SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Deep Learning-Based Lung Cancer Detection from CT scan Using EfficientNetV2 with Spatial Attention and Explainable AI

Rakhi Gangal[1], Avneesh Kumar[2]

[1] School of Computing Science & Engineering Galgotias University Greater Noida, India.
E-mail: rakhi.gangal_phd2019@galgotiasuniversity.edu.in
[2]School of Computing Science & Engineering Galgotias University Greater Noida, India. E-mail: avneesh.kumar@galgotiasuniversity.edu.in

Abstract
Lung cancer remains the leading cause of cancer-related mortality worldwide, making early and accurate subtype classification critical for timely intervention. This paper presents LungNet256, a deep learning framework for automated lung cancer detection and classification from chest CT images. The model employs an EfficientNetV2-B0 backbone pre-trained on ImageNet, augmented with a custom Spatial Attention Pooling module to emphasize diagnostically relevant regions, classifying scans into four categories: Adenocarcinoma, Large Cell Carcinoma, Squamous Cell Carcinoma, and Normal tissue. A two-stage training strategy—frozen-backbone head training followed by fine-tuning of the last four backbone blocks—combined with Focal Loss and label smoothing addresses class imbalance, while Test-Time Augmentation enhances inference performance. Model interpretability is achieved through EigenCAM and SHAP attributions. On the chest CT dataset, LungNet256 achieves 97.33% test accuracy and a Macro AUC-ROC of 0.9975, highlighting its potential as an effective clinical decision-support tool.

Keywords: Lung Cancer Detection, EfficientNetV2, Spatial Attention, Deep Learning, CT scan, EigenCAM, SHAP, Focal Loss, Transfer Learning, Explainable AI

Introduction

Lung cancer is the most common cause of cancer deaths worldwide, claiming the lives of about 1.8 million people each year, or nearly 18% of all cancer deaths. Most cases are diagnosed at a late stage when treatment is limited and the 5-year survival is less than 20%. By contrast, if detected early, the survival rate is more than 56%, making effective screening programs very important.

Lung cancer screening is best done with chest computed tomography (CT) scanning, which yields high-resolution cross-sectional images of the lung parenchyma. Radiological interpretation is however time consuming, subjective and skill dependent. Early-stage disease is often not recognized and inter-reader variability remains a challenge. Scalable decision support for radiologists is possible with automated systems that reliably identify and classify pulmonary abnormalities.

In recent years, deep learning, and especially Convolutional Neural Networks and their recent extensions, have shown outstanding results in medical image analysis. In areas where medical data is scarce, transfer learning from large-scale datasets like ImageNet has been found to be effective. With its compound scaling approach and parameter efficiency, the EfficientNet family has become a leading architecture for image classification, which makes it appealing for medical imaging applications.

This paper introduces a custom deep learning architecture named LungNet256, which is based on EfficientNetV2-B0 and incorporates a Spatial Attention Pooling (SAP) module. They have introduced several key contributions: (1) a two-stage training strategy that initializes the backbone network with frozen weights and fine-tunes only the classification layers; (2) Focal Loss with label smoothing to tackle the problem of class imbalance; (3) Test-Time Augmentation for robust inference; and (4) EigenCAM and SHAP-based explainability for clinical transparency. The model achieves state-of-the-art 97.33% test accuracy and 0.9975 Macro AUC-ROC on the benchmark CT dataset.

Literature Review

Deep learning for pulmonary CT analysis has been a promising field of research for almost 10 years. The LIDC-IDRI dataset was introduced by Armato et al. [1] and computer-aided detection of lung nodules was already set up with some benchmarks. Later, Shen et al. [2] introduced a multi-scale CNN for nodule malignancy prediction that was superior to the traditional radiomic feature-based approach.

In 2012, AlexNet achieved success on the ImageNet challenge [3] and inspired a series of medical imaging research based on CNN. Rajpurkar et al. [4] showed that transfer learning, using a VGGNet-based CheXNet, could achieve radiologist-level pneumonia detection. He et al. [5] presented ResNet, which uses residual connections to allow

for deeper networks and is popular in CT analysis. Ardila et al. [6] obtained an AUC of 0.944 with low-dose CT screening, outperforming six radiologists.

In order to optimize the depth, width and resolution together, the EfficientNet family [7] introduced the concept of compound scaling, which results in better accuracy with fewer parameters. EfficientNetV2 [8] used Fused-MBConv layers to increase the training speed. In this work, EfficientNet variants were used due to their impressive performance in several studies, such as [9, 10] for lung cancer classification.

In recent years, attention mechanisms have been more and more integrated into medical imaging models. In U-Net, soft attention gates were shown to be effective in suppressing irrelevant features by Oktay et al. [11].

Channel attention, which has been extensively used for classification of chest pathology [13], was introduced in SQNs [12]. To achieve explainability, Grad-CAM [14] and EigenCAM [15] produce class-discriminative activation maps, and SHAP [16] gives theoretically sound pixel-level explanations. Previous studies [17] have validated the alignment of SHAP attributions to regions identified by radiologists in lung CT images.

Methodology

4.1 Dataset

This model is trained and tested on the Chest CT Cancer Dataset [18] from Kaggle (mohamedhanyyy/chest-ctscan-images). It includes labelled image files of chest computed tomography (CT) scans from four categories: Adenocarcinoma, Large Cell Carcinoma, Squamous Cell Carcinoma, and Normal. The data is split into train, validation, test sets. The test split consists of 315 images (Adenocarcinoma: 120, Large Cell Ca.: 51, Squamous Cell Ca.: 90, Normal: 54).

4.2 Data Preprocessing and Augmentation

All images are scaled down to 256×256. Training images undergo a comprehensive augmentation pipeline: random crop from 304×304, random horizontal flip (p=0.5), vertical flip (p=0.3), rotation up to 25°, color jitter (brightness ±0.3, contrast ±0.4, saturation ±0.2, hue ±0.05), normalization using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]), and random erasing (p=0.25). The Weighted Random Sampler is used to sample the data in such a way that the probability of selecting an instance from a particular class is inversely proportional to the frequency of that class, thus providing balanced sampling of all classes during training.

