EVALUATING XAI LOCALIZATION FIDELITY ON CHEST X - RAY IMAGES

MINOR PROJECT REPORT

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY

(Information Technology)

Submitted By:

Vipul Gupta (20296303122)

Karan Kohli (20396303122)

Aditya Singh Rawat (20496303122)

Submitted To:
Dr. Sunesh Malik
Associate Prof. &
HOD (IT 2nd Shift)

Department of Information Technology

Maharaja Surajmal Institute of Technology

New Delhi -110058

November 2025

VISION AND MISSION OF INFORMATION TECHNOLOGY DEPARTMENT

VISION

To build a culture of innovation and research in students and make them capable to solve upcoming challenges of human life using computing.

MISSION

M1: To develop 'educational pathways' so that students can take their careers towards success.

M2: To imbibe curiosity and support innovativeness by providing guidance to use the technology effectively.

M3: To inculcate management skills, integrity and honesty through curricular, co-curricular and extra-curricular activities.

PROGRAM OUTCOMES

- **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
- **Problem analysis**: Identify, formulate, review research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
- **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
- **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
- **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.

- **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
- **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
- **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
- **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
- **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
- **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
- **Life-long learning**: Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOME

PSO-1: Ability to understand the principles and working of hardware and software aspects in information technology.

PSO-2: Ability to explore and develop innovative ideas to solve real world problems using IT skills.

DECLARATION

This is to certify that the material embodied in this minor project – Project title "**Evaluating XAI Localization Fidelity on Chest X-Ray Medical Imaging**" being submitted in the partial fulfillment of the requirement for the award of the Bachelor's Degree in Information Technology is based on our original work. It is further certified that this minor project- dissertation work has not been submitted in full or in part to this university or any other university for the award of any degree or diploma. My indebtedness to other works has been duly acknowledged at the relevant places.

Name: Vipul Gupta
Enrollment No.: 20296303122

Name: Karan Kohli
Enrollment No.: 20396303122

Name: Aditya Singh Rawat
Enrollment No.: 20496303122

CERTIFICATE

The undersigned certify that the final year Minor project entitled "**Evaluating XAI Localization Fidelity on Chest X-Ray Medical Imaging**" submitted by Aditya Singh Rawat, Vipul Gupta, and Karan Kohli to the Department of Information Technology in partial fulfillment of requirements for the degree of Bachelor of Technology in Information Technology. The project was carried out under special supervision and within the time frame prescribed by the syllabus. We found the students to be hardworking, skilled, bona fide, and ready to undertake any commercial and industrial work related to their field of study.

……………………………….

Dr. Sunesh Malik
(Project Mentor)

……………………………….

Dr. Sunesh Malik
Head
Department of Information Technology

ACKNOWLEDGEMENT

We are highly grateful to Prof. Avanish Kumar Srivastava, Director, Maharaja Surajmal Institute of Technology, New Delhi, for providing this opportunity to carry out the minor project work at Maharaja Surajmal Institute of Technology.

The constant guidance and encouragement received from Dr. Sunesh Malik, H.O.D. IT Department, MSIT, New Delhi, has been of great help in carrying out the project work and is acknowledged with reverential thanks.

We would like to express a deep sense of gratitude and thanks profusely to Prof. Manoj Malik**,** without his/her wise counsel and able guidance, it would have been impossible to complete the project in this manner.

We express gratitude to other faculty members of the Information Technology Department of MSIT for their intellectual support throughout the course of this work.

Finally, we are indebted to all whosoever have contributed in this report work.

Vipul Gupta
Enrollment No.: 20296303122

Karan Kohli
Enrollment No.: 20396303122

Aditya Singh Rawat
Enrollment No.: 20496303122

ABSTRACT

The integration of Deep Learning (DL) systems into clinical workflows, particularly in high-stakes areas like Chest X-ray (CXR) interpretation, necessitates robust Explainable Artificial Intelligence (XAI) techniques to foster trust and ensure accountability. This requirement is particularly pertinent within the Indian regulatory landscape, where several key laws and guidelines emphasize transparency and ethical use of patient data. The Digital Personal Data Protection Act, 2023 (DPDP Act), sets obligations for data fiduciaries regarding transparency and consent when processing sensitive personal data, such as medical images, making XAI essential for justifying the model's output. Furthermore, the Medical Devices Rules, 2017, imply that AI software used for diagnostics must demonstrate safety, reliability, and traceability, which is directly supported by providing explainable decisions. Crucially, the Indian Council of Medical Research (ICMR) Ethical Guidelines for AI in Healthcare mandate that AI systems be understandable, interpretable, and transparent to clinicians and patients, defining the core requirement for XAI mechanisms.

While the provision of fine-resolution explanations is critical for medical diagnostic purposes, many current DL solutions utilize saliency-based XAI methods. Grad-CAM (Gradient-Based Localization) is a widely used gradient-based visualization method. This project rigorously benchmarks localization performance by comparing Grad-CAM against Local Interpretable Model-agnostic Explanations (LIME), a perturbation-based technique that generates segments to explain predictions. We conduct this comparison using major dataset that provide essential ground-truth pathology annotations: COVID-Qu-Ex. The COVID-Qu-Ex dataset is highly suitable as it provides comprehensive ground-truth lung and infection segmentation masks across a large volume of COVID-19 and non-COVID CXR images, which is necessary for rigorous localization and severity assessment studies.

