# Lung Cancer Detection Using CoAtNet with Optuna-Based Hyperparameter Optimization

*Sanjuktarani Jena[1,2], Yash Shah[2], Upendra Kumar[3], Sai Kiran Oruganti[4]*
[1]Lincoln University College Malaysia
[2]Department of Computer Engineering Sardar Patel Institute of Technology Mumbai, India
3Department of Computer Science Engineering Dr A P J Abdul Kalam Technical University Lucknow, Uttar Pradesh, India
sanjuktarani.jena@spit.ac.in; upendra.ietlko@gmail.com; upendra.ietlko@gmail.com; saisharma@lincoln.edu.my

***Abstract*** - Deaths associated with cancers are on the rise, and one of the primary causes of these fatalities is lung cancer. This makes accurate and dependable detection systems all the more important. This work presents a framework for the detection of lung cancer that builds upon the CoAtNet model in combination with Optuna-based hyperparameter optimization. Experiments were performed using the IQ-OTH/NCCD dataset, which comprises 1190 CT scan slices from approximately 110 patients. The model was trained for 30 epochs, and the performance was then compared with three widely used architectures: ResNet-18, ResNet-50, and Swin Transformer Tiny. Among these, the CoAtNet-Optuna approach demonstrated the best performance, achieving an accuracy of 99.39%. These results highlight that combining convolution-attention networks with automated hyperparameter tuning can substantially enhance the accuracy of lung cancer detection. This suggests the potential application of this method in supporting clinical and computer-aided diagnosis.

*Index Terms - Optuna optimisation, Convolutional Neural Networks (CNN), Transfer Learning, Computed Tomography (CT), CoAtNet*

## I. INTRODUCTION

Among the cancer related deaths, lung cancer is one of the leading causes of mortality throughout the world, despite the advancements made in processes for prevention, early detection and therapy. Nearly 1.8 million deaths were associated with lung cancer in 2022 alone and there are over 2 million new cases annually, with 20 diseases (source- Times of India, Aug. 1, 2025). Early computer aided diagnosis (CAD) [1] systems depended upon very specific handcrafted features like morphological descriptors and texture analysis which was followed by traditional classifiers (SVM, random forests, kNN). The drawback of these approaches is that it lacks the generalization across various imaging settings. The establishment of benchmarks that are standardized, especially the LIDC-IDRI dataset and LUNA16 challenge, accelerated CAD development by significant amounts. Radiomics introduced high- dimensional quantitative image descriptors for nodules, which enables statistical risk modelling. However limited replicability of feature and insufficiency of external validation hindered clinical adoption. Deep learning [22], [16], especially convolutional neural networks (CNNs) [2], enabled end-to-end learning from raw image data, which resulted in achieving strong performance in 2D and 3D nodule detection and malignancy classification. Yet, CNNs [21] [25] in volumetric scans CNNs often struggle to capture a universal contextual relationship. Vision Transformers (ViTs) [23], [11] overcome this limitation by modeling long range dependencies, despite the fact that they require large datasets and are less adept at fine grained local feature extraction. An emerging solution that has come up in recent times has been Hybrid architectures that combine CNNs and Transformers. CoAtNet amplifies this approach, integrating convolutional layers for more efficient local feature capture with global context modeling- this is well suited for staging as well as malignancy prediction for lung cancer. Recent advances in large language models (LLMs) [24] encompass medical AI into multimodal domains, integrating Economic Health Records (EHR). Regulated domain specific LLMs on lung CT report corpora enable radiomics text fusion, classification enhancement performance by incorporating both imaging biomarkers and clinically relevant contextual data. This work proposes a hybrid CNN- Vision Transformer

approach based on CoAtNet for lung cancer classification. The method leverages convolutional blocks to extract high-resolution local features from pulmonary nodules and attention modules to include global semantic information along with optuna optimization technique, thus enhancing prediction of malignancy. Evaluations have been made based on IQ-OTH/NCCD dataset and demonstrate its abilities over conventional CNNs, ViTs, and other hybrid models in terms of accuracy and sensitivity to variations in imaging protocols. The sections ahead present a summary of the related literature. Section III provides the dataset description and experimental with Section IV carrying the outlines, methodology and proposed block diagrams. Section V consists of results and analysis with the Section VI concluding the study. and highlights directions for future research.

## II. LITERATURE SURVEY

Traditional CNNs [13] [14] widely adopted because they effectively capture local features in medical images like CT scans and chest X-rays. However, they struggle to model long range relationships. To fill this gap, Transformer-based models [11], Though, initially built for natural language processing, they have been tailored over time to support applications in medical imaging. The models show considerable improvements in understanding global context. Recently, hybrid CNN and Transformer frameworks have appeared, combining the spatial sensitivity of CNNs with the global attention methods of Transformers. Structures like CoAtNet [26], [19] merge convolution and self-attention, providing both scalability and interpretability across different resource environments. These advancements have led to better accuracy for lung cancer detection.

