Evaluating XAI Localization Fidelity on Chest X-Ray Images

Maharaja Surajmal Institute of Technology

Affiliated to GGSIPU | NAAC Accredited 'A' Grade | NBA (CSE, IT, ECE,EEE) | Approved by AICTE | ISO 9001:2015 Certified

Student Name(s):

- Aditya Singh Rawat (20496303122)
- Karan Kohli (20396303122)
- Vipul Gupta (20296303122)

Mentor Name: Dr. Sunesh Malik

Date: 3.12.2025

CONTENTS

I. Introduction

The project is situated in the domain of Explainable Artificial Intelligence (XAI) for Medical Image Classification. It focuses on leveraging advanced computational techniques to analyze medical images, specifically Chest X-ray (CXR) images, for the diagnosis of respiratory diseases.

- The integration of Deep Learning (DL) models in high-stakes environments, such as medical imaging analysis, faces a major hurdle known as the **"Black Box" problem**. These complex models impede direct interpretation of their predictions, which is a major barrier to clinical trust and adoption. When automated systems fail, they often do so "spectacularly disgracefully without warning or explanation," emphasizing the need for reliable explanations.

- **Mandate for Interpretability and Transparency: The Indian Council of Medical Research (ICMR) Ethical Guidelines for AI in Healthcare** (2023) mandate that AI systems must be understandable, interpretable, and transparent to clinicians and patients, defining the core requirement for XAI adoption in healthcare.

II. Problem Statement

COVID Classification Infection Mask Creation with LIME

Deep neural networks work well in classifying medical images however clinicians cannot justify the classification of the diagnosis.

We propose to address this "black-box" limitation by developing an **explainable AI model for medical image classification with infection mask generation**.

Specifically, we will use a models like **deep learning architectures** with XAI techniques like **LIME** (Local Interpretable Model-agnostic Explanations)

Combining the strengths of both worlds to enhance trust and transparency in the medical image diagnosis process. This will help both sides the doctors and the patients understand the diagnosis.

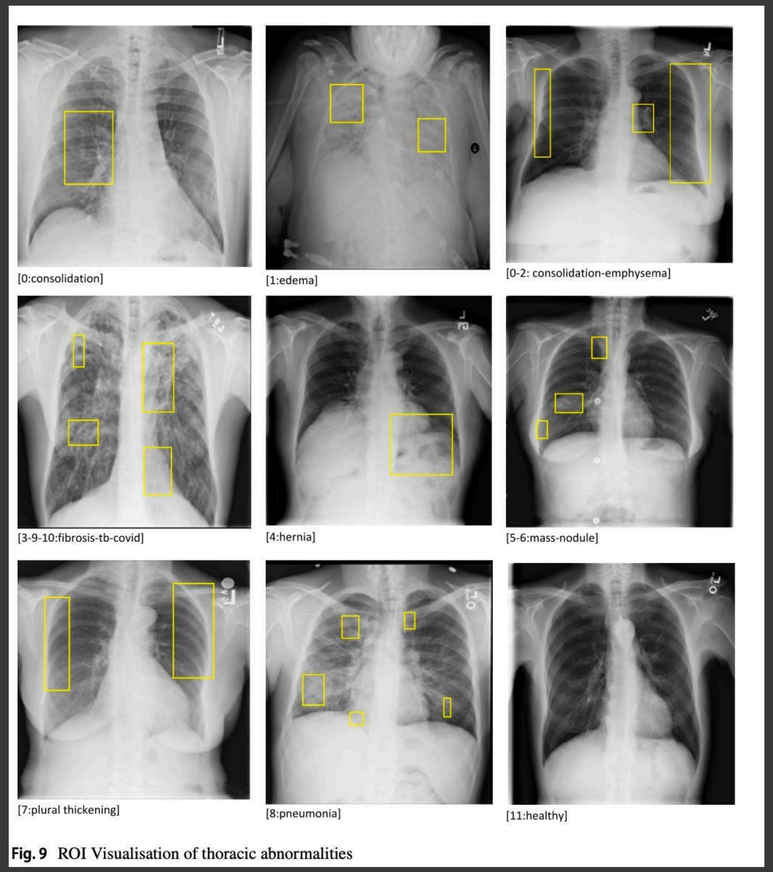


Fig. 9 ROI Visualisation of thoracic abnormalities

III. Objectives

01 : Make DL model

To develop and fine tune a deep learning model, capable of accurate binary and multi-class classification.

02 : Evaluate XAI

Applying LIME on the trained model to see how the model is deriving the classification results.

03 : Documentation

Documenting the progress and comparisons enabling us to evaluate which XAI technique works best for the task of medical Imaging

04 : Infection Mask

Using XAI techniques we seek to develop infection mask with unsupervised pipeline lowering cost of and increasing accuracy of these saliency maps.

