ÇANKAYA UNIVERSITY

CENG 407 – FINAL REPORT

Project title : LungXai

Authors

Orkun Oğuztürk-202111078

Can Berk Meşe-202111045

Ömer Faruk Şahin-202111073

Barbaros Murat Dönmez-202011019

Furkan Çoban-202011204

Arda Kaan Bakır-202111064

Elif Güngör-202111077

CONTENT

Abstract

Lung cancer is one of the leading causes of cancer-related mortality worldwide, and early detection of malignant pulmonary nodules plays a critical role in improving patient outcomes. Although low-dose computed tomography (LDCT) enables early identification of lung nodules, manual assessment remains challenging due to large scan volumes, subtle visual differences between benign and malignant lesions, and inter-observer variability among radiologists. These challenges motivate the development of reliable and interpretable artificial intelligence–based decision support systems.

In this study, we present LungXai, a modular deep learning framework for lung nodule malignancy classification using chest CT images. The system adopts a 2.5D input representation, in which a small stack of adjacent axial slices is treated as a multi-channel input to a two-dimensional convolutional neural network. This approach captures limited volumetric context while maintaining computational efficiency. A ResNet-18 architecture is employed as a baseline classifier within the MONAI medical imaging ecosystem, ensuring standardized preprocessing, reproducibility, and extensibility.

The proposed pipeline processes raw DICOM images and XML-based radiologist annotations from the LIDC-IDRI dataset to construct annotation-guided 2.5D samples. Experimental evaluation demonstrates that meaningful discrimination between benign and malignant nodules can be achieved even with a lightweight baseline model. The analysis further highlights that ROC-AUC is a more reliable evaluation metric than accuracy in this imbalanced medical classification setting. In addition, a prototype user interface is developed to visualize model outputs and demonstrate practical system integration.

Rather than presenting a fully coupled end-to-end clinical solution, LungXai follows a baseline-first and progressive development strategy. More advanced extensions are intentionally deferred to later stages, allowing the current work to establish a reproducible and extensible foundation for anatomically grounded and interpretable lung nodule analysis.

1. Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide [1]. Although low-dose computed tomography (LDCT) enables the early detection of pulmonary nodules, clinical evaluation remains challenging due to the large number of slices per scan, subtle visual differences between benign and malignant nodules, and inter-observer variability in malignancy assessment. These challenges motivate the development of computer-aided decision support systems that can assist radiologists by providing consistent and reproducible predictions.

In recent years, deep learning–based approaches have become widely adopted for pulmonary nodule analysis. However, medical imaging introduces domain-specific constraints that are not typically encountered in natural-image applications, including heterogeneous voxel spacing, intensity calibration in Hounsfield Units (HU), and the use of specialized data formats such as DICOM and XML-based annotations. Addressing these constraints requires a controlled and standardized experimental pipeline.

In this study, we present **LungXai**, a modular research-oriented framework for lung nodule malignancy classification using chest CT images. The current phase of the project focuses on establishing a reliable baseline classification pipeline based on (i) a 2.5D input representation constructed from adjacent axial slices, (ii) a ResNet-18 classification model, and (iii) a prototype user interface (UI) for visualizing model outputs. The system is implemented using the MONAI framework to support standardized preprocessing and reproducible experimentation.

Rather than introducing a fully integrated end-to-end clinical system, this work follows a baseline-first and progressive development strategy. More advanced components - such as segmentation-assisted region-of-interest refinement, stronger classification backbones, and explainable artificial intelligence (XAI) techniques - are outside the scope of the current implementation and are not included in this phase.2. Previous Studies

2.1 Previous Studies on Deep Learning–Based Lung Nodule Analysis

This section reviews prior studies related to deep learning–based lung nodule classification, focusing on the evolution of model architectures, comparative performance against clinical standards, the role of contextual information and explainability, and the overall progression of research in this domain.

2.1.1 Evolution of Deep Learning Architectures

Deep learning (DL) has fundamentally transformed medical image interpretation, replacing traditional radiomics pipelines that relied on handcrafted features with data driven Convolutional Neural Networks (CNNs). Early DL approaches utilized 2D CNNs, but these lacked volumetric awareness across contiguous CT slices [3]. This limitation led to the adoption of 3D CNNs as the standard for volumetric tasks like lung nodule detection. However, 3D convolution presents significant practical barriers, including cubic computational cost and the requirement for extensive annotated datasets, scaling poorly for large CT volumes [2].

To overcome the tradeoff between volumetric context and computational efficiency, researchers introduced 2.5D architectures. These models utilize a small stack of consecutive slices (multi-slice context) as input to a 2D network, effectively capturing local 3D context with reduced complexity [2]. Comparative analyses confirm the efficacy of this approach, showing that 2.5D DenseNet models can achieve a diagnostic performance comparable to that of thoracic radiologists [2]. Furthermore, a 2.5D DenseNet model demonstrated superior overall performance and discrimination, achieving a

significantly higher Area Under the Curve (AUC) of 0.921 compared to a 3D DenseNet (AUC 0.835) and a size based logistic model (AUC 0.836) in differentiating invasive pulmonary adenocarcinomas [2]. The 2.5D approach also showed a high specificity of 88.2% at 90% sensitivity [2].

2.1.2 DL Performance Compared to Clinical Standards

Recent systematic reviews and meta analyses validate the superior performance of DL-based models in predicting the risk of malignancy in CT-detected pulmonary nodules compared to conventional clinical tools [6]. A comprehensive meta analysis across seventeen externally validated studies (comprising 9,884 nodules) yielded the following key findings [6]:

- DL vs. Physician Judgement: DL-based models demonstrated a 1.6% increase in sensitivity compared to physician judgement alone [6]. While specificity was similar (pooled DL specificity 0.77 vs. physician 0.81), DL models showed superior pooled AUC of 0.86 compared to physician readers alone (AUC 0.83) [6].
- DL vs. Clinical Risk Models: DL models were 14.5% more sensitive and 7.4% more specific than clinical risk models (such as the Brock and Mayo models) alone [6]. The pooled AUC for DL models (0.86) was significantly superior to clinical risk models (AUC 0.79) [6].

This evidence suggests that DL-based models, such as DenseNet-121, which achieved an accuracy of 90.39% and specificity of 93.65% in comparative studies, are already justified for routine deployment alongside experienced physicians [4, 6].

2.1.3 The Role of Nodule Context and Explainability

While CNNs are effective in feature extraction, studies indicate that high performance often relies on hybrid architectures and considering the nodule's surrounding environment [3, 5].

- Impact of Fibrotic Microenvironment: Novel research has demonstrated that incorporating the surrounding pulmonary context, particularly pulmonary fibrosis (fibrotic microenvironment), significantly improves nodule malignancy classification [5]. A 3D Attention Gated Network (3D AG-Net) trained with semantic fibrosis metadata achieved the best results with 80.84% accuracy and 0.89 AUC [5]. This study confirmed that models trained with both the nodule and its microenvironment outperformed models trained on the nodule alone, with the ability to detect malignant nodules increasing by 9.21% when fibrosis data was presented [5].

- Need for Hybrid Models and XAI: The conventional convolution process may be ineffective for feature extraction alone, leading to the development of hybrid models utilizing attention blocks to overcome these limitations [3]. Furthermore, studies focusing on Explainable AI (XAI) emphasize that predictive estimates are sometimes not understood or interpreted by clinicians [3]. The development of visualizable models, such as the 3D AG-Net, which uses attention gates to filter and visualize features at different network depths, provides crucial interpretability that is highly valued in clinical settings [5].

In conclusion, the literature points toward the efficacy of 2.5D and attention based architectures for achieving high performance with computational efficiency [2], and highlights the necessity of integrating these systems into a standardized, clinically translatable workflow (e.g., using frameworks like MONAI) that can handle complex data constraints and provide the required diagnostic accuracy and explainability [2, 6, 5].

2.1.4 Evolution of Research from Past to Present

Research on AI-based lung nodule analysis has evolved from early 2D CNN approaches toward methods that better capture volumetric context and clinical usability. While 2D models are computationally efficient, their inability to represent inter-slice continuity motivated 3D CNNs, which improved spatial awareness but introduced substantial computational cost. To balance these trade-offs, 2.5D architectures emerged as a practical compromise by stacking adjacent slices as multi-channel input, capturing local 3D context with 2D-level efficiency. More recently, attention mechanisms and explainability techniques (e.g., Grad-CAM) have gained importance to address the clinical “black box” concern, while unified ecosystems such as MONAI have become central by standardizing preprocessing, training, evaluation, and deployment pathways for real-world medical imaging workflows.

2.2 Overview of Existing Frameworks

The nnU-Net framework introduced automated configuration for segmentation pipelines, standardizing preprocessing, architecture selection, and training parameters. It has achieved state of the art results in organ segmentation challenges but is task specific and lacks built in modules for classification, explainability, or deployment. Integrating nnU-Net into classification workflows typically requires additional frameworks.

TorchIO focuses on preprocessing and augmentation, enabling advanced data transformations such as elastic deformations, bias field simulation, and random noise. Although highly useful for dataset preparation, TorchIO does not provide model architectures, evaluation metrics, or deployment infrastructure.

Both nnU-Net and TorchIO contribute valuable utilities; however, they function as partial solutions. In contrast, MONAI integrates multiple components commonly required in medical AI research within a single coherent ecosystem. For a project that aims to progress from research to clinically usable outcomes, such integration ensures maintainability, reproducibility, and compliance with hospital IT standards.

2.3 MONAI: The Unified Medical AI Ecosystem

The Medical Open Network for AI (MONAI), founded jointly by NVIDIA and King's College London, represents a purpose built ecosystem for medical imaging research and deployment. It is fully open source (Apache 2.0 license) and implemented in Python using the PyTorch backend. MONAI's modular architecture consists of several tightly integrated subsystems.

The monai.transforms module provides domain specific image transformations, including intensity normalization, resampling, cropping, random affine transformation, and Gaussian noise addition. The monai.networks library contains deep learning architectures optimized for medical imaging, such as DenseNet, UNet, AttentionUNet, and SwinUNETR, supporting 2D, 2.5D, and 3D inputs. The monai.metrics module implements standardized evaluation metrics including Dice coefficient, Hausdorff distance, AUC, and FROC. The monai.inferers package enables efficient sliding window and patchbased inference for large 3D volumes, while monai.visualize provides utilities for Grad-CAM, feature maps, and training visualizations. Finally, monai.apps includes pre trained models and ready to use pipelines for segmentation, classification, and detection tasks.

MONAI's data pipeline supports flexible input dimensionality. Through its composable transforms, multiple adjacent CT slices can be concatenated along the channel dimension to simulate a 2.5D stack. The framework's CacheDataset and SmartCacheDataset classes allow efficient on thefly augmentation of slice groups, while sliding_window_inference enables inference on large 3D scans by processing 2.5D patches sequentially—an efficient hybrid strategy between 2D and 3D.