Yang et al. [3] developed a multi-view CNN framework enhanced with a squeeze-and-excitation module to improve feature representation. LIDC-IDRI dataset was used for achieving a binary classification accuracy of 96.04%. The model effectively captured the spatial heterogeneity of lung nodules and addressed challenges associated with multi-view variability, leading to more robust diagnostic performance.

Zia Ur Rehman et al. [4] introduced a customized CNN architecture combined with an attention mechanism, which is allowing this model to focus on the most important features during analysis. The approach was evaluated on the LUNA16 dataset, where it demonstrated superior performance across multiple evaluation metrics in comparison with the existing state-of-the-art methods. Chen et al. [5] presented a TransUNet-based framework for thyroid nodule detection, incorporating a dual-loss function that simultaneously optimized both localization and classification objectives. Their approach achieved a 3.9% improvement in accuracy compared to conventional CNN-based models, highlighting its effectiveness in medical image analysis.

The Lung Nodule-SSM model [7] was introduced in a separate study, utilizing self-supervised learning on DINOv2 for enhancing both detection as well as classification of lung nodules. In this approach, DINOv2 was pre-trained on unlabelled CT scans from the LUNA16 dataset to extract robust feature representations, and subsequently fine-tuned with a Transformer for nodule detection and diagnosis. The proposed approach attained an accuracy of 98.37% in detection and classification of the nodules.

Ramezani et al. [8] proposed a Deformable Detection Transformer (DETR)-based framework to address the challenges to detect rare lung nodules in predominantly normal datasets. The method integrates DETR, Maximum Intensity Projection (MIP), and a customized focal loss function to improve sparse nodule detection. Specifically, a 7.5 mm MIP was applied to reduce nodule sparsity and enrich spatial context, thereby enhancing feature representation. To mitigate class imbalance, the model employed a tailored focal loss, which facilitated accurate bounding box prediction around nodules. Evaluated on LUNA16 dataset, the approach achieved 94.2% F1-score, with 95.2% recall and 93.3% precision, under conditions where the test dataset exhibited only 4% nodule sparsity.

Yang et al. [9] investigation was carried out on eight pretrained Transformer models belonging to the BERT, RoBERTa, and ALBERT families, assessing their ability to handle tasks such as extracting clinical concepts, identifying relations, and detecting negations. To assess their effectiveness, the Transformer-based models were benchmarked against classical approaches like RNNs and bidirectional LSTMs. RoBERTa stood out among the tested models with 0.9279 as F1-score for nodule extraction, while ALBERT-base and GatorTron achieved superior results with F1-score as 0.9737. In addition, seven out of the eight Transformer models reached a perfect F1-score of 1.0 for negation detection, and the framework as a whole reported an F1-score of 0.8869.

Yadav et al. [10] proposed EDTNet, was presented as an encoder–decoder Transformer model capable of capturing extended spatial dependencies for lung nodule detection. The architecture features Transformer layers for downsampling/upsampling, a patch-expansion unit for resolution recovery, and a cross-attention mechanism in the decoder. To improve feature representation, ESLA (Spatially Aware Local Attention) was incorporated in the encoder–decoder stages to refine edges and small-nodule details, while ESGA (Spatially Aware Global Attention) in the bottleneck captured global context. When evaluated against models such as UNet, ResUNet++, DeepLabV3+, and Swin-UNet, EDTNet achieved strong results, with 96.27% precision, 95.81% IoU, and 96.15% Dice score on DS1, and 98.84% sensitivity, 96.06% IoU, and 97.85% Dice score on DS2, outperforming existing baselines.

Saha et al. [12] investigated Interstitial Lung Disease (ILD) pattern recognition using a Vision Transformer (ViT) framework that employed multi-head self-attention to model both the local as well as the global spatial dependencies. Even after adjusting hyperparameters, which includes the number of attention heads and hidden units, the model attained a notable level of test accuracy of 82.75%, indicating room for further optimization.

Gao et al. [15] proposed a Distanced Long Short-Term Memory (DLSTM) model designed for addressing the challenges of irregular longitudinal sampling in medical data. This framework incorporates a Temporal Emphasis Model (TEM), enabling effective learning across both regularly and irregularly sampled time intervals. The model was evaluated on three datasets: simulated sequences, the National Lung Screening Trial (NLST) CT scans, and 1,420 cases acquired clinically with heterogeneous and irregular temporal patterns. Experimental results showed consistent improvements, with the NLST dataset reporting an increase in F1-score from 0.6785 with a standard LSTM to 0.7085 using DLSTM. In external validation on irregular clinical data, DLSTM demonstrated strong generalization, achieving an AUC of 0.8905, outperforming both CNN-based feature methods (0.8350) and conventional LSTMs (0.8380).