IV. Literature Review

- **Deep Learning Models:** The most common approach involves Deep Convolutional Neural Networks (CNNs) such as VGG16, ResNet and custom architectures like COVID-Net. These models are trained to perform binary classification (e.g., COVID-19 vs. Normal) or multi-class classification on CXR and CT scan images. Many of these systems achieve high accuracy, with some reporting scores over 95%. [1][4]

- **Transfer Learning:** To improve performance and overcome the limitations of small medical datasets, many methods employ transfer learning. This involves using a model pre-trained on a large-scale dataset, such as ImageNet, and fine-tuning it for the specific medical classification task. [5][3]

- **Ensemble and Fusion Methods:** More advanced solutions use ensemble learning, which combines the predictions from multiple different models to enhance accuracy and robustness. Similarly, feature fusion techniques merge the features extracted by several CNNs to create a more powerful and discriminative feature set for the final classifier.[3]

- **Visual/Saliency-Based Methods:** The most frequently used techniques are such as saliency maps, Class Activation Mapping (CAM), and Gradient-weighted Class Activation Mapping (Grad-CAM).[2][4] These methods generate heatmaps that highlight the regions that were most influential in the model's decision-making process. [1][2][4]

- **Perturbation-Based Methods:** Local Interpretable Model-agnostic Explanations (LIME) is another widely adopted technique. LIME explains a single prediction by creating variations (perturbations) of the input image and learning a simpler, interpretable model (like a linear model) that approximates the complex model's behavior in that local vicinity.[6][7][8]

V. Methodology

In this update of the project we had three goals:

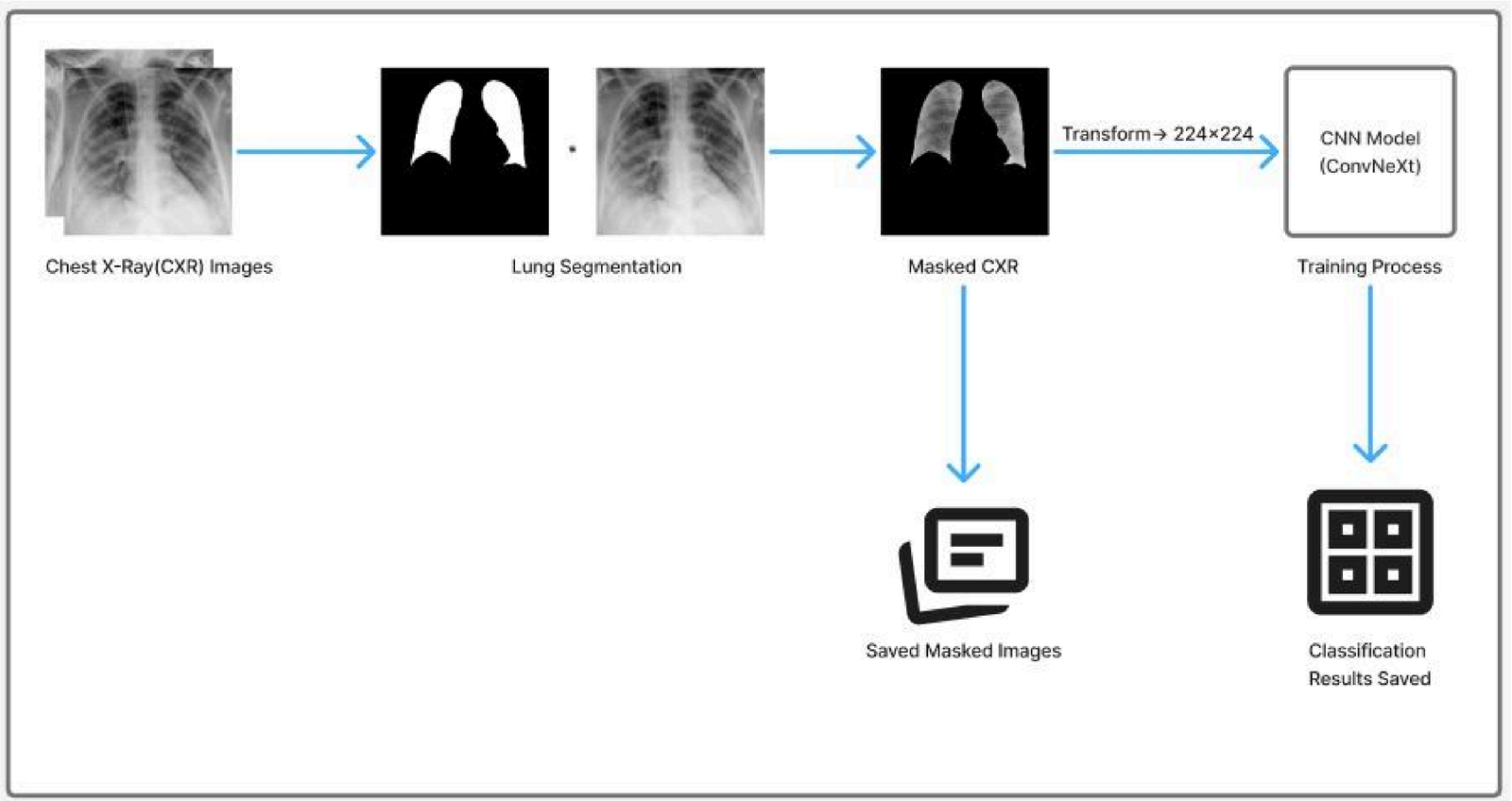
1. Push the accuracy of our model further from 91 using ConvNeXt and 2-Label Classification
2. Evaluate the ability of LIME and Grad-CAM to create infection masks from ConvNeXt Pre-trained model
3. Compare metrics of LIME and Grad-CAM proving LIME is better than Grad-CAM

VI. Technologies Used

01

ConvNeXt

ConvNeXt is a recent Convolutional Neural Network (CNN) architecture (2022) included in the evaluation specifically to drive novelty, as it was not commonly used in existing papers, particularly for localization tasks in medical imaging.