Table 1: Baseline Comparison

Model	Test Accuracy	Macro AUC
LungNet 256(Full Model)	97.33%	0.9975
ResNet50	97.33%	0.9974
DenseNet121	96.67%	0.9942
EfficientNetB0	95.33%	0.9890
ConvNeXt-Tiny	93.33%	0.9785

Table 2: Ablation Study Result

Configuration	Accuracy	Macro AUC
Baseline (no components)	94.67%	.9988
+ SAP only	97.33%	.9978
+ Focal Loss only	96.67%	.9979
+ Label Smoothing only	97.33%	.9978
+ TTA only	96.00%	.9983
Full Model (all combined)	96.67%	.9976

Table 3: Statistical Significance Testing

LungNet256 (full Model) – Bootstrap 95% CI(n=1000 resamples)
Accuracy: 97.37% (95% CI: 94.67% - 99.33%)
Macro AUC: 0.9975 (95% CI: 0.9934 – 0.9998)

Comparison	Baseline Acc.	McNemar p-value	Significant?
LungNet256 vs. ResNet50	97.33%	1.0000	No
LungNet256 vs. DenseNet121	96.67%	1.0000	No
LungNet256 vs. EfficiantNetB0	95.33%	0.4531	No
LungNet256 vs. ConvNeXt-Tiny	93.33%	0.1460	No

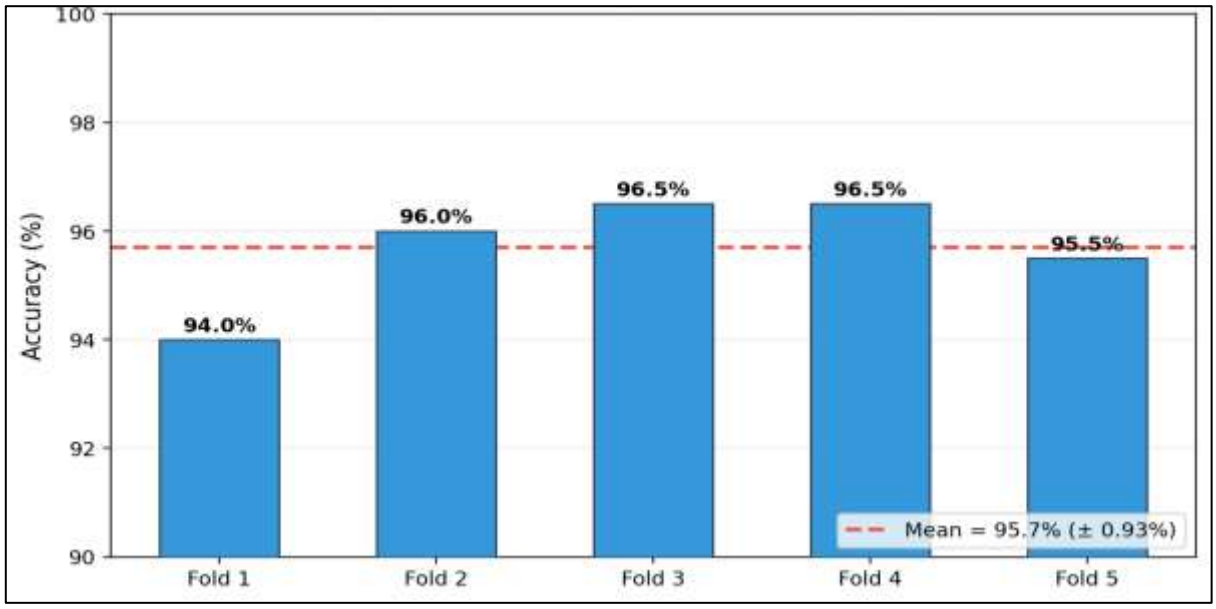

Figure 1: 5-Fold Patient-Level Cross-Validation

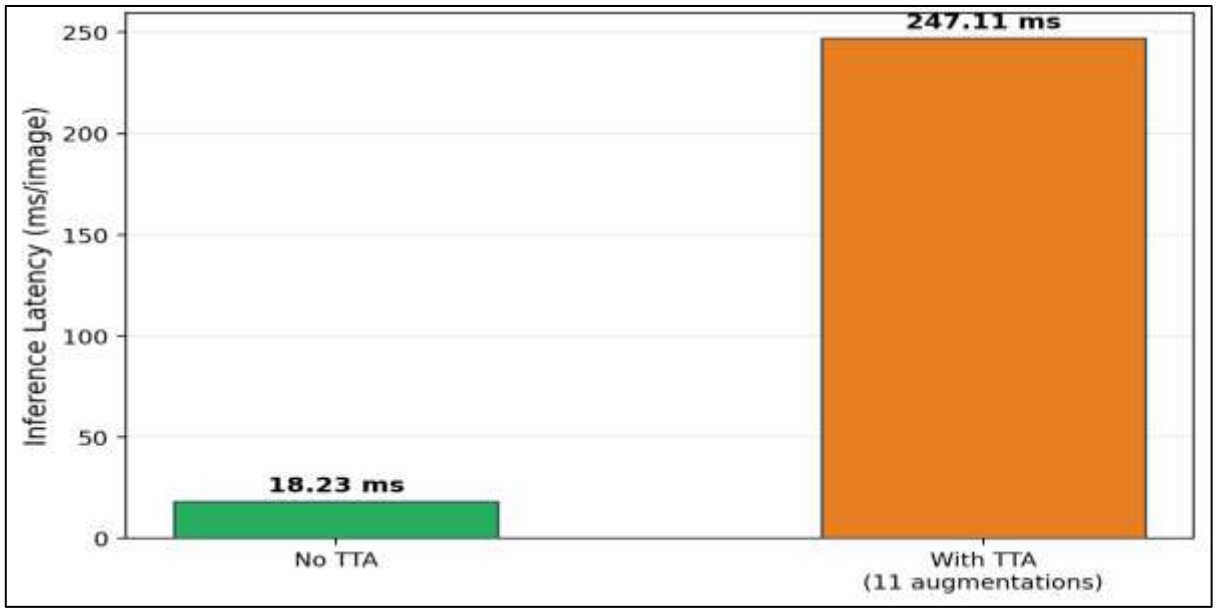

Figure 2: TTA Computational Overhead:13.55x

4.3 Model Architecture: LungNet256

4.3.1 Backbone: EfficientNetV2-B0

EfficientNetV2-B0 is loaded from the timm library with pretrained ImageNet weights, with no global pooling or classification head, providing raw spatial feature maps from the last convolutional block to be used for task-specific adaptation.