L. Sun et al. [17] conducted an analysis using the LIDCIDRI dataset, where a 3D U-Net was employed to segment lung nodules and minimize background interference. To enhance feature representation, class features, image features, and complex attribute features were aligned through innovative learning. Furthermore, a tailored loss function optimization in the proposed Nodule-CLIP framework improved classification performance, achieving an accuracy of 90.6% and a recall of 92.81%. Another researcher [18] applied the YOLOv11 model to the dataset of 1,608 CT scan images, comprising 623 cancerous and 985 non-cancerous cases. The model outperformed previous YOLO versions, achieving a mAP of 96.26%, IoU of 95.76%, precision of 98.11%, and recall of 98.83% on the test set. In comparison, YOLOv10 achieved a mAP of 95.23%, while YOLOv9 reached 95.70%.

Chen et al. [20] analyzed a publicly available chest X-ray dataset using Efficient Net-b5 and CoAtNet-0-rw as backbone networks. To improve interpretability, Group-score-weighted Class Activation Mapping (Group-CAM) was employed for visual explanations of predictions. Results demonstrated that the pretrained CoAtNet-0-rw combined with Lours achieved the highest overall AUROC of 0.842, significantly outperforming the ResNet50 + LWBCE baseline (AUROC: 0.811, $p = 0.037$).

## III. DATASET AND EXPERIMENTAL SETUP

This study utilizes the IQ-OTH/NCCD Lung Cancer Dataset, which is a collection of Computed Tomography (CT) scans and was gathered at the Iraq-Oncology Teaching Hospital and the National Center for Cancer Diseases. Complete dataset comprises 1,190 CT scan slices from 110 patient cases, which were marked by specialist oncologists and radiologists. The cases fall into three categories: malignant (40 cases), benign (15 cases), and normal (55 cases). For our experiment, we used a version of the dataset containing 1,099 images. We partitioned this data using a stratified split to create a training set (70%, 769 images), a validation set (15%, 165 images), and a test set (15%, 165 images). This stratification ensures that the class distribution is consistently represented across all data subsets, which is crucial for training a well-balanced model.

The experimental setup was configured on the Kaggle platform using an NVIDIA P100 GPU. First, the images were resized to 224×224 pixels to match the input requirements of the CoAtNet model. Data augmentation technique is used to enhance the model's ability to generalize and prevent overfitting, exclusively to the training set. These augmentations include random horizontal flips as well as random rotations of up to 10 deg. Both the training and validation/test sets were subsequently converted to PyTorch tensors and normalized using the standard ImageNet mean ([0.485,0.456,0.406]) and standard deviation ([0.229,0.224,0.225]). To address the inherent class imbalance

within the dataset, class weights were calculated and integrated into the loss function, ensuring that the model did not favor the majority class. The model was trained with a batch size of 32 for 30 epochs. Hyperparameter optimization was conducted using Optuna over 50 trials, with each trial running for 5 epochs to efficiently determine the optimal configuration before commencing the full training regimen. Early stopping was implemented based on validation loss with a patience of 10 epochs to prevent overfitting.

## IV. METHODOLOGY

This section presents our comprehensive approach to lung cancer detection, detailing the proposed CoAtNet-based architecture and the experimental framework used to evaluate different deep learning models.