Furthermore, MONAI provides native support for explainable AI via Grad-CAM and Integrated Gradients, allowing visualization of malignancy driving regions directly on CT slices. Deployment-oriented components such as MONAI Deploy are considered future work and are not part of the current implementation.

Unlike modular utilities such as TorchIO or nnU-Net MONAI supports a unified workflow from preprocessing to inference within a single standardized environment." This eliminates compatibility issues between libraries, reduces code redundancy, and improves experiment traceability for research purposes. Therefore, the current study exclusively adopts MONAI for all stages of development[7]. Table 2.3.1 summarizes the key differences between MONAI, nnU-Net, and TorchIO, highlighting MONAI's advantage as a unified end-to-end medical AI framework.

Framework Comparison

Table 2.3.1: Comparison of MONAI, nnU-Net, and TorchIO frameworks in terms of functionality, flexibility, and deployment support.

Criterion	**MONAI**	**nnU-Net**	**TorchIO**
Goal	An end to end framework for medical imaging and deep learning.	Self adapting automatic segmentation system.	A library for medical image preparation and augmentation.
Input Formats	DICOM, NIFTI, PNG, JPG, 3D volumetric data.	NIFTI heavy.	NIFTI, DICOM, etc.; focus: preprocessing.
Preprocessing Features	Medical image normalization, spacing adjustment, intensity transforms, lung windowing, DenseNet support.	Automatic voxel spacing adjustment, normalization, resampling.	Built in loading, non rigid transforms, random intensity, bias field noise, elastic deformation.
Model Architecture	Systems supporting 2D / 3D U-Net, SegResNet, DenseNet, Swin-UNETR, Vision Transformer based systems.	2D/3D UNet/3D cascade UNet automatic configuration.	Model is robust, users add their own models.
Core Feature	Education -	Automatic	Modern augmentation and

	Assessment - XAI - Inference - Clinical deploy pipeline.	hyperparameter selection and preprocessing.	data pipeline control.
2D-3D Support	Native 2D and 3D pipeline.	2D, 3D, 3D cascade.	2D & 3D
Explainability (XAI)	GRAD-CAM, IG, SHAP integration.	Not natively present; external.	No
Annotation / GUI	MONAI Label (3D Slicer / OHIF support).	No annotation support.	Annotation support.
Hospital / PACS Integration	Yes (MONAI Deploy DICOM, HL7, FHIR).	No	No
Flexibility	Segmentation + classification + detection + registration.	Only segmentation.	Preprocessing library for every model type.
Optimization Strategy	Customizable loss functions, mixed precision, distributed training.	Fully automatic model selection and hyperparameter optimization.	Data pipeline acceleration (patch based training support).
Advantage	Complete ecosystem supporting research to clinical deployment.	SOTA benchmark in segmentation.	Ready made modern medical augmentation set.
Limitation	High learning curve.	Preprocessing and segmentation only.	Does not include model training.

2.4 Commercial Clinical AI Systems

Clinical deployment has been exemplified by Qure.ai and Optellum. Qure.ai's qCT-LN Quant automates volumetric nodule quantification and serial growth tracking, integrating into PACS and providing structured reports. Optellum's Virtual Nodule Clinic specializes in malignancy risk stratification, offering dashboards with color coded risk scores and follow up management tools[8][9].

Both systems are FDA-cleared, but their proprietary architectures prevent academic reproducibility. These products illustrate the importance of combining high algorithmic performance with interpretability and workflow usability core design principles also emphasized in our system. Table 2.4.1 provides a comparative overview of Qure.ai and Optellum Virtual Nodule Clinic, highlighting differences in clinical focus, output structure, and workflow integration.

Commercial System Comparison

Table 2.4.1: Comparison of Qure.ai and Optellum Virtual Nodule Clinic in terms of clinical functionality and workflow integration.

Criterion	**Qure.ai**	**Optellum Virtual Nodule Clinic**
Medical Use Area	Automatic findings and reports for Chest X-ray and brain CT.	Malignancy risk assessment of lung nodules.
Data Input	X-ray, CT.	CT (thorax).
Output Structure	Automatic report draft, finding annotation, heatmap.	Nodule based risk score, case management panel.
Clinical Workflow Model	Triage + automatic reporting.	Risk stratification + clinical decision support.
Interface	Report editor + visual finding annotation.	Nodule list + malignancy risk matrix + follow up plan.
Performed Task	Automatic writing of findings, decision acceleration.	Clinical probability support for nodule malignancy.
Decision Authority	Physician approved AI recommendation.	Multidisciplinary team assessment.

Most Prominent Feature	Automatic radiology report generation.	Nodule malignancy scoring and patient specific risk planning.
Strength	Operational speed and report standardization.	Risk modeling close to clinical decision.
Limitation	Not focused on nodule based malignancy.	Area of operation is lung nodules only.

3. Research Questions and Approach

Q1. Why is a 2.5D hybrid MONAI approach preferred over purely 2D or 3D CNN architectures?

2D models require significantly less GPU memory and offer faster training times, making them ideal for limited computational environments. However, they lack volumetric contextual awareness. 3D models process entire volumes, improving spatial understanding but demanding large VRAM and prolonged training.

A 2.5D hybrid approach implemented in MONAI provides the optimal balance stacking multiple consecutive CT slices as input channels to a 2D CNN. This configuration captures local 3D context while maintaining training efficiency.

Q2. Which deep learning architectures are most suitable for segmentation and classification tasks in lung nodule malignancy assessment?

In the current phase, a lightweight ResNet-18 model is adopted to establish a stable and reproducible 2.5D baseline under limited hardware resources. In subsequent phases, stronger feature extractors will be explored for comparative evaluation.

- **Segmentation (planned):** architectures such as SegResNet and U-Net will be evaluated for robust mask generation and ROI refinement.
- **Classification (planned extensions):** architectures such as DenseNet121, ResNet50, and EfficientNet variants will be investigated to improve discrimination performance and enable richer representational capacity.

- **Ensemble (optional future work):** combinations of complementary backbones may be explored to balance performance, robustness, and interpretability.

Q3. How will Explainable AI (XAI) techniques be implemented within the system?

Explainability is considered an important extension for clinical interpretability; however, XAI methods are not integrated in the current implementation. The planned integration strategy and candidate techniques are presented in Section 4.2, while an exploratory reliability analysis is reported in Appendix A.surrounding structures. At the current stage, these techniques are not implemented and are considered part of planned future work, following the integration of segmentation and ROI-guided learning.

Q4. How do hardware limitations affect model selection, training, and evaluation strategies?

Hardware constraints directly influenced model selection and experimental design. The current system prioritizes a computationally efficient 2.5D ResNet-18 baseline to validate the data pipeline and training workflow under limited GPU resources. Accordingly, training and evaluation settings were chosen to preserve stability and reproducibility while enabling iterative experimentation.

4. Proposed Contributions

This section presents the studies carried out within the scope of the LungXai project and the outputs obtained. The contributions are organized based on the current stage of development of the project, with completed implementations and planned future studies addressed under separate headings. This structure aims to systematically present the scope of the work and its development process.

4.1 Implemented Contributions (Current Phase)

4.1.1 A reproducible end-to-end 2.5D classification baseline on LIDC-IDRI

In the current phase of this study, an end-to-end pipeline has been developed that transforms raw CT images from the LIDC-IDRI dataset and their corresponding XML annotations into training-ready samples and trains a baseline malignancy classification model. The pipeline begins with loading CT series and parsing annotation metadata, followed by preprocessing, nodule-centered slice selection, and 2.5D sample construction. This baseline provides a reproducible reference infrastructure for future enhancements, including segmentation-assisted learning strategies and the integration of explainability methods.

A key component of this baseline is the malignancy label construction strategy. Since LIDC-IDRI provides malignancy ratings from multiple radiologists, the radiologist scores associated with each nodule are aggregated (e.g., by averaging available scores) to obtain a more stable supervision signal

and reduce inter-reader variability. Following common practice in the literature and consistent with Section 5.2.2, the aggregated malignancy score is mapped to binary labels as follows:

• **Benign (0):** aggregated malignancy score in the range **[1, 2]**
• **Malignant (1):** aggregated malignancy score in the range **[4, 5]**
• **Excluded:** cases with aggregated malignancy score around **3 (indeterminate)** are excluded to reduce label ambiguity and potential label noise.

This consensus-oriented labeling strategy supports more reliable model training and provides a consistent basis for subsequent comparative experiments.

4.1.2 Efficient 2.5D input representation for local volumetric context

To benefit from volumetric information while keeping computational cost under control, the LungXai system adopts a 2.5D input strategy instead of fully three-dimensional models. In this approach, the model input is constructed by combining the central image slice containing the nodule with several neighboring slices located immediately before and after it. These slices are provided to the model as a multi-channel input within a single sample.

This method enables the learning of structural variations of nodules across consecutive image slices, while avoiding the high memory consumption and long training times typically associated with fully three-dimensional models. As a result, spatial context is preserved while the training process remains efficient and accessible. Owing to these properties, the resulting baseline model is suitable for training under limited hardware resources and provides a solid foundation for subsequent development stages.

4.1.3 Implementation of the ResNet-18 basic model with a stable training workflow

A ResNet-18 architecture was used as the baseline model for malignancy prediction. The primary motivation for selecting ResNet-18 is its relatively lightweight structure, which allows for efficient and stable training. In addition, its widespread adoption in the literature enables consistent and fair comparisons with more complex models planned for later stages of the study, such as DenseNet and EfficientNet variants.

The training process includes splitting the dataset into training and validation subsets, applying reproducible preprocessing steps to ensure consistency across experiments, and monitoring performance using standardized evaluation methods. This structured setup allows different experiments to be directly comparable and supports reliable interpretation of the obtained results.

Considering the class distribution imbalance commonly observed in medical datasets, a sampling strategy was employed to reduce the model's bias toward the majority class. This approach ensures that samples from the minority class are sufficiently represented during training, thereby improving the model's generalization capability and enabling a more balanced and reliable evaluation of validation performance.

4.1.4 Preliminary evaluation and metric analysis with a focus on ROC-AUC

Preliminary experiments were conducted to evaluate the performance of the baseline model during training. Model performance on the validation set was monitored using both accuracy and ROC-AUC metrics. The results indicate that these two metrics do not always exhibit a direct correspondence. In particular, the training stage at which the highest ROC-AUC value is achieved does not necessarily coincide with the stage yielding the highest accuracy.

This observation highlights the importance of threshold-independent evaluation metrics in medical classification problems, where class distribution is often imbalanced. Accordingly, ROC-AUC is considered a more informative metric for assessing model performance in this context, and it is therefore adopted as the primary criterion for model selection and comparison in subsequent stages of the project.

4.1.5 CSV-based auxiliary experimental evaluation using an alternative algorithmic pipeline

In addition to the primary image-based deep learning baseline, an auxiliary experimental analysis was conducted using structured information derived from the LIDC-IDRI dataset in a CSV-based format. This analysis utilized tabular data containing radiologist-assigned malignancy scores and selected descriptive attributes associated with individual nodules. Such a representation enables an independent examination of the data by aggregating assessments from multiple radiologists for the same nodule alongside corresponding nodule characteristics.