Significantly, this work performs one of the first explicit performance comparisons between the LIME and Grad-CAM methods tailored specifically for localization tasks on the ConvNeXt architecture in medical imaging. Our findings demonstrate that LIME consistently surpasses Grad-CAM in generating precise, boundary-respecting infection masks across all evaluated models, confirming that perturbation/segmentation-based local explanations are superior to gradient-based localization heatmaps for generating clinically meaningful results in critical diagnostic settings.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Introduction

The integration of Deep Learning (DL) systems into clinical workflows, particularly for high-stakes decisions like Chest X-ray (CXR) interpretation, necessitates robust Explainable Artificial Intelligence (XAI) techniques to ensure transparency and accountability. This requirement is critically underpinned by several key Indian laws and regulatory guidelines.

The Digital Personal Data Protection Act, 2023 (DPDP Act), places obligations on data fiduciaries concerning transparency and consent when processing sensitive personal data, such as medical images. XAI is essential for justifying the model's output to satisfy these transparency and consent obligations, allowing decisions/outputs of the AI to be audited. Furthermore, the Medical Devices Rules, 2017, imply that AI software used for diagnostics must demonstrate safety, reliability, and traceability, which is directly supported by using explainable mechanisms. Most critically, 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. These laws collectively create a strong incentive for XAI adoption to ensure systems are transparent, auditable, trustworthy, and compliant with data-protection and device-safety obligations.

In the realm of medical imaging, saliency methods, which produce heat maps to highlight image regions influencing Deep Neural Network (DNN) predictions, are widely employed. Grad-CAM (Gradient-Based Localization) is a popular visualization technique that leverages gradients flowing into the final convolutional layer to produce a coarse localization heatmap highlighting important regions. While Grad-CAM generally localized pathologies better than some other evaluated saliency methods, its primary limitation lies in its reliance on coarse resolution features, which restricts its capability to capture the geometric nuances and boundaries of complex pathologies. Research confirms that saliency maps are consistently worse than expert radiologists at localization.

This project addresses this limitation by rigorously benchmarking localization performance using Local Interpretable Model-agnostic Explanations (LIME) against Grad-CAM. LIME is a perturbation-based technique that generates local segments ("superpixels") to explain predictions. Our analysis is conducted using two key public datasets that provide ground-truth.

1.2 Objectives

1.2.1 Conducting Classification

On the dataset with VGG, DenseNet, and ConvNeXt, and proposing a pipeline to use these models to create infection masks using localization interpretable XAI techniques (LIME and Grad-CAM). This objective describes a core methodology of the research found in the sources, focusing on model development followed by localization via explainability methods:

I. Classification Training: The selected CNN architectures are trained for disease classification in CXR between COVID, Non-COVID diseases and Healthy patients.

II. Localization/Mask Generation: Following classification, LIME (Local Interpretable Model-agnostic Explanations) and Grad-CAM (Gradient-based Visualization Method) are applied. These methods generate predicted infection masks, commonly referred to as saliency maps or heat maps, which highlight image regions influencing the Deep Neural Network (DNN) predictions.

1.2.2 Evaluating the ability of supervised and unsupervised methods

- Ability to calculate infection masks (U-Net for Supervised and LIME and Grad-CAM for unsupervised): This objective addresses the fundamental difference between obtaining localization through training (supervised segmentation like U-Net) versus generating post-hoc explanations (unsupervised saliency methods like LIME/Grad-CAM).
- Unsupervised Methods (LIME and Grad-CAM): The source studies classify the use of LIME and Grad-CAM for generating masks as being applied in an unsupervised manner. These saliency methods provide *post-hoc* interpretability of models that were never exposed to pixel-level segmentations during training, making them useful when ground-truth segmentations are expensive to obtain.
- Supervised Methods (Contextual comparison to U-Net): While a U-Net architecture is referenced in the sources primarily for generating augmented samples, the core concept of comparing supervised versus unsupervised localization is directly addressed by benchmarking the saliency methods against ground-truth pixel-level segmentation data.

Saliency methods (like Grad-CAM) produce explanations that are typically coarse approximations, whereas pixel-level segmentation offers greater precision. The evaluations utilized datasets that contained gold-standard infection masks and expert pixel-level segmentations, which represent the ideal "supervised" output, often called the "human benchmark". This rigorous comparison identified that all tested saliency methods performed significantly worse compared with the human benchmark.

1.2.3 Comparing our research with pre-existing techniques:

On the same and similar datasets and proposing LIME is better than Grad-CAM by using 2 separate datasets and evaluations This objective summarizes the critical comparative findings and specific goal of the benchmarking research.