### *A. Proposed CoAtNet Architecture*

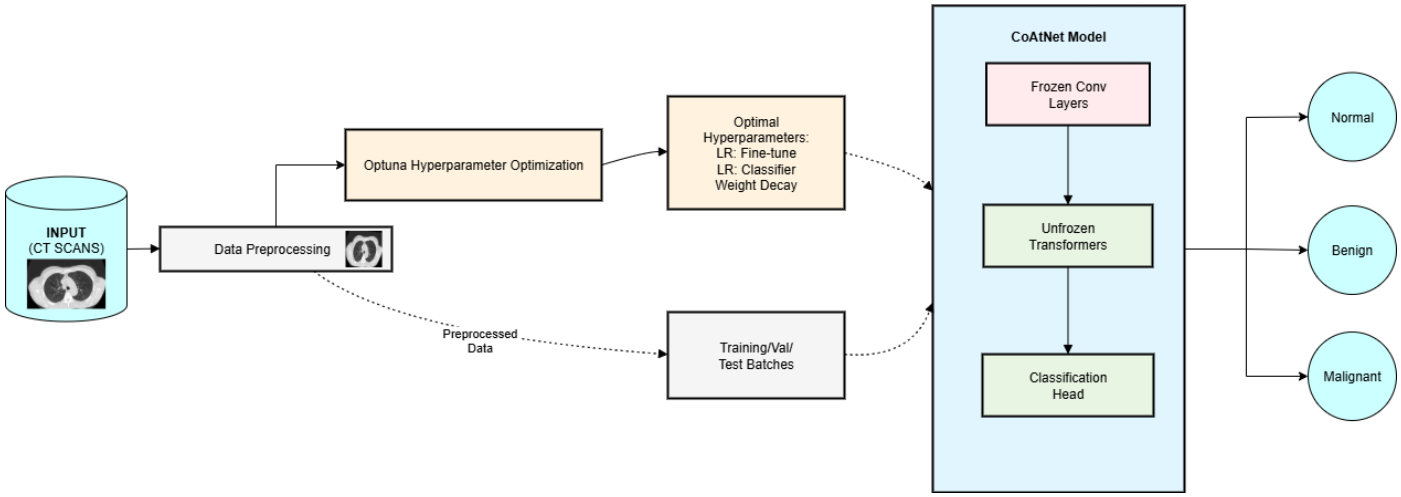
We propose a hybrid CNN-Vision Transformer (ViT) approach based on CoAtNet (Convolution and Attention Network) for lung cancer classification. CoAtNet combines the inductive biases of convolutional neural networks with the global modeling capabilities of the vision transformers, making it particularly suitable for medical image analysis where both local feature extraction and global context understanding are crucial. The architecture leverages the strengths of both paradigms: convolutional layers excel at capturing local patterns and spatial hierarchies essential for identifying morphological features in medical images, while transformer blocks provide superior capability for modeling long-range dependencies and global context that are critical for understanding the overall pathological state.

The CoAtNet architecture consists of four main stages, where Stages 0-1 employ convolutional layers for low-level feature extraction, capturing fine-grained details and local patterns characteristic of lung tissue structures. Stages 2-3 utilize transformer blocks for high-level semantic understanding, enabling the model to integrate information across different regions of the lung image and establish relationships between distant anatomical structures. The classification head comprises a fully connected layer that maps the learned representations to the final prediction classes.

Our fine-tuning strategy implements selective unfreezing of network components to optimize the balance between leveraging pre-trained knowledge and adapting to domain specific characteristics. Initially, all parameters are frozen to preserve the robust ImageNet-learned features that provide a strong foundation for natural image understanding. Subsequently, the final two transformer blocks (Stages 2-3) are selectively unfrozen to enable adaptation of high-level semantic representations to the specific patterns and structures present in lung cancer imagery. The classification head is also unfrozen to facilitate task-specific learning, while earlier convolutional stages remain frozen to maintain the strong low level feature extraction capabilities essential for processing medical images. This selective fine-tuning approach allows the model for leveraging the pre-trained knowledge while adapting the specific characteristics of lung cancer imagery, ensuring optimal performance without compromising the learned representations.

### *B. Hyperparameter Optimization with Optuna*

Fig. 1 Overall methodology flowchart showing the complete pipeline from data preprocessing to model evaluation, highlighting the Optuna optimization loop and comparative analysis framework.




To ensure optimal model performance, we employ Optuna, a hyperparameter optimization framework, for automated hyperparameter tuning. This approach eliminates the need for manual hyperparameter selection and ensures that each model architecture achieves its maximum potential performance through systematic exploration of the hyperparameter space. The optimization process is designed to be both efficient and comprehensive, utilizing advanced sampling strategies to converge on optimal configurations within a reasonable computational budget.

The objective function is comprehensively designed for evaluating model performance based on validation loss, providing a robust measure of generalization capability. The hyperparameter search space is carefully designed based on empirical studies and theoretical considerations for transfer learning in medical imaging applications as shown in Table I.

We implement a sophisticated differential learning rate strategy where different network components receive distinct learning rates, enabling more nuanced adaptation of the pretrained model:

- **Transformer Layers (Stages 2-3):** $LR_{transformer} = LR_{finetune}$ (conservative updates)
- **Classification Head:** $LR_{classifier}$ (aggressive adaptation)
- **Frozen Layers (Stages 0-1):** LR = 0 (no parameter updates)

This strategy allows for more conservative updates to the pre-trained transformer layers while enabling more aggressive adaptation of the classification head, ensuring that valuable pre-trained representations are preserved while allowing task specific adaptation.

| Hyperparameters | Search Range | Scale |
|---|---|---|
| Fine-tuning Learning Rate | $[1 \times 10^{-6}, 1 \times 10^{-4}]$ | Logarithmic |
| Classifier Learning Rate | $[1 \times 10^{-5}, 1 \times 10^{-3}]$ | Logarithmic |
| Weight Decay | $[1 \times 10^{-5}, 1 \times 10^{-2}]$ | Logarithmic |

Table I: Optuna hyperparameter search space configuration.