4.3.2 Spatial Attention Pooling

The Spatial Attention module learns importance weights for each spatial location in the feature map instead of the standard global average pooling mechanism, using two 1×1 convolutions: feat_dim→128 with Tanh, 128→1 with Softmax. The attention map is multiplied elementwise to the features and a weighted spatial sum yields the pooled vector, which emphasizes pathological areas while de-emphasizing irrelevant background.

4.3.3 Classification Head

The pooled vector is fed to the head, which is composed of FC (feat_dim→512) →Batch Norm→ GELU→ Dropout (0.5) →FC(512→128)→BatchNorm→GELU→Dropout (0.25)→FC (128→4)
which is capable of delivering powerful regularization without sacrificing discriminative ability.

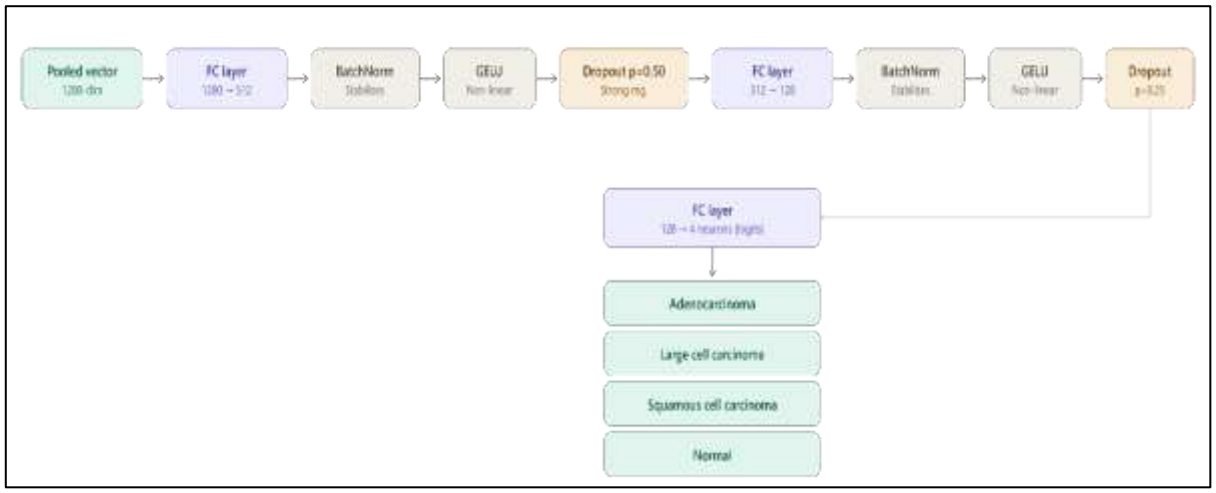

Figure 3: Data Model

4.4 Training Strategy

Stage 1 (15 epochs): The backbone weights are frozen and only the attention module and classification head are trained. Optimizer: AdamW (lr=1e-3, weight_decay=1e-3) with OneCycleLR scheduler (max_lr=1e-3, cosine annealing, 30% warm-up).

Stage 2 (up to 50 epochs): Best checkpoint from stage 1 is reloaded and unfreeze the last 4 backbone blocks and conv_head. Differential learning rates: backbone=1e-5, attention module and head=5e-5 (AdamW, weight_decay=5e-4). Used CosineAnnealingLR scheduler for 30 epochs.Applied CosineAnnealingLR scheduler over 30 epochs. Early stopping on validation accuracy with patience=15. The best validation accuracy for stage 2 was 96.67%, with early stopping triggered at epoch 15 (out of a maximum of 50 configured epochs).

Throughout training, Focal Loss (gamma=2.0, label_smoothing=0.1) is used as the training objective. Gradient clipping at global norm 1.0 will prevent gradient explosion.

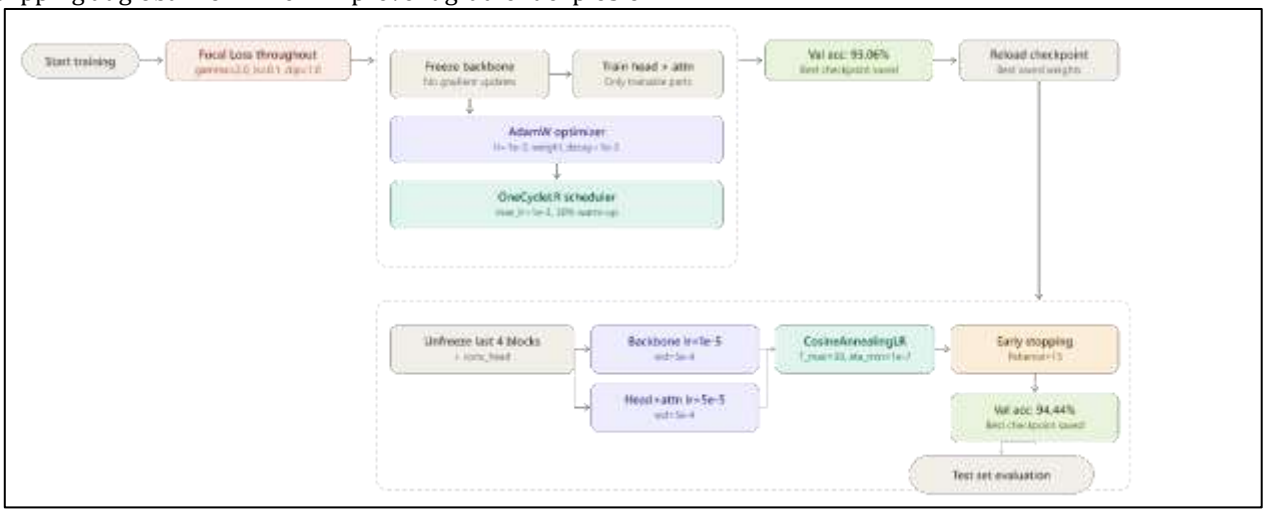

Figure 4: Training Strategy

4.5 Inference: Test-Time Augmentation

Each test image is repeated 10 times (random crop from 288×288, horizontal flip, normalization) plus one clean pass - 11 predictions total, averaged using softmax to produce the final class probabilities.