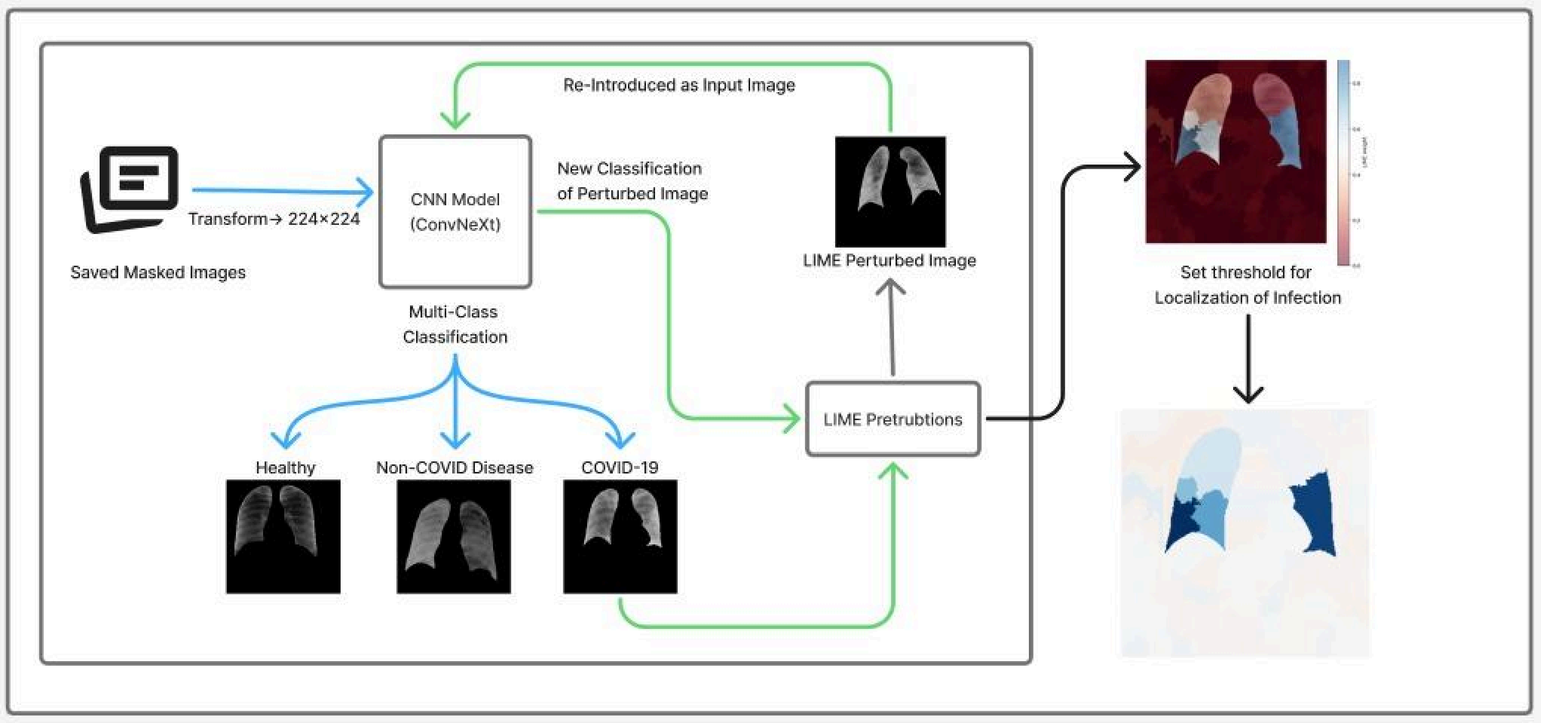


Workflow diagram of training ConvNeXt

VI. Technologies Used

02 LIME

LIME is a perturbation-based technique that seeks to explain predictions of any black-box classifier by locally approximating the complex model with a simpler, interpretable function over user-understandable components like "superpixels".



Workflow diagram of generating Infection Masks

VII. Progress

1. **Dataset:** COVID-QU-Ex
   This specific subset provides comprehensive ground-truth lung and infection segmentation masks. The distribution of images used for infection segmentation comprises

- **1,456 Normal cases**
- **1,457 Non-COVID-19 CXRs**
- **2,913 COVID-19 CXRs**
  with the COVID-19 cases containing corresponding infection masks (sourced from the QaTaCov19 dataset in conjunction with COVID-QU-Ex data).