The Optuna framework employs Tree-structured Parzen Estimator (TPE) sampling to efficiently explore the hyperparameter space through intelligent sequential model-based optimization. Each trial involves model initialization with suggested hyperparameters, training for 5 epochs on the training set to obtain a reliable

performance estimate, validation loss evaluation, and performance feedback to the optimizer for future trial guidance. A total of 50 trials are conducted to ensure

comprehensive exploration of the hyperparameter space while maintaining computational efficiency.

### *B. Pipeline Overview*

Our methodology follows the pipeline shown in Fig 1. Preprocessed CT scans are divided into training/validation/test sets. Optuna optimization runs parallel to model training, iteratively suggesting hyperparameters based on validation performance. The optimized CoAtNet model then processes test images through its hybrid architecture to generate final classifications.

### *C. Baseline Models and Comparative Analysis*

To validate the effectiveness of our proposed CoAtNet approach, we implement and compare against several architectures that represent different paradigms in deep learning for image classification. The baseline models include ResNet- 18, a lightweight residual network with 18 layers that provides efficient computation while maintaining reasonable performance, and ResNet- 50, a deeper residual network with 50 layers and bottleneck blocks that offers enhanced representational capacity through increased depth. Additionally, we evaluate against Swin Transformer (Tiny variant), a hierarchical vision transformer with shifted window attention that represents the pure transformer approach to image classification, providing insight into the relative benefits of hybrid architectures versus pure transformer models.

Each baseline model undergoes optimization using the same Optuna framework to ensure fair comparison and optimal performance across all architectures. This consistent optimization approach eliminates potential bias that could arise from differential hyperparameter tuning efforts, ensuring that observed performance differences can be attributed to architectural advantages rather than optimization disparities.

### *D. Training Strategy and Loss Function*

Given the inherent class imbalance in medical datasets, we implement a weighted cross-entropy loss function to address the disproportionate representation of different classes. The loss function is formulated as $\mathcal{L} = -\sum_{i=1}^{N} w_{y_i} \log(\hat{y}_i)$, where $w_{y_i}$ represents the weight for class $y_i$, computed as $w_i = \frac{N_{total}}{N_{class_i}}$. This weighting scheme ensures that the model receives appropriate penalty signals for misclassifying underrepresented classes, promoting balanced learning across all diagnostic categories.

Table II: Comprehensive performance comparison of deep learning architectures on medical image classification task. Best results are highlighted in **bold**.

| Model | Test Acc. | Best Val Acc. | Final Val. Loss | Precision | Recall | F1-Score |
|---|---|---|---|---|---|---|
| ResNet-18 | 97.58% | 97.58% | 0.1558 | 0.98 | 0.93 | 0.95 |
| ResNet-50 | 97.58% | 98.18% | 0.1874 | 0.96 | 0.94 | 0.95 |
| Swin Tiny | 98.18% | 98.79% | 0.2269 | 0.98 | 0.94 | 0.96 |
| **CoAtNet (Proposed)** | **99.39%** | **100.00%** | **0.0295** | **0.99** | **0.98** | **0.99** |

The optimization strategy employs the AdamW optimizer with decoupled weight decay, which has demonstrated superior performance in transfer learning scenarios. The optimizer configuration includes Optuna-optimized learning rates that are specifically tailored for each model component, Optuna optimized regularization through weight decay parameters, and a batch size of 32 that is optimized for GPU memory constraints while maintaining sufficient gradient estimation quality. This configuration balances computational efficiency with learning stability, ensuring convergence across different model architectures.

To improve model generalization, we apply a comprehensive data augmentation strategy during training. The augmentation pipeline includes random horizontal flipping with 50% probability to account for anatomical symmetries, random rotation within ±10 degrees to simulate natural variations in patient positioning, resizing to 224 × 224 pixels to meet CoAtNet input requirements, and ImageNet normalization using standard parameters (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]) to maintain compatibility with pretrained weights. These augmentations enhance the model's ability to generalize to unseen variations in lung cancer imagery while preserving the essential diagnostic features.

### *E. Model Evaluation*

1) Performance Metrics: We evaluate model performance using the following metrics:

- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$, F1-Score $= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
- AUC-ROC: Area under the Receiver Operating Characteristic curve

Where these metrics are defined by True Positives (TP) and True Negatives (TN) for correct classifications, and False Positives (FP) and False Negatives (FN) for incorrect classifications. The proposed methodology ensures a robust and comprehensive evaluation of different architectures while providing optimal hyperparameter configurations for each model, enabling fair comparison and reliable performance assessment for lung cancer detection.

## V. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of our proposed CoAtNet model against established baseline architectures on the medical image classification task. We evaluate four distinct deep learning architectures: ResNet-18, ResNet50, Swin Transformer Tiny, and our proposed CoAtNet model. All experiments were conducted under identical conditions with optimized hyperparameters to ensure fair comparison.