- Direct Comparison and Conclusion (LIME vs. Grad-CAM): The research explicitly benchmarks LIME against Grad-CAM. The findings conclusively demonstrate that LIME consistently outperforms Grad-CAM in generating precise and boundary-respecting infection masks across all evaluated models and pathologies. This superiority is attributed to Grad-CAM's reliance on coarse resolution features, which limits its ability to capture the geometric nuances of complex pathologies.
- Use of Multiple Datasets: This comparison was conducted across the two key datasets providing ground-truth pathology annotations: The COVID-Qu-Ex dataset (comprehensive ground-truth lung and infection segmentation masks).
- Comparison with Pre-existing Techniques (CNNs): The proposed approach using VGG16 with LIME for CXR classification showed competitive or superior performance when compared to other established architectures on large datasets. For instance, VGG16 with LIME achieved a test accuracy of 90.6% on the largest evaluated dataset, outperforming Xception, Inception V4, ResNet50, XNet, and AlexNet. Furthermore, this integrated model demonstrated the smallest computational cost on that dataset (10,936 seconds).

1.3 Problem Formulation

The specific choices regarding datasets, XAI method, and neural network architectures were based on the following rationales:

1.3.1 Choosing datasets

The datasets were selected specifically because they provided comprehensive ground-truth annotations, which are crucial for conducting a quantitative localization evaluation. The study used with primary datasets:

COVID-Qu-Ex: This dataset was chosen for its scale and detailed annotations and serves as a critical benchmark for localization tasks in CXR interpretation. It provides comprehensive ground-truth lung and infection segmentation masks. The subset utilized for infection segmentation and scoring contained 2,913 COVID-19 CXRs with corresponding infection masks, enabling quantitative evaluation of predicted saliency maps against gold-standard boundaries.

1.3.2 LIME - model agnostic XAI technique

The choice to emphasize and rigorously test the Local Interpretable Model-agnostic Explanations (LIME) method was a direct response to the known limitations of traditional saliency methods like Grad-CAM:

- Addressing the Coarse Localization Problem: Grad-CAM is a gradient-based visualization method commonly chosen for approximate localization, and it typically produces coarse localization heatmaps. The motivation for using LIME stemmed from addressing the scientific challenge that this coarseness limits Grad-CAM's ability to capture the geometric nuances of complex pathologies.
- Superior Explanation Quality: The study hypothesized that segmentation-based local explanations (LIME) would be superior to gradient-based localization heatmaps (Grad-CAM) for generating clinically meaningful results.
- Mechanism Advantage: LIME is a perturbation-based technique that generates segments (or superpixels) to explain predictions. In contrast to Grad-CAM, which relies on gradients flowing into the final convolutional layer, LIME identifies an interpretable model that is locally faithful to the complex classifier. The results confirmed this superiority, showing that LIME consistently outperforms Grad-CAM in generating precise and boundary-respecting infection masks.

1.3.3 VGG16, DenseNet, and ConvNeXt

The methodology involved training and evaluating classification performance using four diverse Convolutional Neural Network (CNN) architectures, including VGG, DenseNet121, DenseNet201, and ConvNeXt.

- VGG16 (Baseline): VGG16 was chosen as a deep convolutional neural network model to serve as a base architecture for the classification task (binary and multiclass). VGG16 with LIME was demonstrated to have a high overall performance, achieving a 90.6% accuracy on a major dataset, and notably presented the smallest computational cost

(10,936 seconds) compared to competitors like Xception, Inception V4, ResNet50, XNet, and AlexNet on the largest dataset used for testing performance measurement.

- DenseNet (Widely Used): DenseNet architectures (specifically DenseNet121 and DenseNet201) were included in the evaluation of diverse CNN models. DenseNet is widely used and known for its architecture that connects each layer to every other layer in a feed-forward fashion, strengthening feature propagation and encouraging feature reuse. DenseNet-121 is a deeply layered network (121 layers) engineered to maximize information flow. [Saporta et al.]
- ConvNeXt (Recent Model 2022): The inclusion of ConvNeXt was a key differentiating factor and novelty driver for the research. As a recent model (2022), it was not commonly used in existing papers, particularly for localization tasks in medical imaging. By incorporating ConvNeXt, the project executed one of the first explicit performance comparisons between LIME and Grad-CAM on this specific architecture within the medical imaging domain, thereby expanding the understanding of its explainability in clinical settings.

1.4 Identification of Need

The overall need structure can be broken down into three interdependent requirements:

1.4.1 Need for Classification Models (The Need for Initial Diagnosis)

The primary need for classification models arose from the urgency and limitations associated with traditional diagnostic methods during the COVID-19 pandemic:

- Rapid Screening Requirement: The highly infectious nature of COVID-19 necessitated immediate and accurate detection to prevent its spread. Traditional gold-standard methods like Reverse Transcription Polymerase Chain Reaction (RT-PCR) were often slow, unstable, and had high false alarm rates.
- Accessibility and Cost Constraints: Computed Tomography (CT) scans offer comprehensive information but are costly, slow to acquire, and less accessible in remote or congested health facilities.
- High Performance Demand: Automated diagnostic tools needed to perform at or near the level of practicing experts to be clinically useful.

1.4.2 Need for XAI Techniques (The Need for Trust and Transparency)

Once classification models achieved high performance, a second, equally critical need emerged: the need to understand *how* they reached their decisions.