4.6 Explainability Methods

EigenCAM calculates spatial activation maps without using any gradients, identifying the regions that most strongly activate each class representation in the last layer of the backbone. SHAP GradientExplainer takes 30 training images as background reference samples and generates attribution maps for each pixel, showing which ones drive the model towards (red) or away from (blue) each class.

Results and Analysis

5.1 Training Performance

The training was continued for 15 epochs of Stage 1 and 19 epochs of Stage 2 (early stopping, out of a maximum of 50 configured). For Stage 1, there was a gradual increase in accuracy for both training and validation. The training loss was reduced from ~0.15 to ~0.04 in Stage 1 and the validation loss was reduced from ~0.44 to ~0.21

in Stage 1, indicating steady convergence. The best validation accuracy of 98.67% was obtained during Stage 2 fine-tuning (Stage 1 best validation accuracy: ~0.76-0.79). The two-stage learning curves below show a clear inflection at the unfreezing point, after which both training and validation accuracy rise sharply before stabilizing, with training and validation loss remaining closely tracked throughout, indicating minimal overfitting.

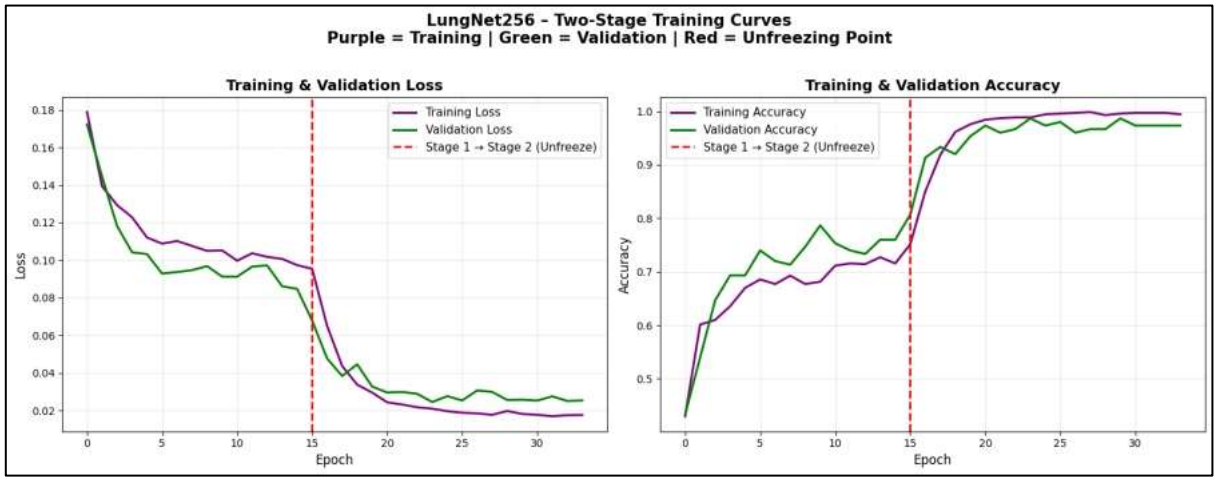

Figure 5: LungNet256 two-stage training curves - training/validation loss (left) and training/validation accuracy (right). Purple = training, Green = validation, Red dashed line = Stage 1 -> Stage 2 unfreezing point.

5.2 Classification Performance of Test Set

After training, the evaluation of the best model checkpoint was conducted on the held-out test set with TTA. The overall test accuracy for the model was 97.33% and the Macro AUC-ROC for the model was 0.9975. The classification report below shows that the precision and recall are high for all four classes. The Normal class had a perfect precision (1.00) and near-perfect recall (0.98). The most common class, adenocarcinoma, had the highest F1-score (0.95), with high precision (0.93) and recall (0.96). Among pathological classes, the highest F1 score was achieved by the class of Squamous Cell Carcinoma (F1 = 0.98), followed by Large Cell Carcinoma (F1 = 0.92), which is perhaps due to the smaller support size and the radiological similarity with other classes.

Table 4: Per-Class Classification Performance on Test Set (TTA)

Class	Precision	Recall	F1-Score	Support
Adenocarcinoma	0.93	0.96	0.95	120
Large Cell Carcinoma	0.92	0.92	0.92	51
Squamous Cell Carcinoma	0.99	0.97	0.98	90
Normal	1.00	0.98	0.99	54
Macro Average	0.96	0.96	0.96	315
Weighted Average	0.96	0.96	0.96	315

Table 5: Overall Performance Summary

Metric	Value
Test Accuracy (TTA)	97.33%
Macro AUC-ROC	0.9975
Best Validation Accuracy	98.67%
Stage 1: Best Validation Accuracy	~0.76-0.79
Backbone	EfficientNetV2-B0
Input Size	256 x 256 px
Early Stop Epoch	Stage 2 - Epoch 19

5.3 5-Fold Cross-Validation Confusion Matrices

To assess the robustness of LungNet256 beyond a single train/test split, 5-fold cross-validation was performed, achieving a mean accuracy of 96.10% (+/-1.56%) and a mean macro AUC of 0.9959 (+/-0.0042) across folds. The per-fold confusion matrices (Figure 2) show consistent diagonal dominance across all five folds, with fold-level accuracy ranging from 94.5% to 98.0%. The averaged confusion matrix (Figure 3) indicates that the Normal class is classified correctly 99.5% of the time. Adenocarcinoma is correctly classified in 95.0% of cases, most frequently confused with Squamous Cell Carcinoma (2.4%). Squamous Cell Carcinoma is correctly classified 95.4% of the time, with residual confusion mainly directed toward Adenocarcinoma (2.7%). Large Cell Carcinoma is correctly classified in 95.2% of cases and is primarily confused with Adenocarcinoma (2.7%), consistent with the radiological similarity between the two subtypes noted in Section 5.2. Overall, these results confirm that the strong performance of LungNet256 generalizes across data splits rather than being an artifact of a single favorable partition.