### *A. Training Dynamics Analysis*

Figure 2 illustrates the training dynamics across all four architectures throughout the 30-epoch training process. The comprehensive comparison reveals distinct convergence patterns and performance characteristics for each model.

The training loss curves (Figure 2a) demonstrate that CoAtNet achieves the fastest convergence, reaching nearzero training loss by epoch 10, significantly outpacing other architectures. The validation loss analysis (Figure 2c) confirms CoAtNet's superior generalization capability, maintaining the lowest validation loss throughout training while avoiding overfitting.

Notably, CoAtNet exhibits exceptional stability in both training and validation metrics, with minimal fluctuation after initial convergence. This stability indicates robust feature learning and superior architectural design for the medical image classification task.

### *B. Quantitative Performance Evaluation*

Table II presents a comprehensive quantitative comparison of all evaluated models across multiple performance metrics.

### *C. ROC Analysis and Discriminative Performance*

We evaluated CoAtNet using Receiver Operating Characteristic (ROC) analysis on the test set. As shown in Figure 3, the model achieved a perfect AUC score of 1.00, indicating flawless discrimination among normal, benign, and malignant cases. This result underscores CoAtNet's reliability


Figure 2: Comprehensive training performance comparison across four deep learning architectures. (a) Training loss convergence, (b) Training accuracy progression, (c) Validation loss evolution, and (d) Validation accuracy trends over 30 training epochs. CoAtNet demonstrates superior convergence speed and stability compared to baseline methods.

in clinical contexts, where minimizing both false positives and false negatives is critical.

*D. Class-wise Performance Analysis*

Table III provides detailed class-wise performance metrics for our proposed CoAtNet model, demonstrating exceptional performance across all medical image categories.

Table III: Detailed class-wise performance metrics for CoAtNet on the test set.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Benign Cases | 1.00 | 0.94 | 0.97 | 18 |
| Malignant Cases | 1.00 | 1.00 | 1.00 | 85 |
| Normal Cases | 0.98 | 1.00 | 0.99 | 62 |

| | | | | |
|---|---|---|---|---|
| **Weighted Average** | **0.99** | **0.99** | **0.99** | **165** |

*E. Computational Efficiency Analysis*

Beyond accuracy metrics, we evaluated computational efficiency across all models. Table IV summarizes the training time and memory requirements.

Table IV: Computational efficiency comparison across evaluated architectures.

| Model | Parameters | FLOPs | Memory (GB) |
|---|---|---|---|
| ResNet-18 | 11.7M | 1.8G | 2.1 |
| ResNet-50 | 25.6M | 4.1G | 3.8 |
| Swin Tiny | 28.3M | 4.5G | 4.2 |
| **CoAtNet (Proposed)** | **23.1M** | **3.7G** | **3.5** |

*F. Qualitative Analysis and Visual Interpretation*

Figure 4 shows representative predictions across all three classes. CoAtNet achieves accurate classification under varying image quality and case complexity, highlighting its ability to generalize to diverse clinical scenarios. Key observations include:

- Accurate Classification: Correct predictions across all categories.
- Robustness: Handles diverse image qualities and anatomical variations.
- Class-Specific Learning: Captures distinct morphological patterns of each class.

These visual results support the quantitative findings and demonstrate CoAtNet's clinical potential in reliable lung cancer detection.

*G. Key Findings and Insights*

CoAtNet demonstrates clear advantages over baseline models:

- Highest Accuracy: 99.39% test accuracy and perfect recall for malignant cases.
- Strong Generalization: Lowest validation loss (0.0295) with rapid convergence.
- Efficient Design: Competitive parameter count (23.1M) and moderate memory usage.

These results highlight CoAtNet's reliability, efficiency, and clinical relevance for lung cancer detection.

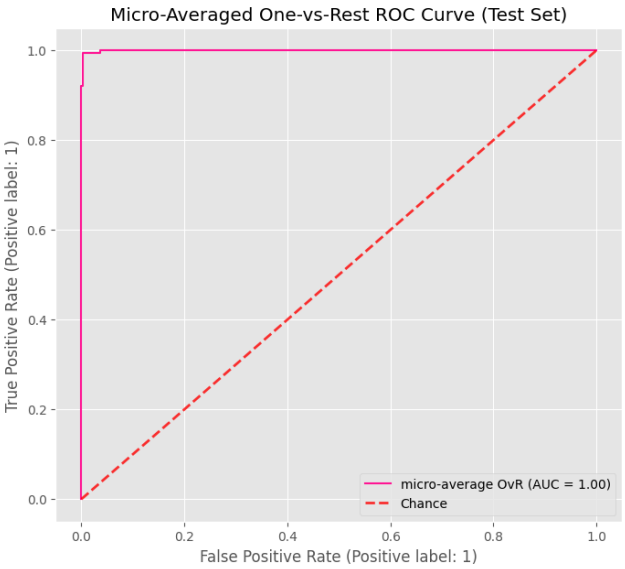



Figure 3: Micro-averaged one-vs-rest ROC curve for CoAtNet on the test set, achieving an AUC of 1.00.