Before

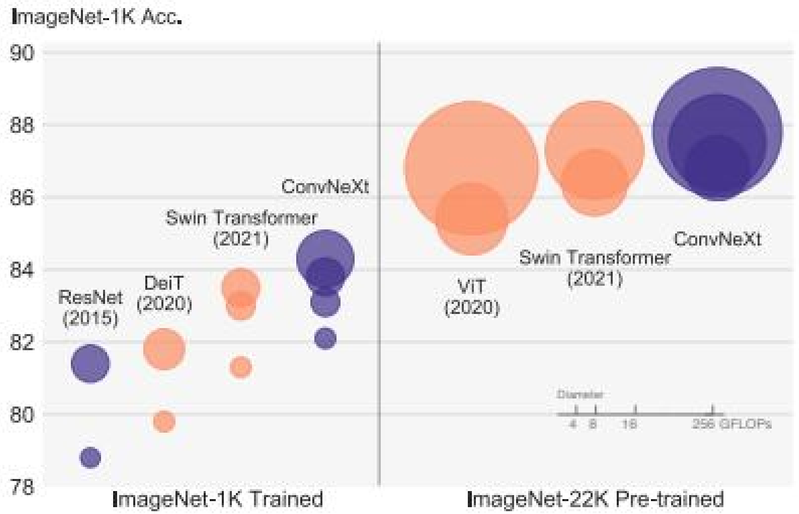
After

Stage 01 : Contributed by Aditya Singh Rawat

Classification with a Fine-Tuned ConvNeXt Model

Model Architecture: ConvNeXt is recognized as a modern Convolutional Neural Network (ConvNet) architecture, initially introduced in 2022. It was included in this project as a recent model to drive novelty and expand the understanding of explainability in the medical domain.

Architecturally, a succeeding family of models, ConvNeXt V2, was developed through the co-design and scaling of ConvNets with self-supervised learning techniques like Masked Autoencoders (MAE). This V2 model family proposes a fully convolutional masked autoencoder framework and introduces a new component: the Global Response Normalization (GRN) layer. The purpose of the GRN layer is to enhance inter-channel feature competition within the ConvNeXt architecture.



1. **Accuracy** is defined as the ratio of correctly classified elements among all the data, reflecting the overall correctness of the model's predictions.

$$Accuracy\ (for\ each\ class) = \frac{TP + TN}{TP+FP+TN+FN}$$

2. **Precision** is the rate of correctly classified positive class samples among all the members classified as positive samples.

$$Precision = \frac{TP}{TP+FP}$$

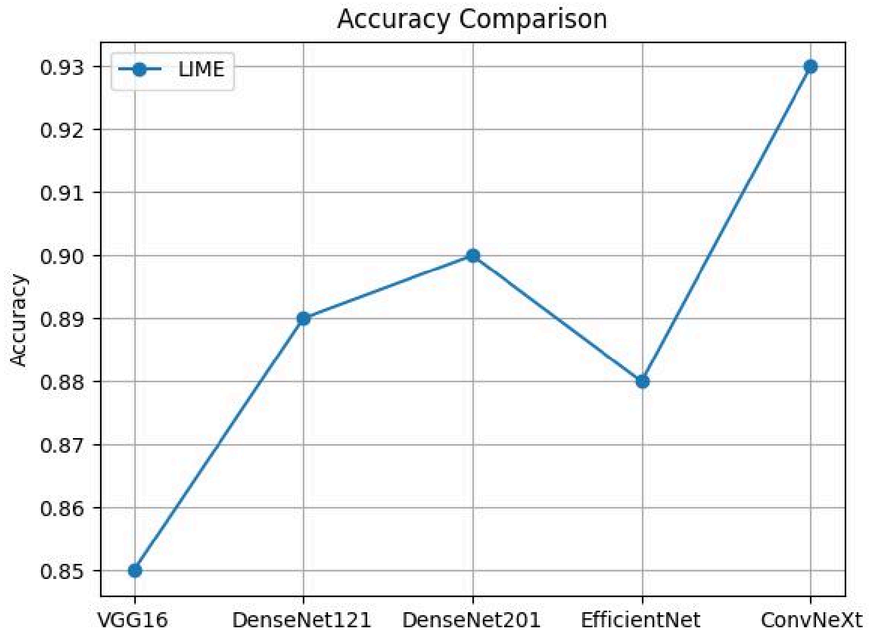
3. **Area Under the Curve (AUC)** is used to assess the model's overall capability to distinguish between positive and negative instances, and is derived from the Receiver Operating Characteristic (ROC) curve.

Stage 02 : Contributed by Aditya Singh Rawat

Classification with a Fine-Tuned DenseNet Model

Once the lung regions are segmented, the processed images will be fed into a deep learning classifier for diagnosis.