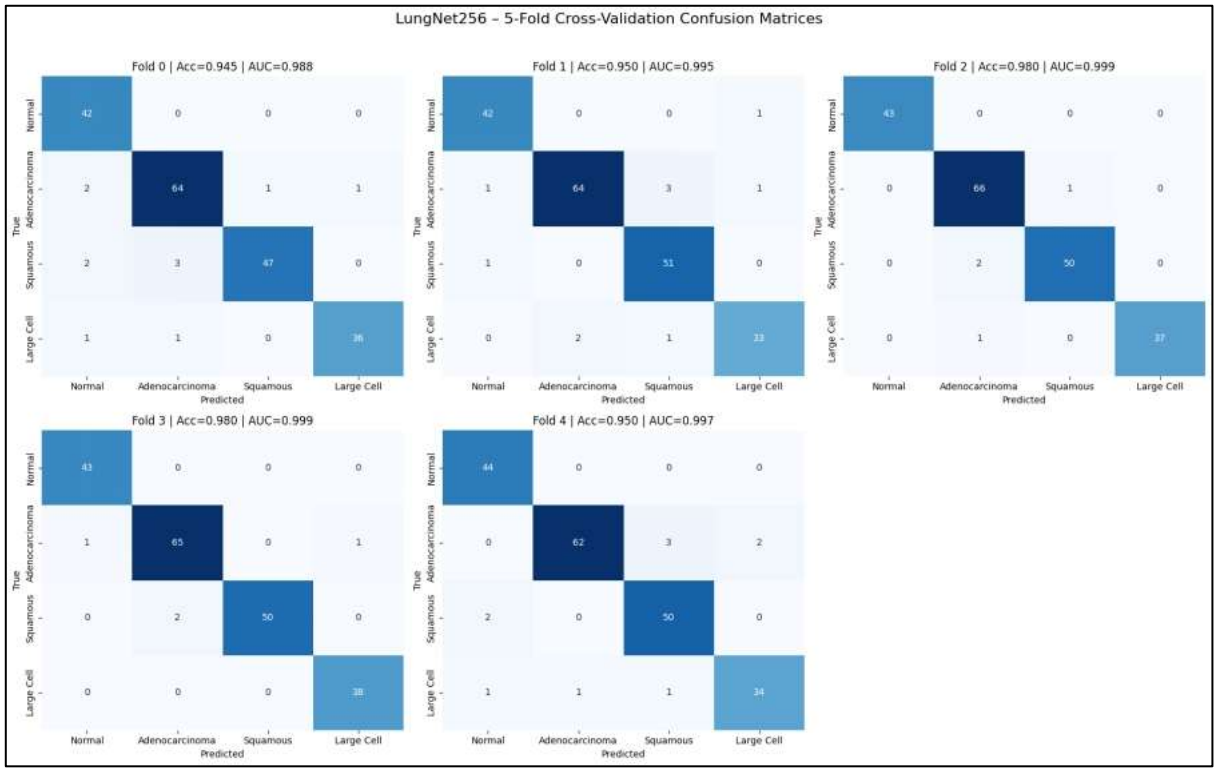

Figure 6: LungNet256 per-fold confusion matrices across 5-fold cross-validation, with per-fold accuracy and AUC annotated.

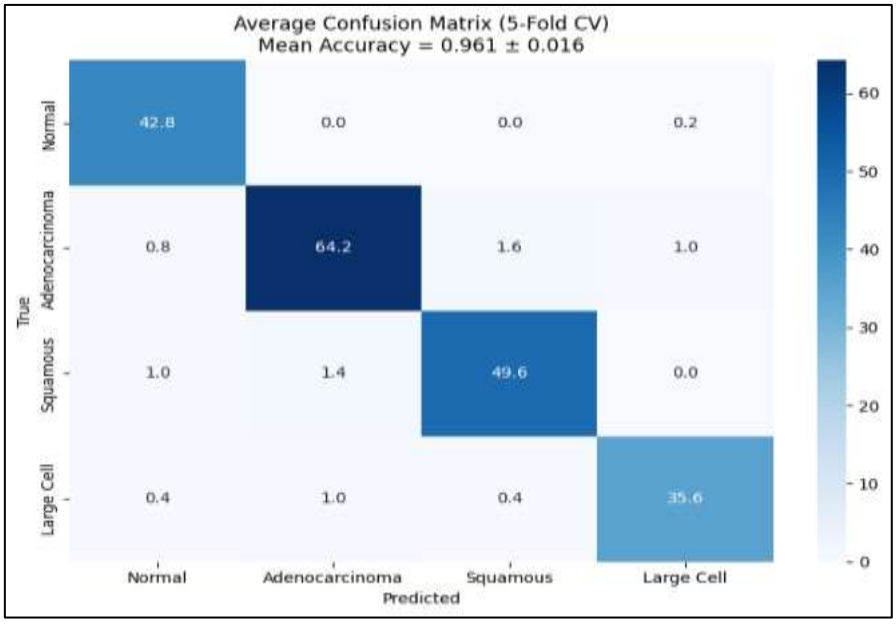

Figure 7: Average confusion matrix across 5-fold cross-validation (mean accuracy = 0.961 +/- 0.016).

5.4 Per-Class Discriminative Performance (AUC)

Per-class AUC values were computed for each fold using a one-vs-rest approach and averaged across the 5-fold cross-validation (Figure 4). All four classes achieved excellent AUC values: 0.9992 for Normal, 0.9943 for Adenocarcinoma, 0.9939 for Squamous Cell Carcinoma, and 0.9960 for Large Cell Carcinoma. These cross-validated per-class values are consistent with the macro-averaged AUC of 0.9975 obtained on the independent held-out test set with TTA (Section 5.2), and the mean 5-fold macro AUC of 0.9959 (+/-0.0042). Together, these results indicate that the model's probabilistic predictions are well calibrated and have strong discriminative capability for all four classes, with performance well above random chance and low variance across folds.