Unlike the image-based deep learning pipeline, this auxiliary study employed an alternative algorithmic approach operating solely on structured data. This design allowed for rapid exploration of the discriminative properties of the available features and facilitated the evaluation of different experimental scenarios without the computational overhead associated with training deep neural networks. As a result, the CSV-based pipeline served as a lightweight and efficient exploratory tool, particularly suited for early-stage analysis and comparative investigations.

The primary purpose of this auxiliary analysis was not to replace the main image-based learning framework, but to provide complementary insights. Through these experiments, the consistency of the dataset was examined, the distribution of radiologist assessments was analyzed, and the relevance of selected features for malignancy discrimination was investigated. The observed patterns were consistent with trends identified in the image-based experiments, providing supportive evidence for the overall learning strategy.

Overall, this auxiliary experimental pipeline contributed additional perspective on the dataset characteristics and informed subsequent design considerations, such as feature selection and experimental setup. By incorporating both image-based and structured-data analyses, the study adopts a broader experimental viewpoint without relying on a single data representation or modeling paradigm.

4.1.6 Hyperparameter configuration and comparative reporting support

To ensure that the experimental studies are conducted in a reliable and comparable manner, the core hyperparameters used during model training were defined in a systematic way and carefully documented. In this context, parameters such as the number of epochs, batch size, and model architecture settings were specified at the beginning of the experimental process and applied consistently across all experiments. This approach ensured that variations in the results were primarily attributable to differences in model architecture or data representation, rather than to random configuration changes.

The defined hyperparameter configurations were shared within the project team and used to generate key performance metrics, including accuracy and ROC-AUC, for multiple model setups. This strategy enabled direct comparison of different models trained on the same dataset under similar training conditions. In particular, during the early stages of experimentation, the effects of varying epoch numbers and training durations on model behavior were observed, and these observations informed the design of subsequent experiments.

This systematic approach to configuration and reporting enhances the transparency and reproducibility of the experimental results presented in this report. Moreover, it allows the existing experimental framework to be extended in a consistent manner when more complex architectures or additional modules—such as segmentation or explainability components—are introduced in later stages. In this respect, the controlled handling of hyperparameters represents an important contribution to the experimental robustness and maturity of the study.

4.1.7 Partial implementation of the user interface (UI) layer

To support interaction with the classification pipeline and improve the interpretability of model outputs, a user interface (UI) layer has been partially implemented as part of the project. At the current stage, the interface functions as a prototype and is intended to demonstrate the overall system workflow rather than to serve as a clinical application.

The UI enables basic user interaction, including input selection, model execution, and visualization of classification results. By presenting model predictions in a visual format alongside numerical outputs, the interface facilitates qualitative inspection of the system’s behavior.

Although limited in scope, the current UI establishes a basic interaction layer that can be extended in future stages to incorporate additional outputs such as segmentation results and explainability visualizations. This component highlights the project’s consideration of not only model development but also the presentation and interpretability of results.

5. Methodology and System Approach

5.1 Distinctive Aspects of the Study

This study adopts a progressive development strategy for the problem of lung nodule classification. While many studies in the literature present complex systems in which segmentation, explainable artificial intelligence (XAI), and deployment-oriented components are integrated simultaneously, this project prioritizes the establishment of a reliable and interpretable baseline classification framework.

At the current stage, the developed system consists of the following components:

- 2.5D slice-based input representation
- ResNet-18–based classification model
- A user interface (UI) for presenting model outputs

At this stage, SegResNet-based segmentation, explainability methods such as Grad-CAM, DenseNet architectures, and deployment or optimization solutions have not yet been integrated into the system. These components are planned to be incorporated in later phases of the project.

The main objectives of this approach are:

- To evaluate the contribution of each module independently,
- To prevent misleading performance gains caused by prematurely coupled complex architectures,
- To establish a reproducible and analyzable **baseline system**.

Current Phase:

- ResNet-18–based classification
- 2.5D slice stacking
- classification pipeline without segmentation
- User interface–assisted inference

Advanced components such as segmentation-based ROI extraction, explainability methods, and deployment-oriented optimizations are outside the scope of the current methodology and are discussed exclusively in Section 11.

5.2 Dataset Description: LIDC-IDRI

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset is one of the most comprehensive publicly available thoracic CT datasets used for pulmonary nodule analysis, malignancy classification, and explainable medical AI research. The dataset consists of

1,018 patient studies, each containing a full chest CT scan stored in DICOM format, along with XML-based annotations provided independently by four experienced thoracic radiologists.

Each study includes:

- A complete axial CT volume (typically consisting of tens to several hundreds of slices, depending on acquisition protocol),
- A corresponding XML annotation file describing radiologist findings,
- Both individual and consensus-level annotations.

Variability in acquisition parameters such as slice thickness and spacing reflects real-world clinical heterogeneity and increases the dataset's suitability for evaluating robust AI systems.

5.2.1 Nodule Annotations

Radiologists independently annotated each CT volume following a standardized two-phase protocol. For each annotated nodule, the following information is provided:

- Slice-wise nodule contours,
- Three-dimensional centroid coordinates (x, y, z),
- A malignancy score on a five-point ordinal scale (1–5),
- Additional semantic characteristics such as texture, margin, and sphericity.

Annotated lesions are categorized into:

- Nodules ≥ 3 mm, which are considered radiologically significant,
- Nodules < 3 mm, which are commonly excluded from deep learning–based malignancy classification tasks.

In this study, only nodules ≥ 3 mm are included to ensure sufficient visual and structural information for learning.

5.2.2 Malignancy Label Definition

The malignancy score provided in the LIDC-IDRI dataset ranges from 1 (highly unlikely malignant) to 5 (highly suspicious for malignancy). Following common practice in the literature, labels are interpreted as:

- **Scores 1–2:** Benign

- **Score 3:** Indeterminate
- **Scores 4–5:** Malignant

Since a malignancy score of 3 represents diagnostic uncertainty, these samples are excluded in this study to reduce label noise and improve the reliability of binary classification. After excluding indeterminate cases, the remaining dataset typically exhibits a moderately imbalanced benign-to-malignant distribution, which is addressed during training through appropriate sampling strategies.

5.2.3 Number of Nodules and Sample Definition

Across the 1,018 patient studies, the dataset contains 2,669 annotated nodules ≥ 3 mm. Each nodule spans multiple axial slices along the z-axis, and patients may contain multiple nodules (typically 1–4 per scan).

In this project, each nodule contributes one training instance centered on the nodule location. Rather than treating individual slices independently, the sample is constructed using a 2.5D representation, as described in the following section.

5.2.4 Slice and Patch Generation for This Study

To adapt the LIDC-IDRI dataset to the proposed 2.5D classification framework, the following strategy is employed:

1. The centroid axial slice corresponding to the annotated nodule is identified using XML metadata.
2. A fixed number of neighboring slices above and below the centroid are selected.
3. These slices are stacked channel-wise to form a 2.5D input tensor.
4. Hounsfield Unit (HU) normalization is applied using a lung-relevant window (e.g., −1000 to 400), followed by scaling to the [0, 1] range.
5. An optional ROI-centered crop (e.g., 64×64 or 128×128 pixels) is applied to reduce background influence.

This process yields one 2.5D training sample per nodule, with binary labels:

- benign (0),
- malignant (1).

After preprocessing and exclusions, the working dataset typically consists of approximately 1,500–1,700 samples, depending on filtering criteria. This representation is fully compatible with MONAI-based 2.5D classification pipelines.

5.2.5 Rationale for Using LIDC-IDRI

The LIDC-IDRI dataset is selected due to:

- Its rich multi-reader radiologist annotations,
- Public availability and strong reproducibility,
- Extensive use as a benchmark dataset in lung cancer AI literature,
- Compatibility with 2D, 2.5D, and 3D deep learning workflows,
- Suitability for explainability (XAI) studies.

For these reasons, LIDC-IDRI is widely regarded as a de facto standard dataset for pulmonary nodule classification research.

5.3 Preprocessing Steps

Before being fed into the model, raw DICOM data undergo a series of standardized preprocessing steps designed to ensure data consistency and stabilize the training process. First, DICOM series are loaded to reconstruct the three-dimensional CT volume, and voxel intensities are converted to Hounsfield Units (HU). Intensity normalization is then applied using a lung-relevant window, followed by scaling to a normalized range for numerical stability.

Nodule centroid coordinates are extracted from the corresponding XML annotation files, and the axial slice containing the nodule center is identified. To incorporate limited volumetric context, a fixed number of neighboring slices above and below the centroid are selected and stacked channel-wise to construct a 2.5D input representation. Optionally, a nodule-centered region of interest (ROI) crop is applied to reduce background influence. These preprocessing steps ensure consistent input formation and provide stable, training-ready samples for the classification model.

5.4 Processing Pipeline

5.4.1 Implemented Pipeline (Current System)

The processing pipeline implemented in the current system is as follows:

DICOM + XML DATA

- Preprocessing and normalization
- Nodule-centered slice selection
- 2.5D slice construction
- ResNet-18 classification
- Performance evaluation (ROC-AUC, accuracy)

At this stage, the system does not include segmentation or ROI masking and focuses directly on the classification task.

5.5 Algorithms Used in the Current System

This section summarizes the algorithms and methodological components implemented in the current phase of the LungXai system. The table provides a concise overview of the system's modular structure as realized in the present baseline implementation. As shown in **Table 5.5.1**, only components that are actively implemented and evaluated in the current study are included. Components that are planned for future development are intentionally excluded from this section and are discussed separately in Section 11.

Table 5.5.1: Algorithms and system components implemented in the current LungXai baseline.

Component	**Current Implementation**
Input Representation	2.5D slice stacking
Classification Model	ResNet-18
Segmentation	-
Explainability	-
Deployment / Optimization	-

5.6 Explainable Artificial Intelligence (XAI) Concept

This section presents explainable AI concepts at a conceptual level and does not describe components implemented in the current system.

In medical image analysis, it is not sufficient for a model to produce accurate predictions alone. It is also essential to understand which image regions the model relies on when making decisions. This requirement highlights the importance of explainable artificial intelligence (XAI) approaches.

XAI methods aim to increase clinical trust by visualizing the internal decision-making mechanisms of deep learning models.

5.6.1 Example XAI Approach: Attention and Grad-CAM

Attention mechanisms guide the model's focus toward spatially relevant regions by adaptively weighting intermediate feature maps and suppressing background information that is less informative for the target task. As shown in Figure 5.6.1.1, attention gates can be integrated into convolutional neural network architectures to emphasize anatomically meaningful regions, such as pulmonary nodules, while reducing the influence of surrounding tissues. This mechanism allows the network to selectively propagate features that are more relevant to the classification decision.