3-Label Classification: COVID, Non COVID Disease, and Healthy



Comparison of Classification Models

Model	VGG16	DenseNet121	DenseNet201	ConvNeXt [Base]	CoroNet [Khan et al.]	COVID-Net [Wang et. al.]
Parameters (millions)	130	8	20	89	33	11.75
Dataset Split	1,456: Healthy 1457: Non-COVID 2,913: COVID-19				500: Healthy, 500: Pneumonia 157: COVID-19	13,975 CXR images from 13,870 patients
Accuracy	0.85	0.89	0.90	**0.93**	0.902	0.93
Precision COVID	0.94	0.92	0.94	**0.97**	97	-
Precision (Non-COVID)	0.75	0.81	0.85	0.86	92	-
Precision (Healthy)	0.82	0.91	0.88	**0.91**	87	-
Precision (Average)	0.83	0.88	0.89	**0.91**	92	-
Macro-AUC	0.94	0.97	0.98	**0.99**	-	-

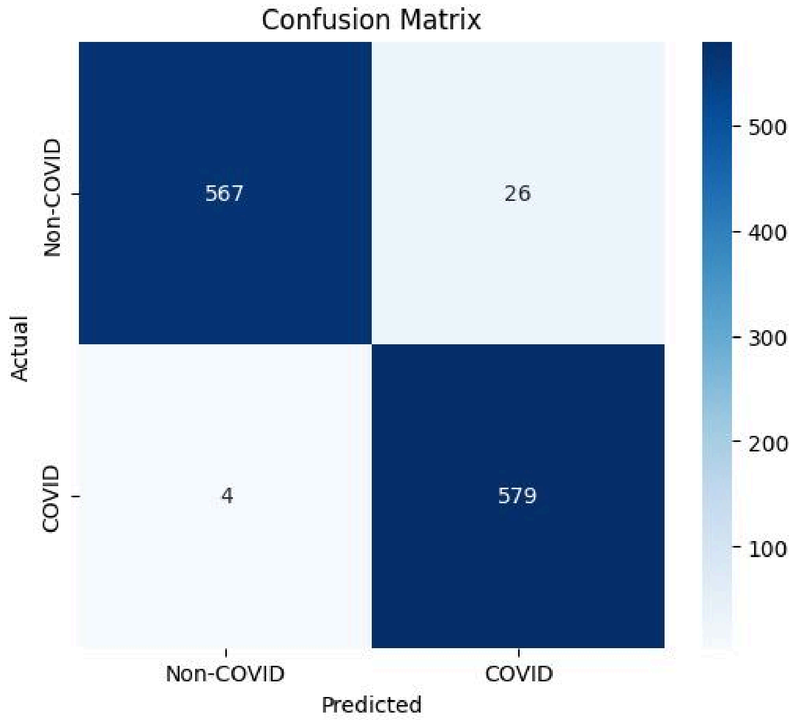
Table: 3 - Class Classification Results

Stage 02 : Contributed by Aditya Singh Rawat

Classification with a Fine-Tuned DenseNet Model

Once the lung regions are segmented, the processed images will be fed into a deep learning classifier for diagnosis.

2-Label Classification: COVID and Non COVID



Model	ConvNeXt	Inception-ResNetV2 [Narin A. et al.]	CoroNet [Khan et al.]
Accuracy	0.97	0.97	99
COVID (Precision)	**0.96**	-	0.97
Non-COVID (Precision)	**0.99**	-	1.00
Precision Average	**0.98**	0.82	98.3
Macro-AUC	**1.00**	-	-

Table: 2 - Class Classification Results

Stage 02 : Contributed by Karan Kohli

Processing with Grad-CAM to obtain Saliency Maps

Gradient-weighted Class Activation Mapping (Grad-CAM) is a popular and widely used gradient-based visualization method for Deep Neural Networks (DNNs). Its primary purpose is to produce 'visual explanations' for model decisions. Consequently, the visualizations produced by Grad-CAM are typically coarse approximations, leading to localization fidelity that is generally inferior to more precise, segmentation-based approaches or the human expert benchmark.

1. **Intersection over Union (IoU):** for a single class (pixel-wise classification) is defined based on True Positives (TP), False Positives (FP), and False Negatives (FN), which is used to evaluate the overlapping ratio:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{FP} + \mathrm{TP}}$$

2. **DICE Score (or Dice Similarity Coefficient, DSC)**: is often utilized to quantify the segmentation performance, and is equivalent to the F1-score for segmentation tasks

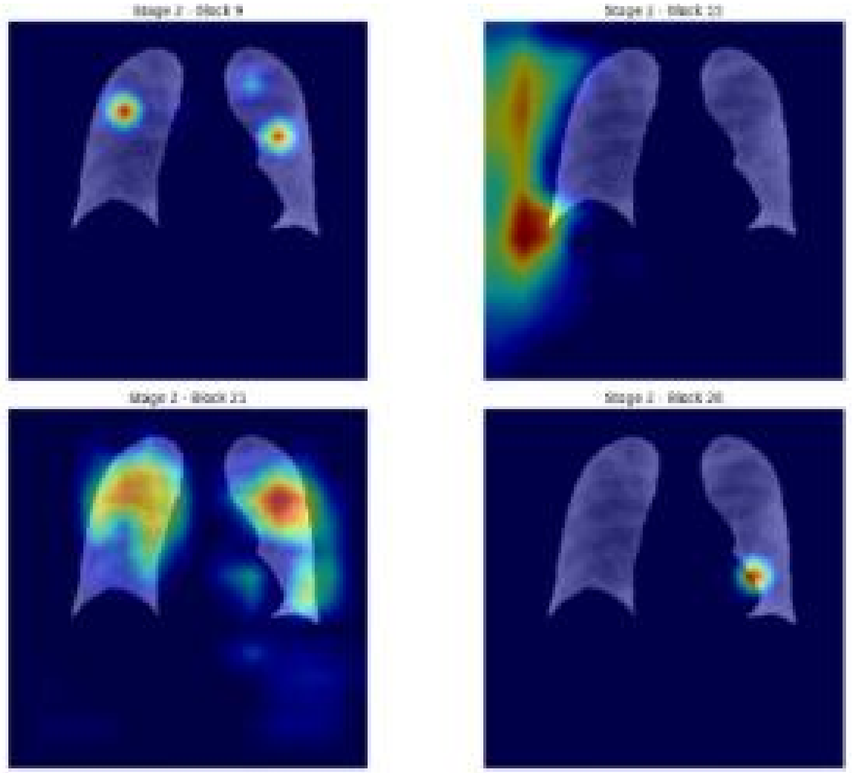
$$\mathrm{DC} = \frac{2|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y}| \cup |\hat{\mathbf{Y}}|}$$