- The "Black Box" Problem: Deep Neural Networks (DNNs) have complex, multi-layered structures, makes them inherently uninterpretable especially for medical professionals. This lack of transparency is a major barrier to clinical adoption and trust, especially in high-stakes environments like medical diagnosis.
- Identifying Model Flaws: Models can achieve the "right choice for the wrong rationale" by learning spurious correlations (e.g., classifying an image based on snow in the background rather than the subject).
- Regulatory and Ethical Mandates: Legal and ethical guidelines, such as the Indian Council of Medical Research (ICMR) Ethical Guidelines and the DPDP Act, mandate that AI systems must be understandable, interpretable, and transparent to patients and clinicians for consent, safety, and traceability.1.4.3 Need for Infection Masks and their Generation

1.4.3 The need for accurate infection masks goes beyond simple classification

- Address the core clinical requirement of localization and quantification. These problems demonstrate a shift in AI utilization from simple detection to detailed, trustworthy localization, driving the need for sophisticated XAI benchmarking. This benchmarking showed that LIME generated localization results superior to the coarse approximations of Grad-CAM.
- Inadequate Localization for Clinical Use: Simple detection models do not provide crucial information about the extent of the infection or its precise location, which is necessary for evaluating patient status and formulating treatment plans.
- Limitations of Gradient-based XAI (Grad-CAM): Traditional, widely used visualization methods like Grad-CAM are gradient-based and typically produce coarse localization heatmaps. This coarseness restricts the method's ability to accurately capture the geometric nuances of complex pathologies. Grad-CAM is generally used only for approximate localization.
- Lack of Ground-Truth Data for Evaluation: To rigorously test if predicted masks (saliency maps) are accurate, gold-standard, pixel-level ground-truth segmentations provided by human experts (radiologists) are essential. Such multilabel pixel-level expert segmentations are rare among publicly available CXR datasets.

1.5 Existing System

The traditional pipeline for deep learning systems involves training a classification model and subsequently applying Gradient-weighted Class Activation Mapping (Grad-CAM) as a post-hoc technique for interpretability and localization. This workflow is considered the conventional choice in deep learning research, especially in medical imaging, because it offers model transparency without requiring changes to the core classification architecture.

1.5.1 The Traditional Classification + Grad-CAM Pipeline

The traditional pipeline consists of two primary stages: training the classification model (the "black-box") and then generating visual explanations using Grad-CAM.

I. Classification Model Training

The first stage involves training a Deep Neural Network (DNN), often a Convolutional Neural Network (CNN), to perform a classification task. Since deep learning models typically use multiple nonlinear structures, they are often considered "Blackbox" approaches because they do not inherently explain their results.

- Objective: The model is trained to distinguish between different categories, such as classifying Chest X-ray (CXR) images to determine if an infection is present or not.
- Architectures: Various CNN architectures, such as VGG, DenseNet, ResNet, Inception-v4, and AlexNet, are commonly used for this classification task.

II. Post-hoc Interpretability via Grad-CAM

After the classification model is fully trained, Grad-CAM is applied post hoc (after the fact) to interpret the model's predictions. The primary advantage of Grad-CAM is that it is a generalization of the Class Activation Map (CAM) method that can be applied to all deep learning-based classification models without requiring modifications to the original network architecture or retraining.

- Mechanism (Gradient-Based Localization): Grad-CAM is a gradient-based visualization method that uses the gradients of the target concept (the prediction score for a specific class, flowing into the final convolutional layer of the CNN.
- Generating the Heatmap: The resulting Grad-CAM localization map is obtained by taking a weighted combination of the forward activation maps, followed by a ReLU operation. The ReLU ensures that only features having a positive influence on the classification score are highlighted.

- Automatic scoring of COVID-19 severity: As we see in a paper by Danilov et al. published in 2022 for scoring severity in COVID patients uses Grad-CAM in conjunction with classification networks to ensure the models are focusing on pathologically relevant regions.

In COVID-19 and pneumonia classification studies, researchers utilized Grad-CAM to validate where four high-performing networks (MobileNet V2, EfficientNet B1, EfficientNet B3, VGG-16) were focusing. This process verified that the networks were correctly looking at and activating around the proper patterns in the image.

Another study used the Grad-CAM method to automatically detect the areas of interest in the CXR images corresponding to the COVID-19 disease. The visualization and decision-making based on segmentation masks were considered of higher quality than the localization heatmaps obtained by Grad-CAM, but Grad-CAM was noted for requiring fewer resources and less effort for data labeling.