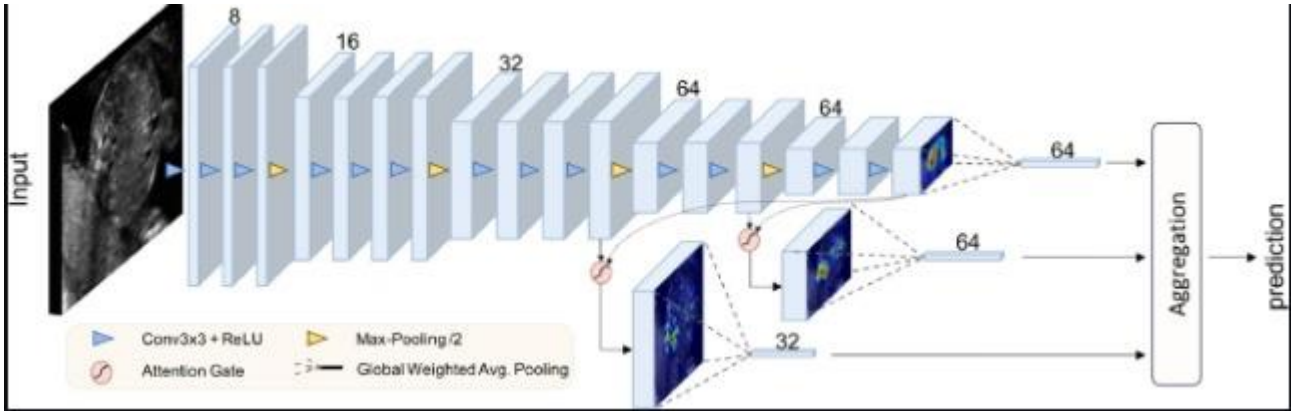

Figure 5.6.1.1: Example architecture of an attention-gated convolutional neural network illustrating multi-scale feature extraction and attention-based aggregation for classification.

Grad-CAM complements this approach by utilizing gradients flowing into the final convolutional layers to generate class-specific activation maps. As illustrated in Figure 5.6.1.2, the gradients are combined with feature maps to produce a heatmap that highlights image regions contributing most strongly to the predicted class. When applied to medical images, such visualizations make it possible to assess whether the model's decision is primarily driven by the lung nodule itself or by irrelevant background structures.

Together, attention mechanisms and Grad-CAM provide an intuitive framework for interpreting deep learning models in medical imaging. Although these approaches are not implemented in the current system, they are presented here to conceptually illustrate how explainability can be incorporated in future stages of the project, particularly after the integration of segmentation and ROI-based learning.

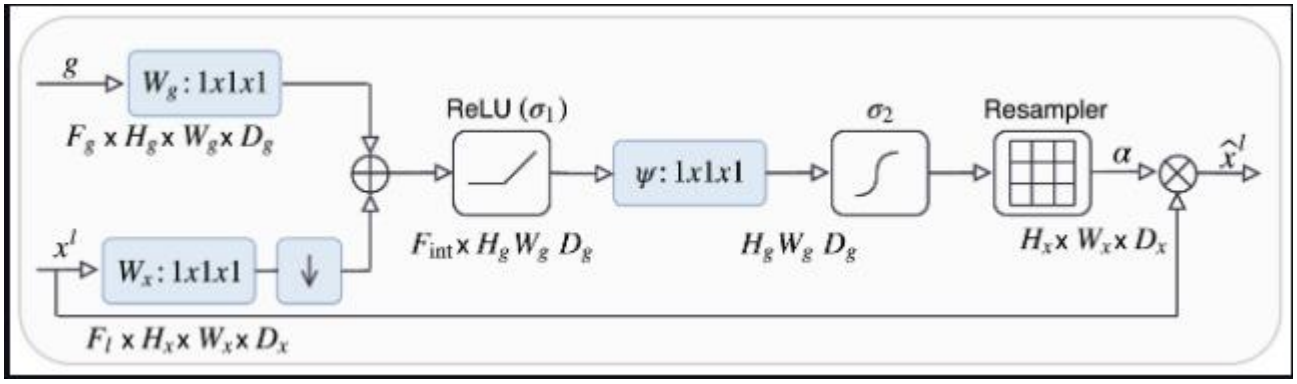

Figure 5.6.1.2: Detailed structure of an attention gate illustrating the combination of gating and input feature maps through 1×1×1 convolutions, non-linear activations, and resampling to generate attention coefficients for feature refinement.

5.6.2 Why XAI Has Not Yet Been Implemented in This Study

The absence of XAI modules in the current system is a deliberate design choice. At the present stage:

- No segmentation infrastructure is available,
- ROI-guided learning has not been implemented,
- The necessary conditions for producing clinically reliable XAI outputs have not yet been established.

Therefore, XAI integration is planned after the segmentation module is completed, enabling more meaningful and interpretable explanations. An exploratory Grad-CAM–based reliability assessment supporting this design decisionis provided in Appendix A.

5.7 Training Process and Epoch-Based Results

During training, validation accuracy and ROC-AUC values were computed at the end of each epoch. Table 5.7.1 reports the validation accuracy and ROC-AUC values obtained at the end of each training epoch.

Table 5.7.1: Validation accuracy and ROC-AUC values across training epochs.

Epoch	Validation Accuracy	Validation AUC
1	0.489	0.674
2	0.529	0.639
3	0.591	0.640
4	0.657	0.600
5	0.623	0.609

The highest validation AUC, approximately 0.674, was achieved at epoch 1.

This observation demonstrates that accuracy may increase while AUC decreases, highlighting that AUC is a more reliable metric for imbalanced datasets.

6 . System Architecture and Implementation

This section presents the end-to-end system architecture and implementation approach of the LungXai project. Rather than focusing on the theoretical aspects of the employed deep learning models, this section aims to describe how the system is structured, how its components interact, and how the overall workflow is managed from raw medical data to final outputs. The proposed architecture demonstrates that the project is not merely a model-centric experiment, but an engineering-oriented research framework designed with reproducibility and extensibility in mind.

6.1 Overall System Approach

LungXai is designed as a multi-stage decision support system for lung nodule analysis using chest computed tomography (CT) images. Instead of directly feeding raw data into the learning model, the system follows a clearly defined and controlled processing pipeline. This design choice improves experimental consistency and reduces uncertainties inherent in clinical imaging data.

Conceptually, the system is composed of logically separated layers responsible for data acquisition, preprocessing, representation learning, modeling, evaluation, and output generation. Each layer is designed with a specific responsibility and interacts with other layers in a loosely coupled manner, enabling modular development and future scalability.

6.2 Input Data and Preprocessing Layer

In the first stage of the system, thoracic CT images are ingested in DICOM format. Since raw pixel intensities vary across scanners and acquisition protocols, these values are not directly suitable for learning-based analysis. Therefore, image intensities are converted into Hounsfield Units (HU), which represent a standardized scale commonly used in medical imaging.

The primary objectives of the preprocessing stage are:

- Standardization of image intensity values,
- Preservation of slice ordering and spatial consistency,
- Construction of a stable and reliable representation suitable for model input.

This stage plays a critical role in the overall architecture, as the quality and consistency of preprocessing directly influence the reliability of the learned features.

6.3 Annotation-Guided Data Representation and 2.5D Input Construction

Preprocessed images are aligned with expert-provided annotation data to construct supervised learning samples. The annotations include information regarding the spatial location of lung nodules and their malignancy assessments.

Instead of employing a fully volumetric (3D) learning approach, this study adopts a 2.5D representation strategy. For each annotated nodule:

- A central slice is identified,
- Adjacent slices above and below the central slice are selected,
- The selected slices are stacked along the channel dimension to form a single input sample.

This approach enables the model to leverage limited volumetric context while maintaining computational efficiency. Compared to purely 2D methods, the 2.5D strategy captures richer contextual information, while remaining more efficient than full 3D models.

6.4 Modeling and Learning Layer

The constructed 2.5D inputs are processed by a convolutional neural network–based classification model. The primary task of the model is to estimate the probability that a given input corresponds to a benign or malignant lung nodule.

The modeling layer encompasses both feature extraction and decision-making processes. It is treated as an independent architectural component, allowing future replacement or extension of the backbone architecture without altering the surrounding system pipeline.

6.5 Training Process and Experiment Management

The training process is managed through a structured control mechanism decoupled from the data processing pipeline. During training:

- Samples are fed to the model in a controlled manner,
- Loss values are computed and optimized through backpropagation,
- Performance metrics are monitored throughout the training process.

Separating experimental configuration from implementation logic allows systematic comparison of different training setups. This approach improves reproducibility and enhances the reliability of experimental findings.

6.6. Resnet-18-Based 2.5D Baseline: Method and Training

This section describes the 2.5D ResNet-18–based baseline model employed in the LungXai framework and the associated training methodology. The primary objective of this baseline is to provide a reliable and computationally efficient reference model for lung nodule classification, enabling systematic evaluation of future architectural and methodological improvements.

6.6.1 Motivation for the 2.5D Approach

Pulmonary nodules are inherently three-dimensional structures that typically span multiple adjacent CT slices. Purely two-dimensional approaches, which analyze individual slices in isolation, fail to capture the spatial continuity and morphological variation of nodules across slices. On the other hand, fully three-dimensional convolutional models, while more expressive, impose significantly higher computational and memory demands and often require substantially larger datasets.

The 2.5D approach adopted in this study aims to strike a balance between these two extremes. By incorporating a limited number of neighboring slices around a central slice, the model gains access to localized volumetric context while maintaining the efficiency and simplicity of two-dimensional convolutional operations. This design choice is particularly well suited to datasets where annotations are primarily slice-based and data heterogeneity is high.

6.6.2 Construction of the 2.5D Input Representation

For each annotated lung nodule, a central axial slice is identified based on expert-provided annotations. To enrich the spatial context, a fixed number of slices immediately above and below the central slice are selected. These slices are combined into a single input sample by stacking them along the channel dimension.

This representation allows the model to simultaneously observe the nodule's appearance at its most representative cross-section and its structural evolution across adjacent slices. As a result, the model can learn features that are more robust to slice-level variability while avoiding the complexity of full volumetric processing.

6.6.3 Backbone Selection: ResNet-18

ResNet-18 is selected as the backbone architecture for the baseline model due to its well-established performance, training stability, and favorable trade-off between depth and computational cost. The residual connections within the architecture facilitate effective gradient propagation, enabling stable optimization even when training data are limited.

Additional motivations for selecting ResNet-18 include its widespread adoption in medical imaging research and its suitability as a reference model for comparative studies. Using a well-known architecture also improves the interpretability and reproducibility of experimental results.

6.6.4 Architectural Adaptation for 2.5D Inputs

The adoption of a 2.5D representation increases the number of input channels relative to standard image classification settings. To accommodate this change, the network's input stage is adapted to accept multi-channel inputs corresponding to the selected slices. Importantly, this modification is restricted to the input level, while the remainder of the network architecture is preserved.

