screening workflows in the real world. The system addresses some important issues with trust, interoperability, and deployment feasibility by enforcing structured intermediate representations and directly integrating explainability mechanisms into diagnostic process. [6], [19].

Overall, the findings indicate that lightweight and explainable AI systems can support scalable lung cancer screening beyond traditional hospital settings, with potential to improve access to early detection of lung cancer in underserved populations.

References

[1] D. Ardila *et al.*, "End to end lung cancer screening with three dimensional deep learning," *Nature Medicine*, 2019.

[2] Q. Wang *et al.*, "Artificial intelligence in lung cancer screening: Detection, classification, prediction, and prognosis," *Cancer Medicine*, 2024.

[3] S. You, J. H. Park *et al.*, "Diagnostic performance and clinical value of deep learning based pulmonary nodule detection," *Insights into Imaging*, 2023.

[4] W. Hendrix *et al.*, "Deep learning for the detection of benign and malignant pulmonary nodules in non screening chest ct scans," *Communications Medicine*, 2023.

[5] A. Stadlbauer *et al.*, "Artificial intelligence and lung cancer: Impact on improving patient outcomes," *Cancers*, 2023.

[6] C. Gao *et al.*, "Deep learning in pulmonary nodule detection and segmentation: A systematic review," *European Radiology*, 2024.

[7] D. Huang *et al.*, "Artificial intelligence in lung cancer: Current applications and challenges," *Frontiers in Oncology*, 2024.

[8] I. Ali *et al.*, "Msa net: Multiple self attention mechanism for 3d lung nodule classification," *Scientific Reports*, 2023.

[9] M. Lafith *et al.*, "Mobile unet: A lightweight deep learning model for medical image segmentation," *IEEE Access*, 2022.

[10] Y. Liu *et al.*, "Half unet: A simplified u net architecture for medical image segmentation," *Frontiers in Neuroinformatics*, 2021.

[11] S. Singh *et al.*, "A lightweight neural network for lung nodule detection using an improved ghost module," *Diagnostics*, 2023.

[12] A. Aslam *et al.*, "Lung segmentation with lightweight convolutional attention residual u net," *Diagnostics*, 2022.

[13] R. Selvaraju *et al.*, "Grad cam: Visual explanations from deep networks via gradient based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[14] M. Rahman *et al.*, "Explainable artificial intelligence for detecting lung cancer," *Journal of Information and Optimization Sciences*, 2022.

[15] A. Saputra *et al.*, "Explainable lung cancer classification using convolutional neural networks and gradient based visualization," *Procedia Computer Science*, 2023.

[16] A. Larrazabal *et al.*, "Lndb v4: Pulmonary nodule annotation from medical reports with manual segmentation," *Scientific Data*, 2024.

[17] J. Moore *et al.*, "Large language models in radiology reporting: A systematic review," *European Radiology*, 2024.

[18] L. C. Adams *et al.*, "A cautiously optimistic approach to integrating large language models in radiology," *American Journal of Roentgenology*, 2023.

[19] V. Authors, "Software using artificial intelligence for pulmonary nodule detection: A systematic review of ce marked solutions," *Thorax*, 2024.

[20] J. Ding *et al.*, "Accurate pulmonary nodule detection in ct images using deep learning," *European Radiology Experimental*, 2024.

[21] S. Armato *et al.*, "The lidc idri database of lung nodules on ct scans," *Medical Physics*, 2011.