3. **Hausdorff Distance:** measures how closely the generated boundary matches the true segmentation boundary. It is a distance metric, and therefore, a lower value indicates better performance.

$$HD(Gt, Pd) = mean_{p_p d \in Pd} min_{p_g t \in Gt} \left\| p_{gt} - p_{pd} \right\|^2$$

Where p_{pd} is predicted pixels and p_{gt} is ground truth pixels during segmentation.

Hausdorff distance is another metric that has been seeing increased usage in medical image segmentation (Karimi and Salcudean, 2019).



Picking the Correct Layer

Stage 02 : Contributed by Karan Kohli

Processing with Grad-CAM to obtain Saliency Maps

Gradient-weighted Class Activation Mapping (Grad-CAM) is a popular and widely used gradient-based visualization method for Deep Neural Networks (DNNs). Its primary purpose is to produce 'visual explanations' for model decisions.

While Grad-CAM is often preferred for approximate localization due to its relative computational efficiency and lower resource demands for data labeling, its reliance on coarse resolution features limits its ability to capture the geometric nuances of complex pathologies. Consequently, the visualizations produced by Grad-CAM are typically coarse approximations, leading to localization fidelity that is generally inferior to more precise, segmentation-based approaches or the human expert benchmark.

	LIME			Grad-CAM		
	DenseNet121 [8M]	DenseNet201 [20M]	**ConvNeXt Base [89M]**	DenseNet121 [8M]	DenseNet201 [20M]	**ConvNeXt Base [89M]**
MIOU	0.3336	0.3254	**0.3263**	0.1257	0.2331	0.1293
DICE Score	0.4659	0.4564	**0.4566**	0.1960	0.3452	0.2036
Hausdorff Distance	93.0492	94.3235	**92.5897**	126.3200	112.0827	133.6850

3-Label Classification

	LIME	Grad-CAM
	ConvNeXt Base [89M]	**ConvNeXt Base [89M]**
MIOU	**0.3950**	0.2355
DICE Score	**0.5312**	0.3449
Hausdorff Distance	**81.9963**	92.5856

2-Label Classification

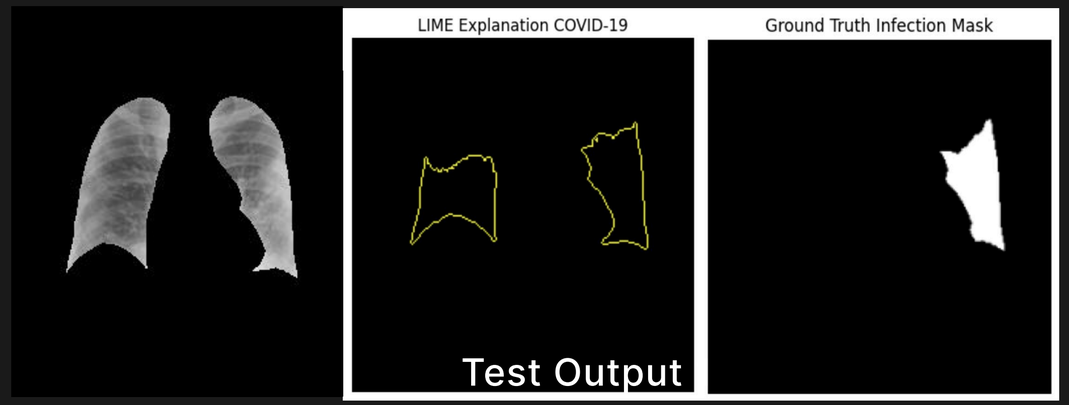
Stage 03 : Contributed by Vipul Gupta

Processing with LIME to obtain Saliency Maps

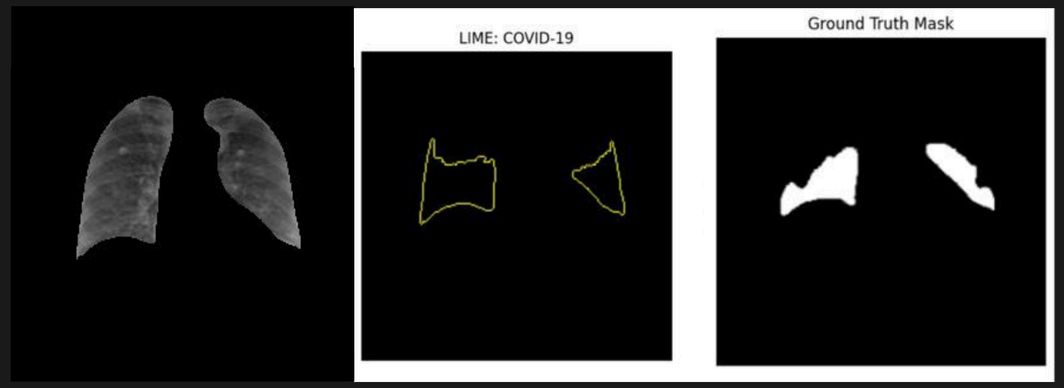
Local Interpretable Model-agnostic Explanations (LIME) is a prominent perturbation-based technique designed to explain the predictions of any classifier or regressor in a transparent manner.