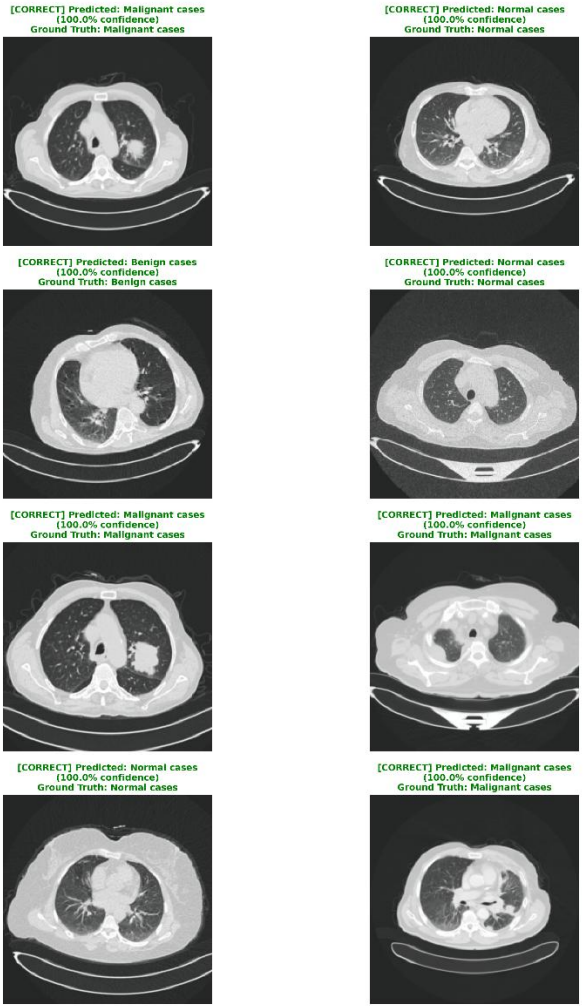



Figure 4: Representative medical image samples with ground truth labels and CoAtNet predictions.

## VI. CONCLUSION AND FUTURE SCOPE

This works consists of a framework for lung cancer detection based on CoAtNet model, which was optimized using Optuna on the IQ-OTH/NCCD dataset. The method has achieved a clearcut accuracy of 99.39%, outperforming other methods such as ResNet-18, ResNet-50, and Swin Transformer Tiny which have similar experimental settings. These results show efficiency of combining convolutional mechanisms with systemic hyperparameter optimization for medical image analysis. This study can be expanded upon by incorporating multimodal information such as CT scans, clinical records, and genomic data to provide a more comprehensive diagnosis. Such integrations can pave the path towards a more reliable, interpretable, and patient-centered system with an earlier detection as well as detection in lung cancer management.

For future work, the study can be extended by incorporating multimodal information such as CT scans, clinical records, and genomic data to provide a more comprehensive diagnosis. Moreover, large language models like Bio-GPT hold promise for analyzing unstructured clinical notes and radiology reports, offering complementary insights to imaging data. Such integration may pave the way toward more reliable, interpretable, and patient-centered decision-support systems in lung cancer detection and management.