By limiting architectural changes, the baseline ensures that observed performance differences can be attributed primarily to the input representation rather than to extensive redesign of the network structure. This design also facilitates future extensions, such as the addition of task-specific output branches.

6.6.5 Learning Objective and Training Loss

The primary learning objective of the baseline model is binary classification, distinguishing between benign and malignant lung nodules. During training, the model is optimized to minimize the discrepancy between predicted class probabilities and ground-truth labels.

This formulation encourages the network to learn discriminative features that capture subtle textural and morphological differences between benign and malignant nodules, which are often difficult to distinguish visually.

6.6.6 Handling Class Imbalance

Lung nodule datasets commonly exhibit class imbalance, with one class occurring more frequently than the other. If not addressed, this imbalance can bias the model toward the majority class, reducing sensitivity to clinically important minority cases.

To mitigate this issue, a sampling-based balancing strategy is employed during training. By adjusting the sampling frequency of training examples, the model is exposed to a more balanced distribution of classes across training iterations. This approach promotes more stable learning dynamics and helps the model develop decision boundaries that better reflect both classes.

6.6.7 Optimization and Training Procedure

Training is conducted using a mini-batch optimization framework, which allows efficient parameter updates and stable convergence. Model performance is monitored throughout training using both loss values and validation metrics to detect potential overfitting or underfitting.

Hyperparameters such as learning rate, batch size, and training duration are managed externally to support reproducibility and controlled experimentation. This setup enables systematic comparison of different training configurations without altering the core implementation.

6.6.8 Evaluation Metric: Emphasis on ROC-AUC

Although classification accuracy is reported for completeness, ROC-AUC is selected as the primary evaluation metric. In medical classification tasks, accuracy alone can be misleading, particularly in the presence of class imbalance.

ROC-AUC provides a threshold-independent measure of separability between classes and captures the trade-off between sensitivity and specificity across all operating points. Consequently, model selection and performance comparisons in this study are primarily based on validation ROC-AUC values.

6.6.9 Role of the Baseline within the LungXai Framework

The 2.5D ResNet-18 model serves as a **baseline reference** within the LungXai system. Its role is threefold:

1. To verify the correctness and stability of the end-to-end pipeline,
2. To establish a quantitative performance reference,
3. To provide a foundation for evaluating future enhancements such as segmentation-guided classification, multitask learning, and spatial localization.

All subsequent experimental improvements are evaluated relative to this baseline configuration.

6.7 Database Schema

The Lung Nodule Detection system utilizes a relational database structure consisting of seven main tables designed to support clinical workflows, AI-based analysis, report generation, and system-level auditing. The overall structure of the database and the relationships between its core entities are illustrated in Figure 6.7.1.

Tables

Users

The users table stores system users. Each user has a unique identifier, username, password, first name, last name, and email address. Users are assigned either Admin or Doctor roles. Professional information such as specialization, department, hospital, and license number is also maintained to support role-based access control.

Patients

The patients table contains patient records. Each patient has a unique patient_id, name, age, and gender. When a patient is registered in the system, a creation timestamp is automatically assigned.

Studies

The studies table represents CT scan studies. Each study is linked to a patient through the patient_id field, which references the patients table. Study date, description, clinical notes, and detected nodule count are stored. The study status (pending, analyzing, completed, reviewed) is maintained in the status field. When a doctor reviews a study, the reviewed_by field references the reviewing physician's identifier in the users table.

Dicom_Files

The dicom_files table stores uploaded medical imaging files. Each file belongs to a study and is linked through the study_id field to the studies table. File path, file name, and DICOM instance number are stored.

Nodules

The nodules table contains lung nodules detected by the AI model. Each nodule belongs to a study and stores anatomical location, size (in millimeters), malignancy risk level (Low, Medium, High), three-dimensional coordinates, and CT slice index. The AI confidence score is stored in the probability field. Physician assessment and notes are recorded in the doctor_assessment and notes fields. The include_in_report flag determines whether the nodule is included in the final report.

Reports

The reports table stores generated reports. Each report is linked to a study and a patient through the study_id and patient_id fields. The user who generated the report is referenced via the generated_by_id field. Total nodule count and included nodule count are stored separately. Report content is maintained in structured JSON format in the report_data field.

Activity_Logs

The activity_logs table records all user actions performed within the system. Each record references the initiating user via the user_id field and stores action type (e.g., LOGIN, UPLOAD, ANALYZE, REPORT) along with additional details for auditing purposes.

Relationships and Data Integrity

The database schema contains eight foreign key relationships. The studies table is linked to the patients table through patient_id and to the users table through reviewed_by. The dicom_files and nodules tables reference the studies table via study_id. The reports table establishes three relationships, linking to studies, patients, and users through study_id, patient_id, and generated_by_id, respectively. The activity_logs table references the users table through user_id.

Cascade delete rules are implemented to preserve data consistency. When a patient is deleted, all associated studies, DICOM files, nodules, and reports are automatically removed. When a user is deleted, related references are set to NULL to prevent unintended data loss while preserving historical records.

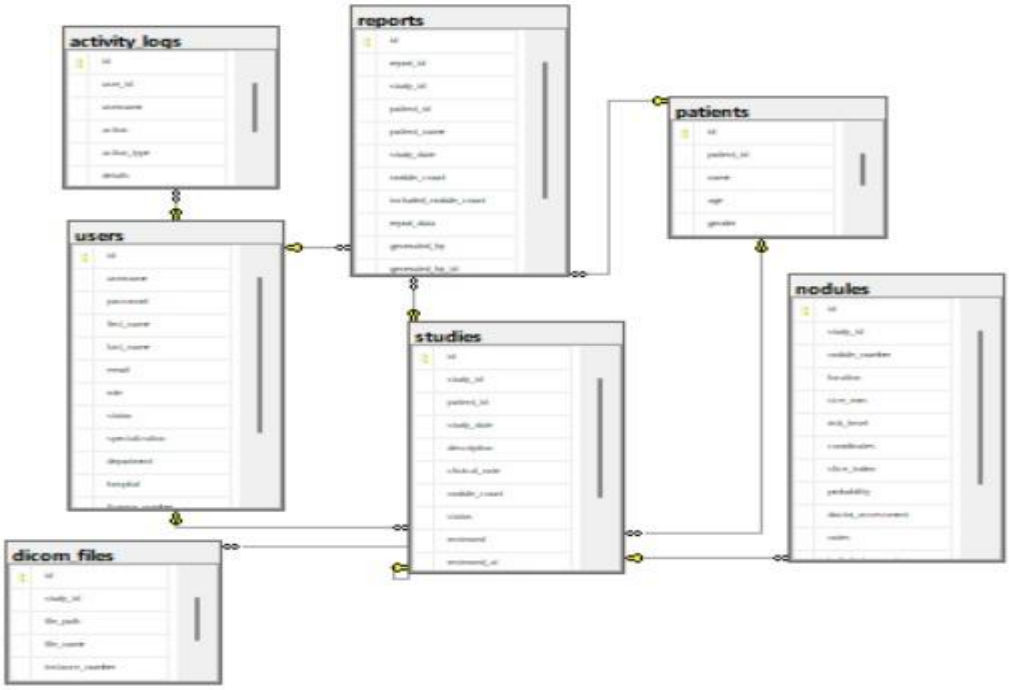

Figure 6.7.1: Relational database schema of the LungXai system.

7. Experimental Results and Analysis

This section presents the experimental results obtained with the proposed 2.5D ResNet-18 baseline and provides an in-depth analysis of model behavior during training and validation. The goal is to assess classification performance using clinically meaningful metrics, justify model selection criteria, and identify potential sources of error that inform future improvements.

7.1 Experimental Setup Overview

All experiments were conducted using the standardized pipeline described in the previous sections. The dataset was split into training and validation subsets with patient-level separation to prevent

information leakage. During training, model checkpoints and validation metrics were recorded at each epoch to enable systematic comparison across training stages.

To ensure a fair assessment in the presence of class imbalance, evaluation focused on threshold-independent metrics in addition to conventional accuracy. Across the conducted experiments, the highest validation ROC-AUC (≈ 0.674) was achieved at the first training epoch, while the corresponding validation accuracy was approximately 0.489. These values indicate that the baseline model is able to discriminate benign and malignant nodules above random chance, despite limited accuracy at the default decision threshold.

7.2 Training Dynamics and Epoch-wise Performance

Model performance was monitored throughout training by tracking training loss, validation accuracy, and validation ROC-AUC at each epoch. Rather than relying solely on loss convergence, particular attention was given to validation metrics to assess generalization behavior.

The results indicate that while validation accuracy continued to improve or plateau in later epochs, validation ROC-AUC peaked at an earlier stage. This observation highlights the importance of metric selection when determining the optimal stopping point.

7.3 Best Epoch Selection Criterion

The best-performing epoch was selected based on the highest validation ROC-AUC rather than maximum accuracy. This choice is motivated by the clinical relevance of ROC-AUC, which reflects the model's ability to distinguish malignant from benign nodules across all classification thresholds.

Selecting the model based on accuracy alone could favor a decision boundary that performs well at a single threshold but generalizes poorly across different sensitivity–specificity trade-offs. Therefore, ROC-AUC–based selection provides a more robust criterion for medical decision support scenarios.

7.4 ROC-AUC versus Accuracy: Comparative Interpretation

Accuracy and ROC-AUC capture fundamentally different aspects of model performance. Accuracy measures the proportion of correctly classified samples at a fixed decision threshold, whereas ROC-AUC evaluates the model's ranking ability independent of threshold choice.

In the conducted experiments, cases were observed where:

- Accuracy remained stable or slightly improved across epochs,
- ROC-AUC decreased after reaching a peak.

This divergence suggests that while the model continued to correctly classify a similar number of samples at the default threshold, its overall ability to separate classes deteriorated. Such behavior is particularly common in imbalanced datasets and reinforces the necessity of ROC-AUC as the primary evaluation metric.

7.5 Error Analysis and Performance Limitations

Despite achieving meaningful classification performance, several factors may contribute to residual errors:

1. **Limited Volumetric Context**
   The 2.5D representation captures only a small neighborhood of slices and may fail to represent nodules with elongated or irregular 3D structures.
2. **Annotation Variability**
   Malignancy labels are derived from expert assessments, which may exhibit inter-observer variability. This introduces label noise that can limit achievable performance.
3. **Background Influence**
   In the absence of explicit segmentation, surrounding lung tissue and non-nodular structures may introduce irrelevant features, potentially distracting the classifier.
4. **Exclusion of Indeterminate Cases**
   Nodules with intermediate malignancy scores are excluded to reduce label ambiguity, but this also reduces dataset diversity and may limit generalization.
5. **Class Imbalance Effects**
   Although sampling strategies mitigate imbalance during training, residual effects may persist, particularly in borderline cases near the decision boundary.

7.6 Discussion of Observed Results

Overall, the experimental results demonstrate that the 2.5D ResNet-18 baseline provides a stable and interpretable performance level suitable for use as a reference model. The observed gap between ROC-AUC and accuracy underscores the importance of selecting evaluation metrics aligned with clinical priorities.