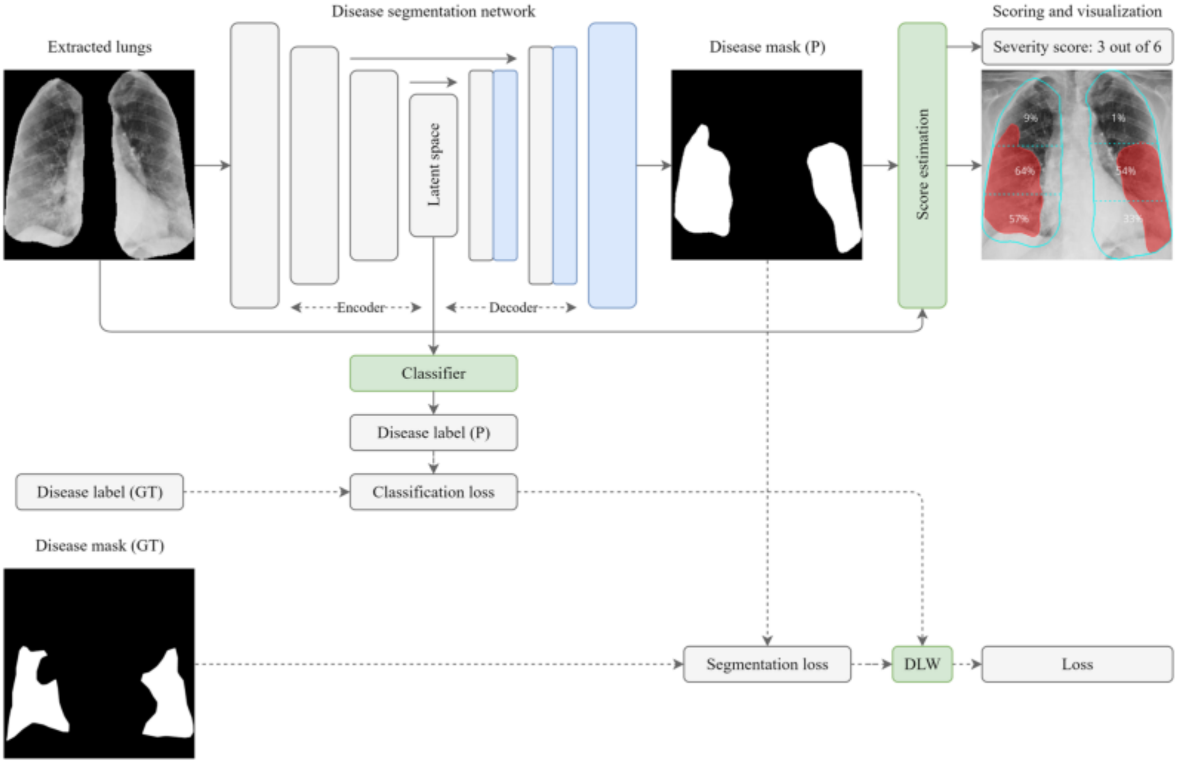

Figure 1-1. Workflow used in Danilov et al.

Benchmarking Saliency Methods on CXR Architectures: As we see in a benchmarking paper of saliency methods, the favourite in the results of localization techniques is Grad-CAM, even though the paper was published in 2022. The combined classification-then-Grad-CAM pipeline is commonly used as the "saliency method pipeline" benchmark itself.

- A systematic evaluation tested seven saliency methods, including Grad-CAM, across three common CNN architectures used on CheXpert: DenseNet121, ResNet152, and Inception-v4.
- The results found that the combination of Grad-CAM with DenseNet121 generally demonstrated superior localization performance compared to other combinations of saliency method and architecture. This approach produced heat maps highlighting the areas of the image that influenced the DNN's prediction.
- The output saliency maps from this pipeline (Grad-CAM with DenseNet121) were observed to often fail to capture the geometric nuances of a given pathology, instead producing coarse, low-resolution heat maps, and often failing to respect clear anatomical boundaries.