## REFERENCES


1. Mastouri, R., Khlifa, N., Neji, H., & Hantous-Zannad, S. (2020). Deep learning-based CAD schemes for the detection and classification of lung nodules from CT images: A survey. Journal of X-ray Science and Technology, 28(4), 591–617.
2. Dey, R., Lu, Z., & Hong, Y. (2018, April). Diagnostic classification of lung nodules using 3D neural networks. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp. 774–778). IEEE.
3. Yang, Y., Li, X., Fu, J., Han, Z., & Gao, B. (2023). 3D multi-view squeeze-and-excitation convolutional neural network for lung nodule classification. Medical Physics, 50(3), 1905–1916.
4. Ur Rehman, Z., Qiang, Y., Wang, L., Shi, Y., Yang, Q., Khattak, S. U., ... & Zhao, J. (2024). Effective lung nodule detection using deep CNN with dual attention mechanisms. Scientific Reports, 14(1), 3934.
5. Chen, L., Chen, H., Pan, Z., Xu, S., Lai, G., Chen, S., ... & Zhang, Y. (2023). Thyroidnet: A deep learning network for localization and classification of thyroid nodules. Computer Modeling in Engineering & Sciences, 139(1), 361.
6. Niu, C., & Wang, G. (2022). Unsupervised contrastive learning based transformer for lung nodule detection. Physics in Medicine & Biology, 67(20), 204001.
7. Noreen, M., & Shaukat, F. (2025). Lung Nodule-SSM: Self-Supervised Lung Nodule Detection and Classification in Thoracic CT Images. arXiv preprint arXiv:2505.15120.
8. Tang, J., Chen, X., Fan, L., Zhu, Z., & Huang, C. (2025). LN-DETR: An efficient Transformer architecture for lung nodule detection with multi-scale feature fusion. Neurocomputing, 633, 129827.
9. Yang, S., Yang, X., Lyu, T., Huang, J. L., Chen, A., He, X., ... & Bian, J. (2024). Extracting pulmonary nodules and nodule characteristics from radiology reports of lung cancer screening patients using transformer models. Journal of Healthcare Informatics Research, 8(3), 463–477.
10. Yadav, D. P., Sharma, B., Webber, J. L., Mehbodniya, A., & Chauhan, S. (2024). EDTNet: A spatial aware attention-based transformer for the pulmonary nodule segmentation. PloS One, 19(11), e0311080.
11. Ali, H., Mohsen, F., & Shah, Z. (2023). Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review. BMC Medical Imaging, 23(1), 129.
12. Saha, S., Kumar, A., & Nandi, D. (2024). ViT-ILD: a vision transformerbased neural network for detection of interstitial lung disease from CT images. Procedia Computer Science, 235, 779–788.
13. Faizi, M. K., Qiang, Y., Wei, Y., Qiao, Y., Zhao, J., Aftab, R., & Urrehman, Z. (2025). Deep learning-based lung cancer classification of CT images. BMC Cancer, 25(1), 1056.

14. Paez, R., Kammer, M. N., Balar, A., Lakhani, D. A., Knight, M., Rowe, D., ... & Maldonado, F. (2023). Longitudinal lung cancer prediction convolutional neural network model improves the classification of indeterminate pulmonary nodules. Scientific Reports, 13(1), 6157.
15. Gao, R., Huo, Y., Bao, S., Tang, Y., Antic, S. L., Epstein, E. S., ... & Landman, B. A. (2019, October). Distanced LSTM: time-distanced gates in long short-term memory models for lung cancer detection. In International Workshop on Machine Learning in Medical Imaging (pp. 310–318). Cham: Springer.
16. Agnes, S. A., Anitha, J., & Solomon, A. A. (2022). Two-stage lung nodule detection framework using enhanced UNet and convolutional LSTM networks in CT images. Computers in Biology and Medicine, 149, 106059.
17. Sun, L., Zhang, M., Lu, Y., Zhu, W., Yi, Y., & Yan, F. (2024). NoduleCLIP: Lung nodule classification based on multi-modal contrastive learning. Computers in Biology and Medicine, 175, 108505.
18. Abdulqader, A. F., Abdulameer, S., Bishoyi, A. K., Yadav, A., Rekha, M. M., Kundlas, M., ... & Farhood, B. (2025). Multi-objective deep learning for lung cancer detection in CT images: enhancements in tumor classification, localization, and diagnostic efficiency. Discover Oncology, 16(1), 529.
19. Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems, 34, 3965–3977.
20. Chen, Y., Wan, Y., & Pan, F. (2023). Enhancing multi-disease diagnosis of chest X-rays with advanced deep-learning networks in real-world data. Journal of Digital Imaging, 36(4), 1332–1347.
21. Jena, S. R., & George, S. T. (2020). Morphological feature extraction and KNG-CNN classification of CT images for early lung cancer detection. International Journal of Imaging Systems and Technology, 30(4), 1324– 1336.
22. Jena, S. R., George, S. T., & Ponraj, D. N. (2021). Modeling an effectual multi-section You Only Look Once for enhancing lung cancer prediction. International Journal of Imaging Systems and Technology, 31(4), 2144–2157.
23. Jena, S. R., Avaiya, U., Kumar, U., & Oruganti, S. K. (2025). TransResNet: A Dual-Stream Deep Model for Precision Pulmonary Lesion Classification. SGS-Engineering & Sciences, 1(2).
24. Jena, S. R. (2025). LLM in Personalized medicine for Lung Cancer Detection. SGS-Engineering & Sciences, 1(1).
25. Jena, S. R., George, S. T., & Ponraj, D. N. (2021). Lung cancer detection and classification with DGMM-RBCNN technique. Neural Computing and Applications, 33(22), 15601–15617.
26. Ahmed, S. T., Barua, S., Fahim-Ul-Islam, M., & Chakrabarty, A. (2024, May). CoAtNet-Lite: Advancing Mammogram Mass Detection Through Lightweight CNN-Transformer Fusion with Attention Mapping. In 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) (pp. 143–148). IEEE.