The analysis further suggests that incorporating additional spatial constraints—such as segmentation-guided region-of-interest extraction or explicit localization objectives—could substantially improve robustness and interpretability.

8. User Interface (UI) Module

This section describes the User Interface (UI) module developed as part of the LungXai system. The UI serves as an interaction layer between the trained deep learning models and end users, enabling model outputs to be accessed, interpreted, and reviewed in a structured and user-friendly manner. Rather than targeting direct clinical deployment, the interface is designed as a demonstration-oriented and research-focused visualization tool.

8.1 Purpose and Design Goals

The primary objective of the UI module is to demonstrate how the proposed lung nodule classification system can be integrated into a practical software environment. The interface is designed to:

- Provide a clear workflow for reviewing imaging-based analysis results
- Present model predictions in an interpretable and accessible format
- Support inspection of multiple cases and system configurations
- Separate presentation logic from model inference logic

The design prioritizes clarity, modularity, and extensibility, allowing future integration of additional outputs such as segmentation masks or localization results.

8.2 UI Architecture Overview

The UI is implemented as a modular, component-based web application. Its architecture follows a clear separation of concerns:

- Layout Components:

A persistent layout structure organizes the interface into navigation and content regions. This includes a top navigation bar and a side navigation panel, enabling consistent access to system pages.

- Page-Level Views:
  Individual pages correspond to distinct system functionalities, such as work list management, case review, system settings, and user management.
- Service Layer:
  A dedicated communication layer handles interactions with backend services or APIs, abstracting data retrieval and submission from the presentation layer.
- Utility Layer:
  Supporting utilities manage domain-specific operations such as DICOM-related handling and data formatting.

This layered UI architecture ensures maintainability and allows interface elements to evolve independently from backend or model-level changes.

8.3 User Workflow and Interaction Scenario

A typical user interaction with the LungXai UI follows the workflow outlined below:

1-Case Selection:

The user accesses a work list containing available imaging cases. Each entry represents a study or sample ready for review. As shown in Figure 8.3.1, users are required to authenticate through a dedicated login interface before accessing any system functionality.

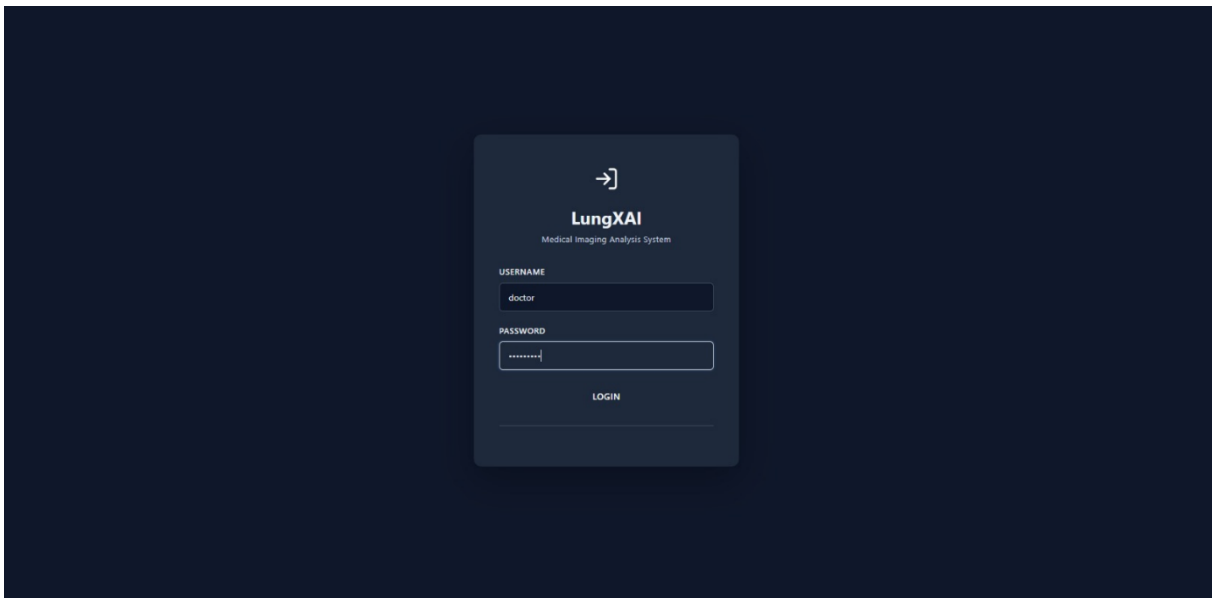

Figure 8.3.1: Login interface of the LungXai system, providing role-based authentication for authorized users.

2-Case Review:

Upon selecting a case, the user is navigated to a review interface where imaging data and associated metadata are displayed. After authentication, the user initiates a new study by entering patient information and uploading DICOM files, as shown in Figure 8.3.2.

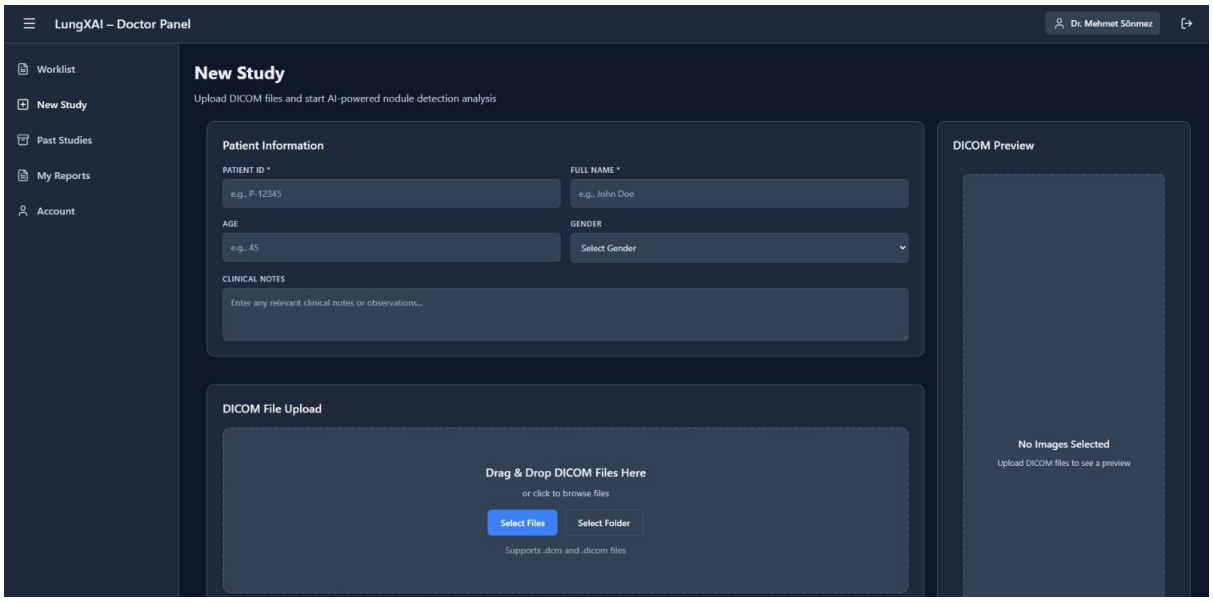

Figure 8.3.2: Interface for creating a new study, including patient information entry and DICOM file upload for AI-based lung nodule analysis.

3-Model Output Visualization:

The UI presents the model's classification output, including predicted class labels and confidence-related information.

4-System Navigation:

The user may navigate between different sections of the interface, such as system configuration panels or administrative pages, without interrupting the review workflow.

This scenario reflects a realistic usage pattern for decision support systems in research or clinical validation settings.

8.4 Presentation of Model Outputs

Model predictions are displayed in a structured and readable manner. Rather than exposing raw numerical outputs, the UI emphasizes interpretability by:

- Clearly indicating the predicted class (e.g., benign or malignant)
- Associating predictions with confidence or probability values
- Presenting results alongside the corresponding imaging context

This design ensures that outputs are not interpreted in isolation but are visually and contextually linked to the underlying CT data.

As shown in Figure 8.4.1, model predictions are presented together with the corresponding CT slice, allowing clinicians to interpret AI outputs in direct anatomical context.

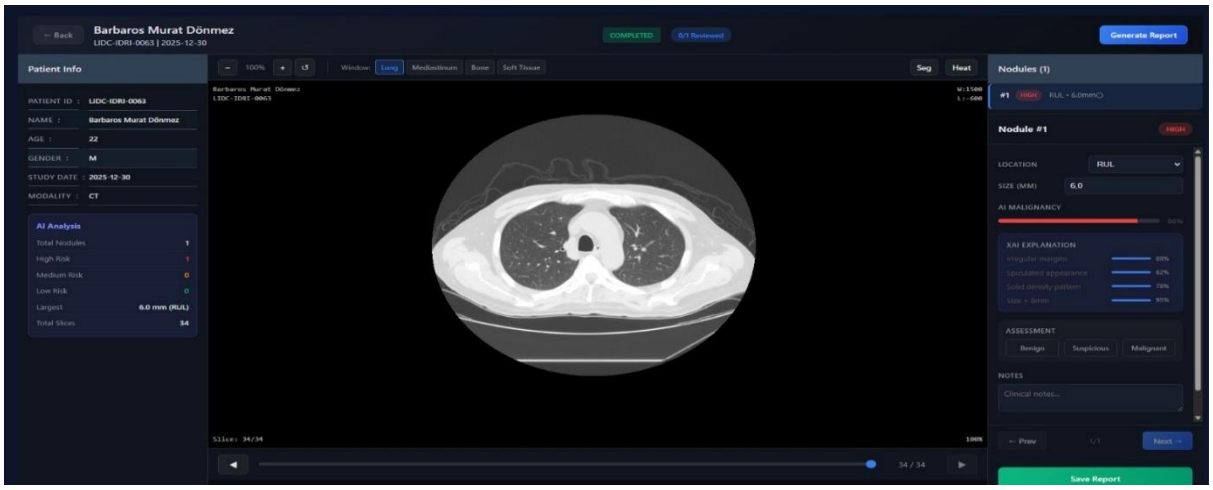

Figure 8.4.1: Visualization of model predictions displayed alongside the corresponding CT slice, including malignancy assessment and confidence indicators.

8.5 Navigation and Interface Components

The interface includes a consistent navigation system that allows users to move efficiently between different functional areas. Key UI components include:

- A sidebar navigation panel for accessing major system modules
- A top bar for contextual actions and system-level information
- Reusable layout elements that ensure visual consistency across pages

Such component reuse simplifies maintenance and provides a cohesive user experience.

Figure 8.5.1 shows the administrative dashboard, which supports system monitoring, user management, and navigation across non-clinical functionalities.