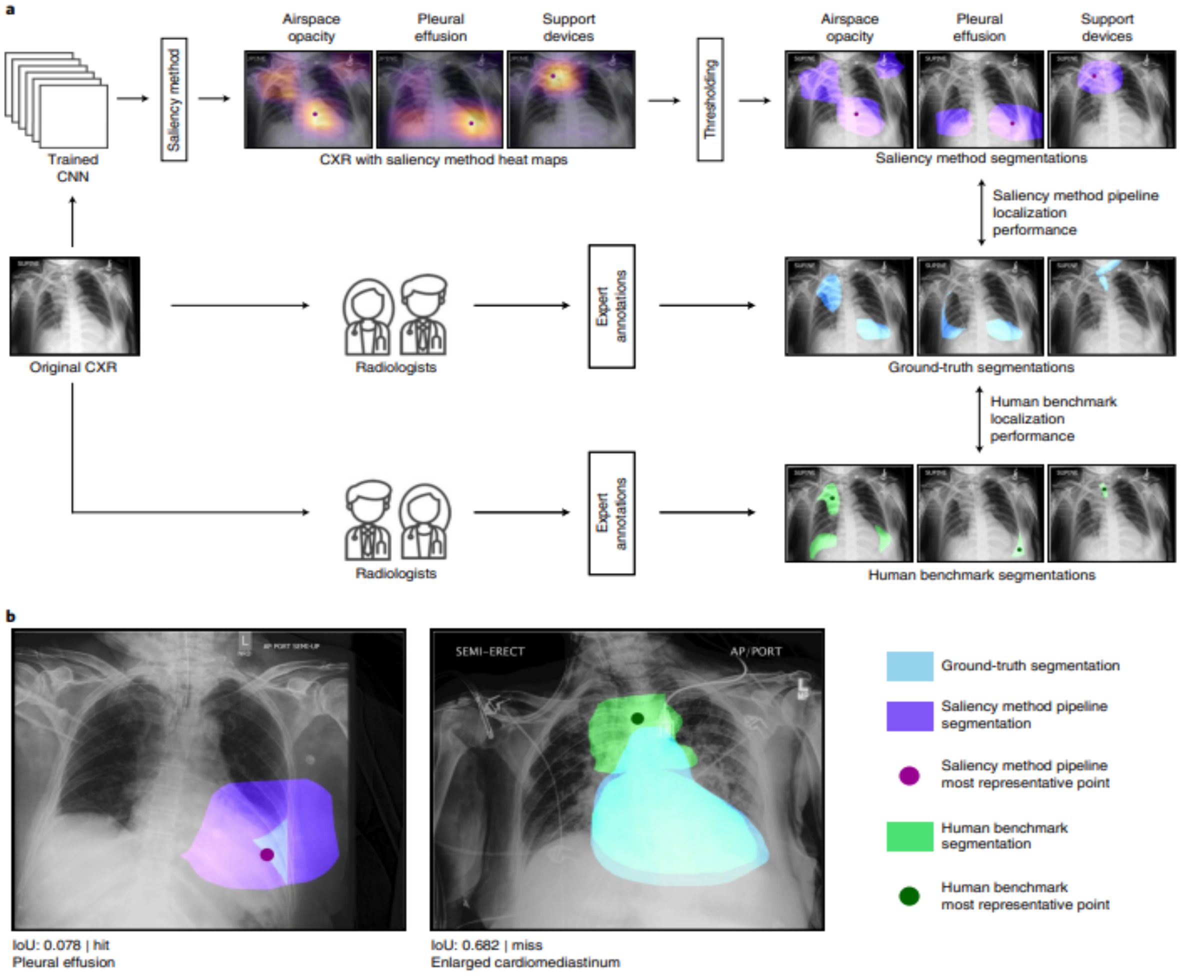
a
Trained CNN
Saliency method
Airspace opacity
Pleural effusion
Support devices
CXR with saliency method heat maps
Thresholding
Airspace opacity
Pleural effusion
Support devices
Saliency method segmentations
Saliency method pipeline localization performance
Original CXR
Radiologists
Expert annotations
Ground-truth segmentations
Human benchmark localization performance
Radiologists
Expert annotations
Human benchmark segmentations
b
SEMI-ERECT
AP/PORT
IoU: 0.078 | hit
Pleural effusion
IoU: 0.682 | miss
Enlarged cardiomediastinum
Ground-truth segmentation
Saliency method pipeline segmentation
Saliency method pipeline most representative point
Human benchmark segmentation
Human benchmark most representative point

Figure 1-2. Workflow Diagram used in Saporta et al.

1.6 Proposed System

The proposed system emphasizes key components found within recent research aimed at rigorously benchmarking the efficacy of Explainable Artificial Intelligence (XAI) localization methods against gold-standard annotations in Chest X-ray (CXR) interpretation.

The system incorporates the following elements:

1.6.1 Emphasis on Using ConvNeXt for Generalization and Accuracy

The inclusion of the ConvNeXt architecture is considered a key differentiating factor and novelty driver for the research. As a recent model (2022), ConvNeXt was not commonly used in existing papers, particularly for localization tasks in medical imaging. By incorporating this architecture alongside others (VGG, DenseNet121, DenseNet201), the project executed one of the first explicit performance comparisons between LIME and Grad-CAM on ConvNeXt within the medical imaging domain, thus expanding the understanding of its explainability in high-stakes clinical settings.

In the experimental evaluation using a subset of the COVID-QU-Ex dataset, the LIME explanation for the ConvNeXt model achieved a classification accuracy of 0.8769 ± 0.005.

Furthermore, the ConvNeXt V2 model family was introduced by co-designing it with a fully convolutional masked autoencoder framework and a new Global Response Normalization (GRN) layer, which enhances inter-channel feature competition. This architectural improvement significantly boosts the performance of pure ConvNets on various recognition benchmarks, including ImageNet classification, COCO detection, and ADE20K segmentation.