The core goal of LIME is to identify an interpretable model (often a sparse linear model) that is locally faithful to the complex black-box classifier. It achieves this by generating simulated data points (perturbed samples) in the vicinity of the input, weighting these samples by their proximity, and then fitting the simple model locally.

We discover for localization in medical imaging, LIME typically outperforms Grad-CAM by generating precise and boundary-respecting infection masks, confirming that segmentation-based local explanations are often superior for producing clinically meaningful results.



Test Output



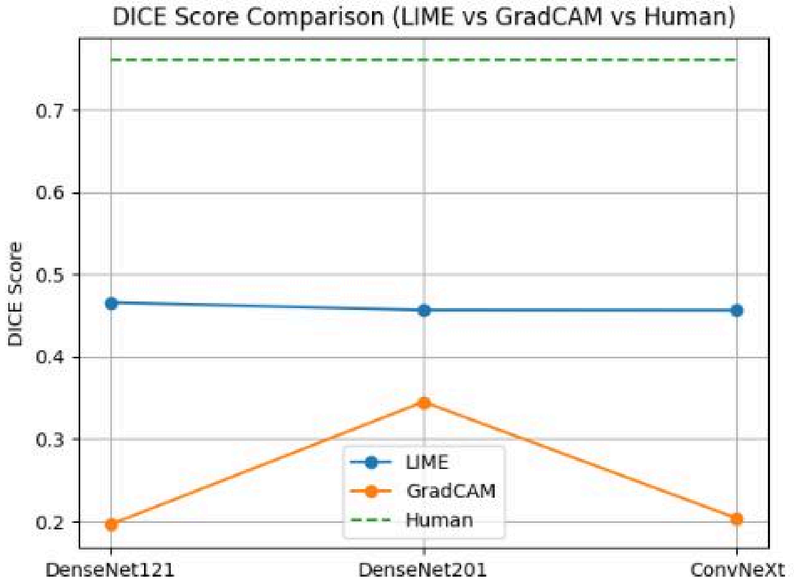
Test Output 2

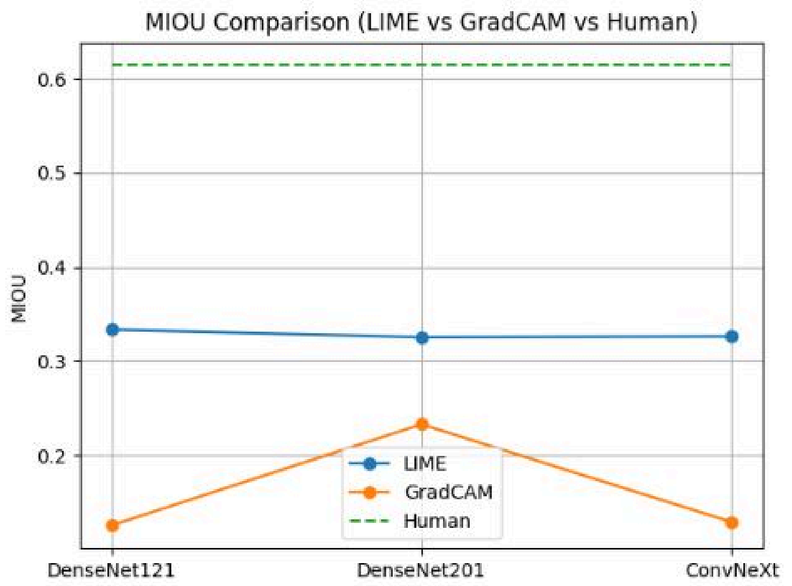
Stage 03 : Contributed by Vipul Gupta

Processing with LIME to obtain Saliency Maps

LIME achieves better MIOU and DICE because it generates superpixel-based explanations that produce smoother, more localized regions matching the actual infected areas. Its perturbation-based approach captures fine-grained spatial patterns, leading to cleaner and more anatomically aligned masks. Since LIME is model-agnostic, it remains stable across architectures and does not depend on deep feature maps. This results in more consistent and accurate segmentation overlaps compared to gradient-based methods.

Grad-CAM performance drops for ConvNeXt because the model uses deeper, more complex hierarchical feature representations, making its final activation maps harder to interpret. ConvNeXt's architecture produces more diffuse and less spatially localized gradients compared to traditional CNNs. As a result, Grad-CAM heatmaps become less focused on the actual lesion regions. This reduces their overlap with ground-truth masks, causing lower MIOU and DICE scores.