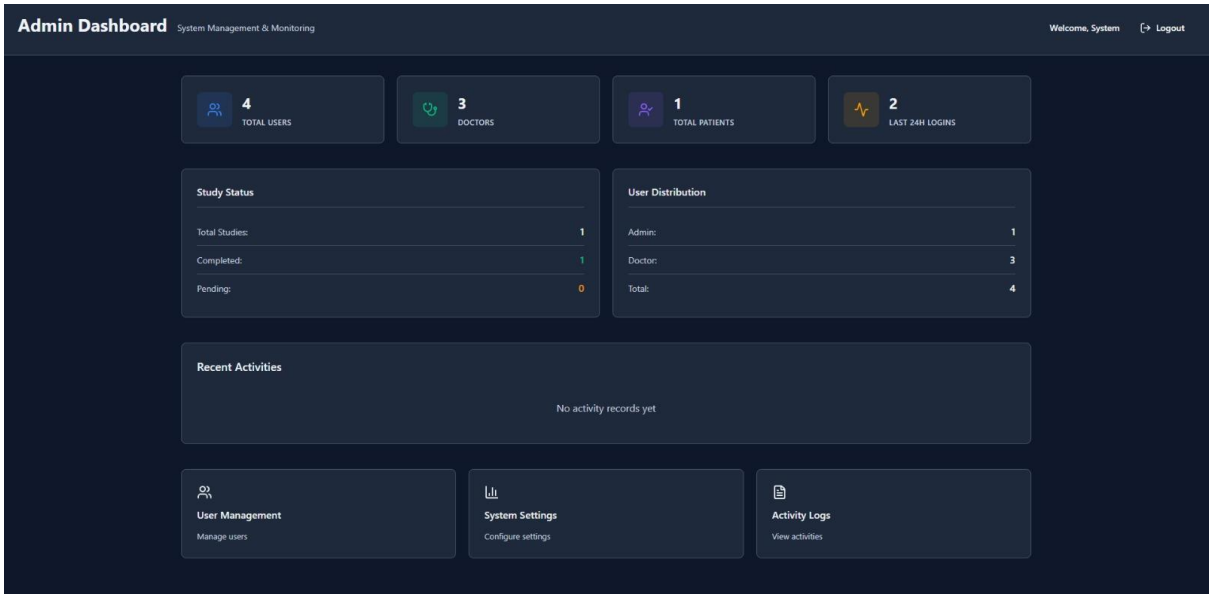

Figure 8.5.1: Administrative dashboard providing system monitoring, user management, and navigation across non-clinical functionalities.

8.6 Role of the UI within the LungXai System

Within the overall LungXai architecture, the UI module functions as a visualization and interaction layer rather than a core analytical component. Its role is to:

- Demonstrate end-to-end integration of data, models, and outputs
- Facilitate qualitative inspection of model behavior
- Support future extensions such as visualization of segmentation masks or localized predictions

By decoupling the UI from model internals, the system remains flexible and adaptable to evolving research objectives.

8.7 Limitations and Future Extensions

The current UI focuses on classification output presentation and system navigation. Planned future enhancements include:

- Visualization of segmentation results overlaid on CT images
- Display of predicted nodule locations (e.g., coordinate-based markers)

- Enhanced case comparison and filtering tools
- Integration of explainability outputs such as heatmaps

These extensions would further improve interpretability and usability, particularly for expert users.

8.8 Use Case Diagram

The use case diagram summarizes the interaction between different user roles and the main functional components of the LungXai user interface. The diagram is designed to present who can perform which actions within the system, rather than illustrating the execution order of operations.

8.8.1 User Roles (Actors)

The LungXai system defines two primary user roles:

- **Doctor**
  Represents clinical users who interact with medical imaging data, initiate analysis sessions, review CT images, and interpret AI-generated malignancy predictions.
- **Admin**
  Represents system administrators responsible for managing users, configuring system settings, and monitoring overall system activity and performance.

These roles are intentionally separated to ensure a clear distinction between clinical decision-making workflows and administrative system management.

8.8.2 Common Use Case: Authentication

Authentication is a mandatory prerequisite for accessing any system functionality. Both Doctor and Admin actors must successfully authenticate through the Login use case before interacting with protected features.

In the use case diagram, authentication is modeled as a shared mandatory behavior using <<include>> relationships. This modeling choice reflects role-based access control (RBAC) rather than execution order, ensuring that system functionalities are available only to authorized users.

8.8.3 Doctor Use Cases

- **Create New Study:** Allows the doctor to initiate a new analysis session by entering patient information and preparing the case for evaluation.
- **Upload DICOM Files:** Enables the uploading of medical imaging data required for AI-based lung nodule analysis.
- **Preview DICOM Files:** Allows optional visual inspection of uploaded images prior to analysis.
- **View CT Images:** Enables interactive browsing and inspection of CT slices.
- **View AI Malignancy Prediction:** Displays malignancy risk estimates, confidence indicators, and explainable AI outputs.
- **Review Case:** Provides an integrated environment for clinical assessment by combining CT image visualization and AI-generated malignancy predictions.
- **Generate Report:** Allows the doctor to generate a structured medical report summarizing imaging findings and AI-assisted assessments.
- **Save Report:** Persists the generated report for later access and review.

Mandatory clinical actions, such as viewing CT images and AI predictions, are modeled using <<include>>, while optional behaviors, such as report generation and clinical note entry, are modeled using <<extend>> relationships.

8.8.4 Admin Use Cases

Admin use cases are associated exclusively with system management and monitoring functionalities:

- **View Dashboard:** Provides a high-level overview of system usage and activity summaries.
- **Manage Users:** Allows the creation, modification, and management of user accounts and roles.
- **View System Statistics:** Enables monitoring of performance metrics and usage data.
- **View Activity Logs:** Supports auditing and traceability through access to system event logs.
- **Configure System Settings:** Enables adjustment of system-level configurations without interfering with clinical workflows.

Administrative use cases are intentionally isolated from medical decision processes to maintain system security and workflow integrity.

8.8.5 Use Case Diagram Overview

Figure 8.8.5.1 presents the use case diagram of the LungXai system, illustrating role-based interactions between clinical and administrative users. The diagram emphasizes functional access rights through the use of <<include>> and <<extend>> relationships, while workflow sequencing and user interface navigation are addressed in subsequent sections.

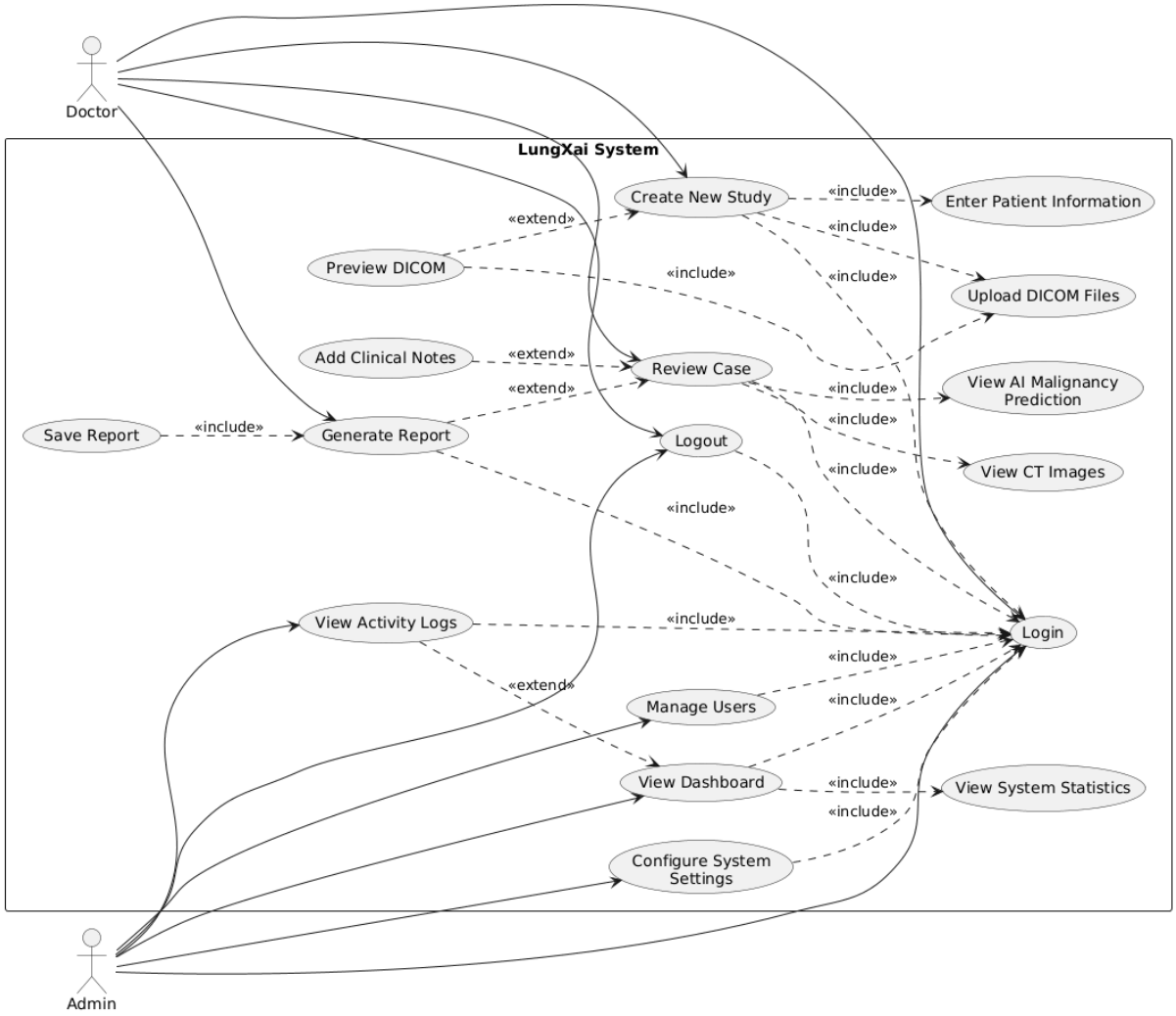

Figure 8.8.5.1: Use case diagram of the LungXai system illustrating role-based interactions of Doctor and Admin users with system functionalities.

9. Optimization / Acceleration Study (Exploratory)

This section reports an **exploratory optimization study** focusing on the computational efficiency of the LungXai system. Unlike previous sections, which address model design and learning methodology, this part specifically examines **runtime and resource-related considerations**

encountered during experimentation. The goal is not to present a fully optimized system, but to document early observations and design decisions related to performance under limited hardware conditions.

9.1 Motivation for an Exploratory Study

During baseline experimentation, it became evident that even relatively lightweight 2.5D classification pipelines introduce non-trivial computational overhead when applied to medical imaging data. CT-based workflows differ from natural image tasks due to volumetric data handling, medical preprocessing requirements, and evaluation with probability-based metrics.

Therefore, an exploratory optimization study was conducted to:

- Identify practical efficiency constraints,
- Understand which components dominate runtime,
- Guide future design decisions before integrating more complex modules such as segmentation and explainability.

This section intentionally avoids claiming final performance gains and instead focuses on engineering insights gained during development.