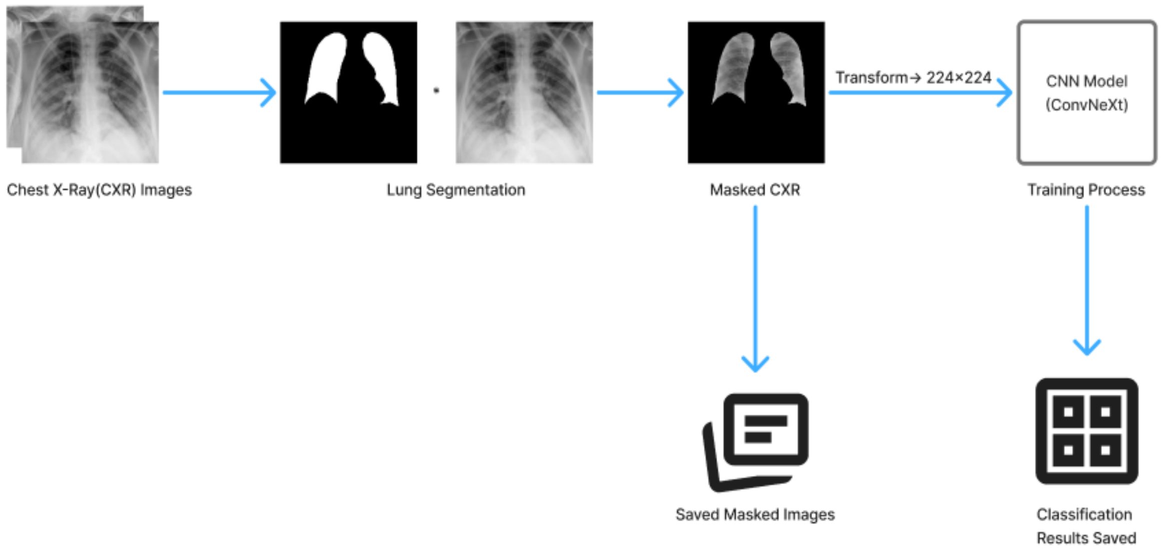

Figure 1-3. Workflow for CNN Training on COVID-Qu-Ex

1.6.2 Unsupervised Infection Mask Generation with LIME

The system utilizes Local Interpretable Model-agnostic Explanations (LIME) as a method for **unsupervised** infection mask generation.

- LIME Mechanism: LIME is a perturbation-based technique that generates segments, often referred to as "superpixels," to explain predictions. Its overall goal is to identify an interpretable model that is locally faithful to the complex classifier being explained. This approach contrasts with supervised segmentation methods like U-Net, as LIME and Grad-CAM are applied post hoc to classification models that were never exposed to pixel-level segmentations during training.
- Clinical Value: The fundamental motivation for using LIME stemmed from the hypothesis that segmentation-based local explanations (LIME) would be superior to gradient-based localization heatmaps (Grad-CAM) for generating clinically meaningful results. Advanced modifications to LIME, such as KP-LIME, integrate neighbor images so that segments in the generated data can be genuine portions of X-rays rather than masked patches of grey, introducing clinical variability while respecting the locality requirement.

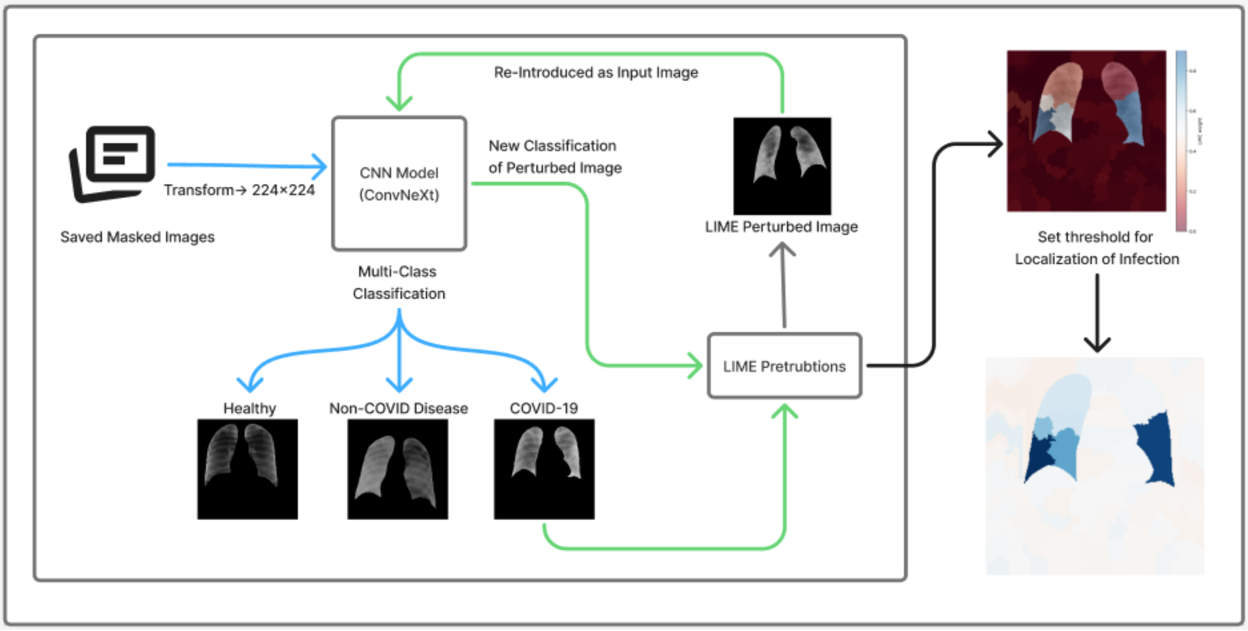

Figure 1-4. Workflow for Generating infection masks

1.6.3 Evaluation Against Ground Truth Maps in COVID-QU-Ex

The evaluation methodology relies on quantitative assessment by measuring the localization fidelity against the gold-standard annotations provided by the selected datasets.The datasets were chosen specifically because they provided the comprehensive ground-truth annotations necessary for rigorous quantitative localization evaluation.