Classification benchmark

- Random Forest (handcrafted features) — 94.7% accuracy on a ~6.6k test split (COVID-QU-Ex). MedRxiv(2023)
- LSE-Net (segmentation + ensemble CNNs) — 92.7% accuracy, 96.7% recall, AUC ≈ 0.944. researchgate.net(2023)

XAI benchmark

- Grad-CAM mIoU ≈ 0.248 (Saporta 2023) - 10 class clasification - DenseNet Ensemble.

VIII. Future Work / Timeline

1. Utilizing Interpretability for Automated Severity Scoring

2. Leveraging VAEs for Semi-Supervised Infection Mask Generation

3. Developing Contrastive (Counterfactual) Explanations

Apply ConvNeXt	Obtain infection masks from ConvNeXt	Evaluate infection masks generated	Evaluating 2-Label Classification	Improve Metrics with advanced architecture	Documentation
DONE	DONE	DONE	DONE	DONE	DONE

IX. Challenges

1. Computational Cost
2. Lung x-ray noise in images
3. Lung segmentation mask for datasets
4. Infection Masks for datasets
5. Comparison statistics for masking
6. Hit Rate Calculation with LIME

X. Conclusion

1. Enhanced Interpretability

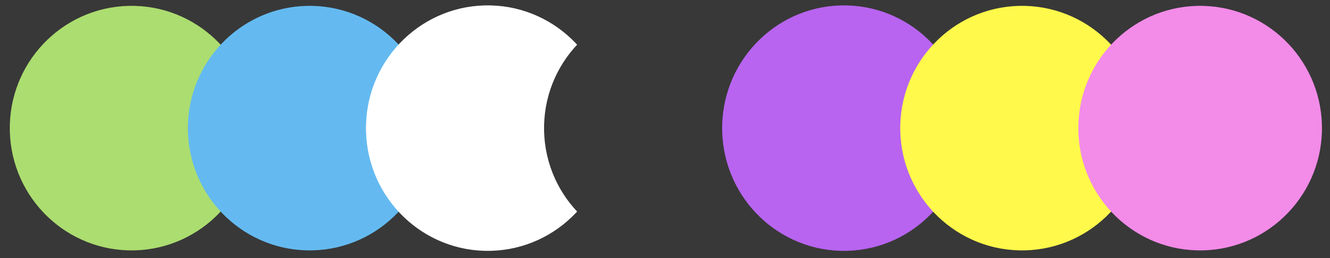
Provide clear, visual explanations (heatmaps) through LIME, making the "black-box" decision-making process of the CNN transparent and understandable to human users.

2. Clinical Acceptance

Increased Clinical Trust and foster greater trust and confidence among clinicians in AI-assisted diagnostic tools by offering actionable insights into model reasoning.

3. Infection Masks

Generate infection masks highlighting affected regions in medical images, enabling doctors to provide interpretable, localized insights into disease detection.



XI. References

1. R. Ghnemat, S. Alodibat, and Q. Abu Al-Haija, "Explainable Artificial Intelligence (XAI) for Deep Learning Based Medical Imaging Classification," *Journal of Imaging, vol. 9, no. 9, p. 177, Aug. 2023.*
2. K. Borys, et al., "Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches," *European Journal of Radiology, vol. 162, p. 110786, Mar. 2023.*
3. Y. H. Bhosale, et al., "Thoracic-net: Explainable artificial intelligence (XAI) based few shots learning feature fusion technique for multi-classifying thoracic diseases using medical imaging," *Multimedia Tools and Applications, Oct. 2024.*
4. X. Zhang, et al., "CXR-Net: An Encoder-Decoder-Encoder Multitask Deep Neural Network for Explainable and Accurate Diagnosis of COVID-19 pneumonia with Chest X-ray Images," *arXiv preprint arXiv:2110.10813, 2021.*
5. P. Afshar, et al., "COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images," *Pattern Recognition Letters, vol. 138, pp. 638-643, Oct. 2020.*
6. M. T. Ribeiro, S. Singh, and C. Guestrin, "“Why Should I Trust You?”: Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 2016, pp. 1135–1144.*
7. J. W. Lua, A. Socolov, and A. Szep, "Modifying LIME for Medical Images," in *Proceedings of Machine Learning Research 6.871 Machine Learning for Healthcare.*
8. J. Adebayo, et al., "Sanity checks for saliency maps," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 2018.*

9. S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, “ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders,” arXiv preprint arXiv:2301.00808, 2023.
10. S. Saporta, X. Gui, A. Agrawal, et al., “Benchmarking saliency methods for chest X-ray interpretation,” Nature Machine Intelligence, vol. 4, pp. 867–878, 2022.
11. Ma, Y., and W. Lv, “Identification of pneumonia in chest X-ray image based on transformer,” International Journal of Antennas and Propagation, 2022, Article 5072666.
12. A. I. Khan, J. L. Shah, and M. M. Bhat, “CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images,” Computer Methods and Programs in Biomedicine, vol. 196, p. 105581, 2020.
13. L. Wang, Z. Q. Lin, and A. Wong, “COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images,” Scientific Reports, vol. 10, no. 1, pp. 1–12, 2020.

THANK YOU