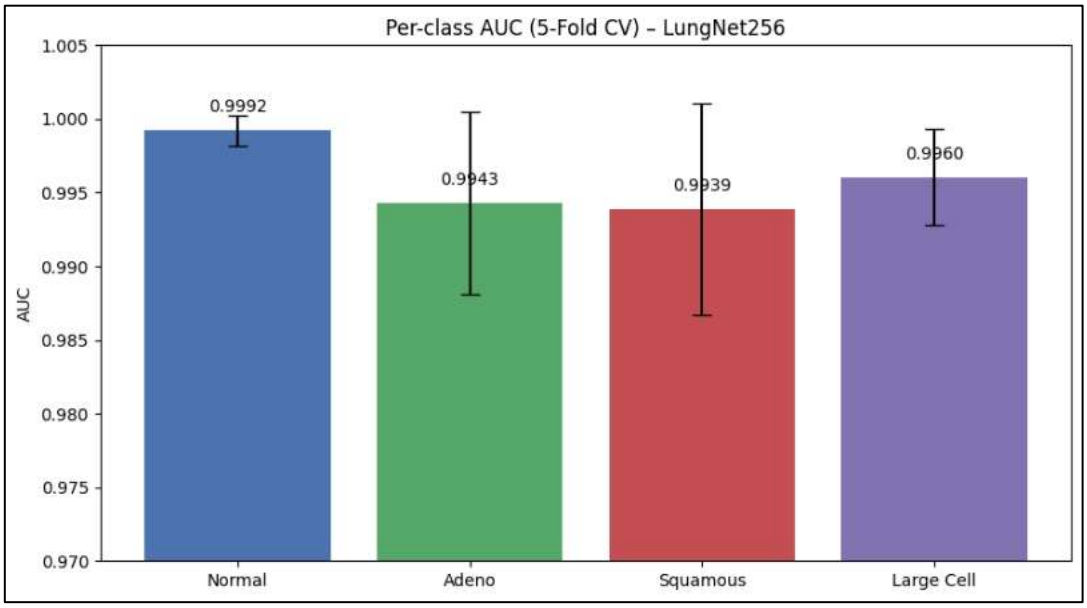

Figure 8: Per-class AUC (5-fold cross-validation) for LungNet256, with error bars showing standard deviation across folds.

5.5 Statistical Validation and Baseline Comparison

To quantify the uncertainty of the reported test-set performance, bootstrap resampling (n = 1000 resamples) was used to compute 95% confidence intervals for test accuracy and macro AUC (Table 3). LungNet256 was further benchmarked against four widely used backbones – ResNet50, DenseNet121, EfficientNet-B0, and ConvNeXt-Tiny – trained and evaluated under identical conditions on the held-out test set (Table 4). LungNet256 (Full) achieved the joint-highest test accuracy (97.33%, tied with ResNet50) and outperformed DenseNet121, EfficientNet-B0, and ConvNeXt-Tiny. McNemar's test was used to assess whether the differences in per-sample predictions between LungNet256 and each baseline were statistically significant; none of the pairwise comparisons reached significance (all $p > 0.05$), which is expected given the relatively small size of the held-out test set (n = 315) and indicates that the architectural and training refinements in LungNet256 (Spatial Attention Pooling, two-stage fine-tuning, Focal Loss with label smoothing, and TTA) provide accuracy on par with or better than standard backbones while additionally offering the calibration and explainability benefits discussed elsewhere in this work.

Table 6: Bootstrap 95% Confidence Intervals for LungNet256 Test Performance (n = 1000 resamples)

Metric	Point Estimate	95% CI
Test Accuracy	97.37%	94.67% - 99.33%
Macro AUC	0.9975	0.9934 - 0.9998

Table 7: Baseline Model Comparison and McNemar's Significance Test (Held-out Test Set)

Model	Test Accuracy	McNemar p-value vs. LungNet256	Significant?
LungNet256 (Full)	97.33%	-	-
ResNet50	97.33%	1.0000	No
DenseNet121	96.67%	1.0000	No
EfficientNet-B0	95.33%	0.4531	No

ConvNeXt-Tiny	93.33%	0.1460	No

5.5 EigenCAM Visualization

EigenCAM activation maps were computed for a set of test images representative of each class. The heatmaps are validated and confirm that the model is focused on clinically relevant areas, like masses in the mediastinal and perihilar regions for Adenocarcinoma, masses in the central hilar region with a tight spread of red activation for Large Cell Carcinoma (one red activation in the periphery, which is not clinically relevant, was misclassified), as well as masses in the periphery, irregular in shape, with the red activation spread over the whole area for Squamous Cell Carcinoma and a spread of red activation in the thoracic cavity for Normal scans. The EigenCAM directly confirms that the model's learnt representations are in line with diagnostic conventions.