9.2 Observed Computational Constraints

Through empirical experimentation, the following constraints were identified:

- **Preprocessing overhead dominates early experimentation**, especially when repeatedly reconstructing samples from raw medical data.
- **GPU utilization is not always the primary bottleneck**; inefficient data preparation can limit throughput even with a lightweight backbone.
- **Memory usage scales with input dimensionality**, making future extensions (e.g., segmentation or multitask heads) sensitive to VRAM constraints.

These observations motivated targeted investigations rather than global architectural changes.

9.3 Explored Efficiency-Oriented Strategies

Several efficiency-oriented strategies were examined at both conceptual and experimental levels to better understand the computational characteristics of the proposed system.

9.3.1 Reuse of Preprocessed Data

Since many preprocessing steps are deterministic, reusing intermediate representations across epochs was explored to reduce redundant computation. This strategy is particularly relevant for medical datasets where raw input processing is significantly more expensive than model forward passes.

9.3.2 Training Configuration Sensitivity

Training stability and runtime were observed to be sensitive to configuration choices such as batch size and evaluation frequency. Rather than maximizing throughput, the focus was placed on identifying stable configurations that minimize unnecessary computation while preserving reliable validation behavior.

9.3.3 Early Consideration of Precision and Memory

Precision-aware training strategies were considered from an early stage to assess their potential impact on memory consumption and iteration speed. While not fully integrated, these considerations informed later architectural planning.

9.4 Relation to Future System Extensions

The findings from this exploratory study directly inform future development stages. In particular:

- **Segmentation-based pipelines** are expected to increase both memory and compute requirements, making early efficiency awareness critical.
- **Multitask learning and localization heads** will further increase model complexity, reinforcing the need for careful resource management.
- **Deployment-oriented scenarios** (e.g., UI-driven inference) will benefit from early identification of runtime-sensitive components.

Thus, this section serves as a bridge between the current baseline system and more computationally demanding future extensions.

10. Future Work and Next Phase Plan

This section outlines the planned extensions of the LungXai framework that build upon the established baseline system. The future work focuses on improving anatomical awareness, model capacity, interpretability, computational efficiency, and deployment readiness.

10.1 Segmentation-Guided Learning (SegResNet)

A SegResNet-based lung nodule segmentation module is planned to be integrated into the system to provide anatomically accurate nodule masks derived from XML annotations. These masks will support ROI-guided classification by reducing background influence and enforcing spatial focus on the nodule region.

Segmentation performance will be evaluated using Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). Classification performance before and after ROI refinement will be compared to quantify the contribution of segmentation to malignancy discrimination.

10.2 Expansion of Classification Backbones

While ResNet-18 serves as a lightweight baseline, stronger architectures such as DenseNet and EfficientNet variants will be evaluated under identical experimental conditions. These comparisons will assess the impact of backbone selection on ROC-AUC, sensitivity, specificity, training time, and inference efficiency.

10.3 Explainable Artificial Intelligence (XAI)

Explainability methods such as Grad-CAM and attention-based visualization will be integrated to highlight regions contributing to model predictions. These outputs will enable qualitative assessment of whether the model focuses on anatomically relevant nodule regions rather than spurious background patterns.

An exploratory Grad-CAM reliability analysis conducted in this study (Appendix A) motivates deferring full XAI integration until segmentation and ROI-guided learning are established.

10.4 Multitask Learning and Nodule Localization

Future extensions include multitask learning formulations in which the model jointly performs malignancy classification and spatial localization by predicting nodule center coordinates. This

approach is expected to improve generalization and enable direct visual marking of nodules within the UI.

10.5 Model Acceleration and Efficiency-Oriented Studies

Further studies will focus on optimizing training and inference efficiency through data caching strategies, precision-aware training, and pipeline-level optimizations. These efforts aim to support scalability as more computationally demanding modules are introduced.

10.6 Deployment-Oriented Structuring and Viewer Compatibility

Long-term development goals include structuring the LungXai pipeline to be compatible with medical imaging viewers such as OHIF and 3D Slicer. This will allow classification results and future explainability outputs to be reviewed alongside original CT images in clinically familiar environments.

11. Conclusion

This study presented LungXai, a modular and reproducible deep learning framework for lung nodule malignancy classification based on chest CT images. Rather than proposing a fully integrated and complex system from the outset, the project deliberately followed a baseline-first and progressive development strategy, prioritizing correctness, interpretability, and engineering robustness over premature architectural complexity.

In the current phase, an end-to-end classification pipeline was successfully established using a 2.5D input representation and a ResNet-18 backbone within the MONAI ecosystem. This configuration enabled the system to capture limited volumetric context while maintaining computational efficiency suitable for constrained hardware environments. The use of standardized preprocessing, annotation-guided sample construction, and structured experiment management ensured reproducibility and facilitated systematic evaluation.

Experimental results demonstrated that meaningful discrimination between benign and malignant nodules can be achieved even with a lightweight baseline model. Importantly, the analysis revealed a clear divergence between accuracy and ROC-AUC across training epochs, reinforcing the necessity of threshold-independent metrics for medical classification problems with class imbalance. Accordingly, ROC-AUC was adopted as the primary evaluation criterion, providing a more reliable measure of clinical discrimination capability.

Beyond model performance, the project emphasized system-level design. The proposed architecture clearly separates data handling, representation learning, modeling, evaluation, and presentation layers. The inclusion of a prototype user interface (UI) further demonstrated how model outputs can be integrated into a practical software environment, enabling structured visualization and interaction without tightly coupling the interface to model internals.

Finally, an exploratory optimization study highlighted key computational considerations encountered during experimentation. Rather than claiming final acceleration gains, this analysis provided early engineering insights that inform future scalability, particularly as more computationally demanding modules are introduced.

References

[1] https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[2] Kim, H., Lee, D., Cho, W.S. et al. CT-based deep learning model to differentiate invasive pulmonary adenocarcinomas appearing as subsolid nodules among surgical candidates: comparison of the diagnostic performance with a size-based logistic model and radiologists. Eur Radiol 30, 3295–3305 (2020). https://doi.org/10.1007/s00330-019-06628-4

[3] Canayaz, M., Şehribanoğlu, S., Özgökçe, M. et al. A comprehensive exploration of deep learning approaches for pulmonary nodule classification and segmentation in chest CT images. Neural Comput & Applic 36, 7245–7264 (2024). https://doi.org/10.1007/s00521-024-09457-9

[4] Saied, M., Raafat, M., Yehia, S. et al. Efficient pulmonary nodules classification using radiomics and different artificial intelligence strategies. Insights Imaging 14, 91 (2023). https://doi.org/10.1186/s13244-023-01441-6

[5] Liu, Y., Hsu, H.Y., Lin, T. et al. Lung nodule malignancy classification with associated pulmonary fibrosis using 3D attention-gated convolutional network with CT scans. J Transl Med 22, 51 (2024). https://doi.org/10.1186/s12967-023-04798-w

[6] Wulaningsih, W., Villamaria, C., Akram, A. et al. Deep Learning Models for Predicting Malignancy Risk in CT-Detected Pulmonary Nodules: A Systematic Review and Meta-analysis. Lung 202, 625–636 (2024). https://doi.org/10.1007/s00408-024-00706-1

[7] https://github.com/Project-MONAI/MONAI

[8] https://www.qure.ai/

[9] https://optellum.com/

Appendix A – Exploratory Explainable AI (XAI) Reliability Assessment

Project: Lung Cancer Detection via CT Scans

Methodology: Gradient-weighted Class Activation Mapping (Grad-CAM)

Model Architecture: DenseNet121 (Timm library)

A.1 Executive Summary

This appendix presents an exploratory explainable artificial intelligence (XAI) reliability assessment conducted as a standalone technical study outside the main LungXai pipeline. The objective of this study is to evaluate the transparency and decision-making behavior of a convolutional neural network when analyzed using Grad-CAM.

Initial testing on synthetic data reveals that while the technical pipeline for explainability is fully functional, the employed ImageNet-pretrained model lacks the clinical domain knowledge required to localize pulmonary nodules. This limitation is quantitatively demonstrated by a localization error of 49.68 pixels, indicating that the generated attention maps do not align with the target lesion.

A.2 Grad-CAM Implementation Details

The Grad-CAM module was applied to transform the trained model from a black-box predictor into a visually interpretable system. The following implementation steps were conducted:

- **Layer Selection:** The target layers were defined as model.features[-1], corresponding to the final convolutional block of the DenseNet121 architecture. This layer captures high-level semantic features relevant to the classification decision.
- **Gradient Tracking:** The Grad-CAM mechanism tracks gradients of the predicted target class (Malignant or Benign) flowing back into the selected feature maps during backpropagation.
- **Heatmap Generation:** A grayscale importance map is generated, where pixel intensity represents the relative contribution of each spatial location to the final classification decision.
- **Visualization (Overlay):** The show_cam_on_image function is used to overlay the Grad-CAM heatmap onto the original grayscale CT slice using a Jet color map (red indicating high importance and blue indicating low importance).

A.3 Experimental Results and Visual Analysis

The observations derived from this analysis are summarized in Table A.3.1.

Table A.3.1: Observations derived from the Grad-CAM analysis on the synthetic lung slice.

Metric	**Observation**

Metric	Observation
Input Signal	Synthetic lung slice with a bright artificial nodule located at coordinates (150, 80).
Model Focus	The Grad-CAM heatmap shows maximum activation at the geometric center of the lung region.
Target Alignment	The model completely bypasses the synthetic nodule, which is marked by a reference red dot.
Localization Error	49.68 pixels, computed as the Euclidean distance between the heatmap maximum and the known nodule center.

These results indicate that the model's attention is dominated by global spatial bias rather than lesion-specific features.

A.4 Root Cause Analysis

The observed failure to localize the target nodule can be attributed to the following factors:

1. **Untrained Weights:** The network relies on ImageNet-pretrained weights, which prioritize general geometric and texture patterns rather than medically meaningful structures such as pulmonary nodules.
2. **Spatial Bias:** The model exhibits a strong center bias, implicitly assuming that the most informative content is located near the image center. This behavior limits sensitivity to small, high-frequency anomalies like nodules.

A.5 Conclusion and Recommendations

The Grad-CAM pipeline is technically correct and successfully exposes the underlying logical limitations of the current model. However, the experiment demonstrates that explainability methods alone do not guarantee clinically meaningful interpretations.

To achieve clinical utility, the following steps are required:

- **Domain-Specific Training:** Fine-tuning the model on medical datasets such as LIDC-IDRI is necessary to shift attention from generic image regions toward anatomically relevant lesions.

- **Iterative Validation:** Repeating the Grad-CAM analysis at regular training intervals (e.g., every 10 epochs) can be used to monitor the progressive alignment of attention maps with the true nodule location.

Overall, this exploratory study supports the decision to defer explainable AI integration within the LungXai system until segmentation and ROI-guided learning are introduced, ensuring anatomically grounded and clinically interpretable explanations.