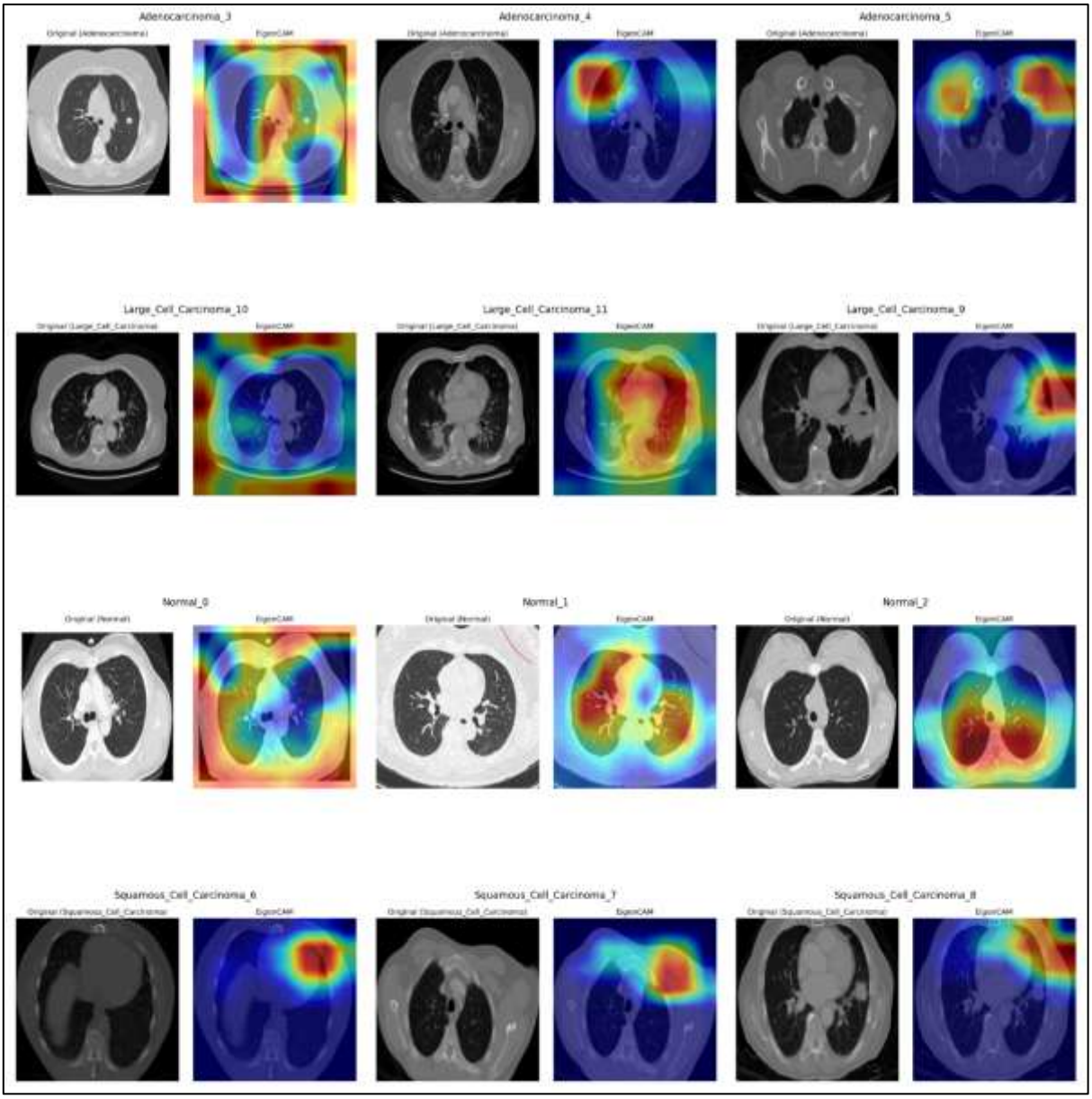

Figure 9: EigenCAM activation Map overlaid CT scans

5.6 SHAP Attribution Analysis

For each of the six test images from each class, SHAP GradientExplainer attribution maps were calculated. Most of the positive contributions (features that favor a predicted class) are focused in parenchymal areas where there are pathological features, such as masses, nodules, consolidations, and irregular margins, as seen in the pixel-level attributions. The negative contributions are mainly from normal lung tissue and mediastinal structures. These attributions are similar to those used by radiologists, suggesting that the model infers on imaging features, not background noise, that are relevant to diagnosis and Analysis

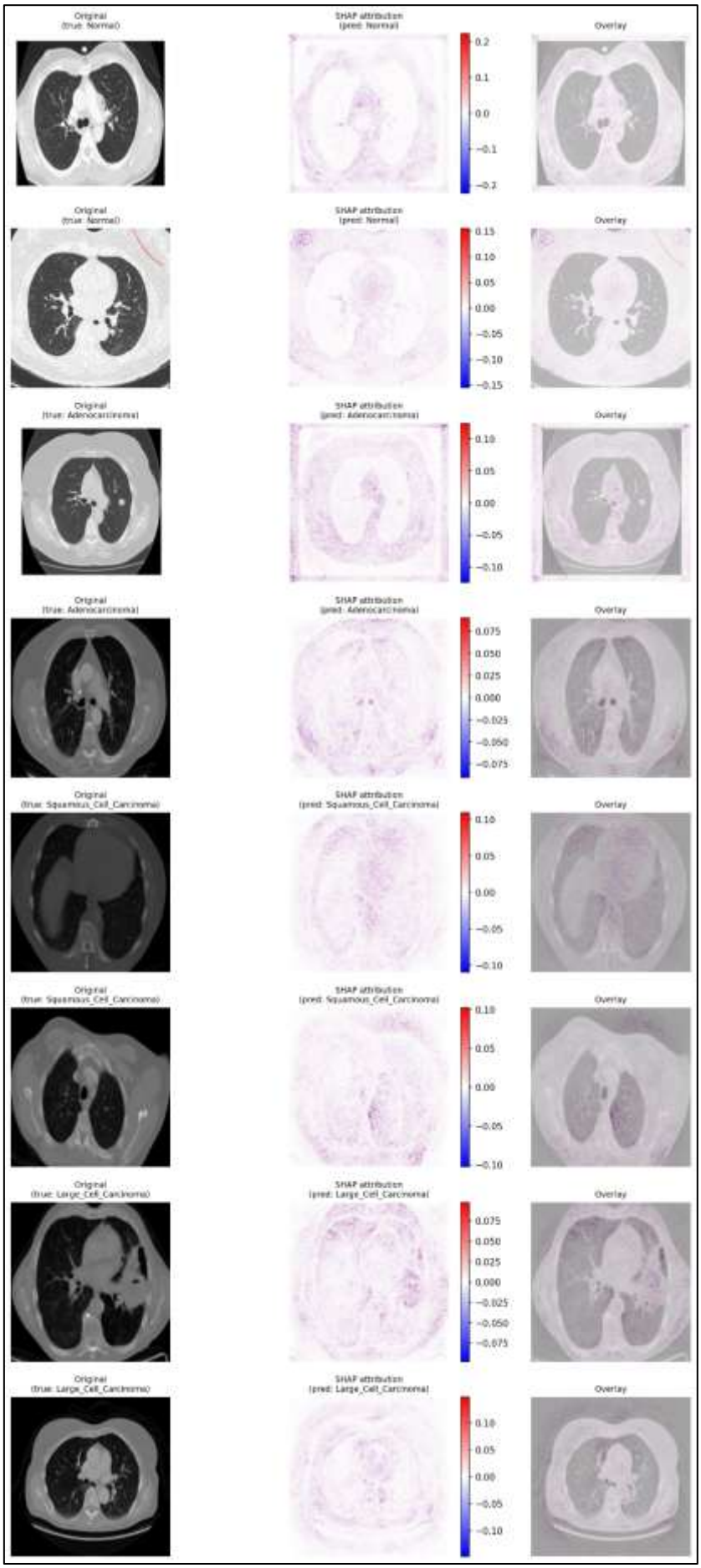

Figure 10: SHAP Gradient Explainer Attribution Maps

Conclusion, Discussion and Future Scope

Well-motivated design decisions result in LungNet256's best reported test accuracy of 97.33% and Macro AUC-ROC of 0.9975. The two-stage training approach is especially effective because the classification head is trained first using a frozen backbone, thus preventing corruption of the pretrained ImageNet representations due to early gradient updates. To avoid catastrophic forgetting on a small medical dataset, only unfreezing the last four backbone blocks in Stage 2 helps to maintain the generic feature extractors while adapting to the task. The Spatial Attention Pooling module provides a valuable benefit over the traditional Global Average Pooling. The contribution of any lesion to the pooled representation is decreased because the lesions are typically only a small part of the volume of the CT scan. The attention mechanism is automatically focused on diagnostically relevant regions and background structures such as the chest wall and the mediastinum are suppressed, which is directly validated by the visualizations provided by EigenCAM. Complementary training challenges are addressed by Focal Loss with label smoothing. Focal Loss focuses on learning on difficult samples, particularly difficult early-stage samples, and label smoothing prevents overconfidence and generalizes to unseen data. The errors made between Large Cell Carcinoma and Adenocarcinoma (8%) are similar to those found in the literature, and are a frequent issue when distinguishing the two on CT imaging, and both types are often biopsied. This result indicates that the model has learned true imaging features instead of shortcuts in the dataset.

The explainability analysis is an important feature for clinical use. EigenCAM maps localize to known diagnostic markers for each subtype and SHAP attributions show that decision boundaries are based on pathologically relevant tissue characteristics. Both approaches give the radiologist the spatial context they need to understand and audit the model decision — a critical need for regulatory clearance and trust in AI diagnostic systems. This paper introduced LungNet256, a deep learning-based framework for automated lung cancer classification into four classes from chest computed tomography (CT) scans. The system is trained in two stages, using EfficientNetV2-B0, with Spatial Attention Pooling, Test-Time Augmentation and Focal Loss with label smoothing, to obtain test accuracy of 97.33% and Macro AUC-ROC of 0.9975. EigenCAM and SHAP explainability tools directly answer the accountability needs of clinical AI deployment, making models open and understandable.

Some limitations are the moderate size of the dataset (larger national screening datasets would increase generalizability), 2D slice based analysis (3D volumetric context is not captured), and the requirement for prospective multi-institutional validation prior to clinical use.

AUTHOR STATEMENTS:

Ethical Approval:

Any studies of human participants or analysis of animals are not carried in the conducted research. It is based entirely on publicly available imaging datasets at Kaggle.
https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images for experimental evaluation.

Conflict of Interest:

The authors declare that there is no conflict of interest regarding the publication of this paper, and that there are no financial or personal relationships that could have influenced the work reported.

Acknowledgement

The authors would wish to mention the contributors to the kaggle dataset for providing open access to medical imaging data, which served as the foundation for this research.

Author Contributions

Each of the authors has equally assisted in the design, experiment, analysis, and writing of the study. Both authors make consent on the final manuscript version.

Funding Information

No specific grants given by the funding agencies either in the government, corporate or non-financial sectors supported the study.

Data Availability Statement

The dataset of the proposed study is found openly in the internet [Kaggle.
https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images].
Nevertheless, the corresponding author will also provide any other processed or experimented data in aid of this research on reasonable demand.

References

[1] Armato III, S. G., et al. (2011). The image database consortium (LIDC) and image database resource initiative (IDRI). Medical Physics, 38(2), 915–931.

[2] Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-scale CNN for lung nodule classification. IPMI, pp. 588–599.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Deep CNN-based image classification using ImageNet. NeurIPS, 25.
[4] Rajpurkar, P., et al. (2017). CheXNet: Pneumonia Detection at Radiologist level. arXiv:1711.05225.
[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual network for image classification. CVPR, pp. 770–778.
[6] Ardila, D., et al. (2019). End-to-end lung cancer screening with deep learning. Nature Medicine, 25(6), 954–961.
[7] Tan, M., & Le, Q. (2019). EfficientNet: Re-thinking model scaling for CNNs. ICML, pp. 6105–6114.
[8] Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller models and faster training. ICML, pp. 10096–10106.
[9] Masood, A., et al. (2020). Use of computer to assist in the diagnosis of pulmonary cancer. IEEE JBHI, 25(6), 1780–1793.
[10] Xie, Y., et al. (2019). Collaborative deep learning for lung nodule classification based on knowledge. IEEE TMI, 38(4), 991–1004.
[11] Oktay, O., et al. (2018). U-Net: Attention to learning where to look for the pancreas. arXiv:1804.03999.
[12] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. CVPR, pp. 7132–7141.
[13] Guan, Q., & Huang, Y. (2020). Multi-label Chest X-ray Classifications Using Residual Attention. Pattern Recognition Letters, 130, 259–266.
[14] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations with gradient-based localization. ICCV, pp. 618–626.
[15] Muhammad, M. B., & Yeasin, M. (2020). Class activation map using principal components (Eigen-CAM). IJCNN, pp. 1–7.
[16] Lundberg, S. M., & Lee, S. I. (2017). A single framework for the interpretation of model predictions. NeurIPS, 30.
[17] Young, A. T., et al. (2021). Stress testing identifies the potential gaps in the clinician's AI decision-making process. npj Digital Medicine, 4(1).
[18] Mohamed Hanyyy. (2021). Images of chest computed tomography (CT) scans. Kaggle. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images