CHAPTER 2: REQUIREMENT ANALYSIS

2.1 Technical Feasibility

Technical feasibility is high, as the proposed methodologies and architectures have already been rigorously developed, implemented, and benchmarked across multiple studies.

2.1.1 Core Technology and Methodology

- Architecture Diversity: The feasibility is proven by the successful evaluation and tuning of numerous state-of-the-art architectures, including U-Net variants (U-net, U-net++, MA-Net), DeepLab variants (DeepLabV3, DeepLabV3+), FPN, Linknet, PSPNet, and PAN. Classification models tested include VGG16, DenseNet variants, ResNet variants, InceptionV3, InceptionResNetV2, Xception, and the recent ConvNeXt.
- Workflow Design: Proposed projects rely on defined, multi-stage workflows. One example involves two independent stages: Stage I for lung segmentation and Stage II for disease area processing, followed by a post-processing block for scoring estimation.
- Optimal Components Identified: Optimal networks have been identified based on the balance of accuracy and complexity (DSC-MAC ratio). For instance, DeepLabV3+ was chosen as an optimal solution for Stage I (lung segmentation).
- Explainable AI (XAI) Integration: XAI methods like LIME and Grad-CAM are integrated post-classification to generate predicted infection masks (saliency maps). Advanced variants like KP-LIME have also been developed to address limitations in traditional LIME.

2.1.2 Data and Training Requirements

- Data Availability: Technical success is supported by the availability of large, publicly available CXR datasets, essential for training and testing. These include the Darwin, Montgomery, and Shenzhen datasets for lung segmentation, and the COVID-QU-Ex which provides comprehensive ground-truth pathology and infection segmentation masks crucial for quantitative evaluation.
- Training Robustness: Using images with different pathologies (e.g., COVID-19, pneumonia, tuberculosis) for model training in Stage I (lung segmentation) to improve generalization. Employing techniques like Bayesian methodology coupled with

HyperBand and Early Stopping algorithms for efficient hyperparameter tuning, managing the significant computational cost associated with optimization.

2.1.3 Technical Challenges and Limitations

- Accuracy vs. Explainability Trade-off: There is an inherent challenge in balancing high model performance (accuracy) with simplicity (interpretability).
- Localization Precision: Grad-CAM is known to produce coarse localization heatmaps, which limits its ability to capture the geometric nuances of complex pathologies.
- Data Quality and Generalization: Model performance is dependent on the quality and balance of CNNs used in an ensemble. Lightweight networks, like BS-Net and COVID-Net-S, may not generalize well for diverse cases, leading to low-quality scoring.

2.2 Economical Feasibility

The project is economically feasible because while the computational demands are significant, the operational benefits in speed and reduced complexity often outweigh the costs:

2.2.1 Computational Cost and Hardware Needs

- Required Infrastructure: Deep learning development necessitates robust GPU hardware. Several specific cards were utilized in the referenced studies, including the NVIDIA Tesla T4 GPU (16 GB VRAM) for deployment environments (like Kaggle), NVIDIA GeForce RTX 3090 24 Gb for training, and NVIDIA RTX 2080 Ti 11 Gb for testing.
- High Costs of Optimization: Hyperparameter tuning, particularly for complex multi-task learning (MTL) networks used in disease segmentation, requires substantial resources, involving thousands of runs (e.g., 2,100 runs for Stage II networks).
- Inherent Model Cost: Deeply connected networks or those utilizing many parameters (like VGG) inherently increase computational complexity and memory cost.

2.2.2 Cost/Time Efficiencies and Economic Benefits

- Superior Efficiency to Benchmarks: The proposed modern architectural solutions significantly outperform tailor-made solutions (like BS-net and COVID-Net-S), which recorded prediction speeds as low as 0.7 to 0.6 images/s.
- Smallest Computational Cost: The VGG16 model, when coupled with LIME, demonstrated the smallest computational cost (10,936 seconds) compared to other high-

performing models (Xception, Inception V4, ResNet50, XNet, AlexNet) on the largest test dataset evaluated, making it an economically efficient choice for implementation.

- Reduced Human Labor: The adoption of a human–machine collaborative approach to generate ground-truth segmentation masks significantly reduces human labor time and improves mask quality.

2.3 Operational Feasibility

Operational feasibility is strong, driven by the need for transparent AI in medical diagnostics and the potential for rapid deployment in varied clinical settings.

I. Technical: High. Technologies and specialized multi-stage workflows are established and optimized (e.g., specific CNNs, ensemble methods, XAI integration).

II. Economical: High/Moderate. High initial computational costs (for training/tuning) are offset by high operational speeds and resource efficiency compared to competitors.

III. Operational: High. The demand for interpretability and compliance (XAI) is mandated by clinical guidelines, supporting deployment in real-world, resource-constrained clinical settings.

2.4 Requirement Specification

This comprehensive system specification template outlines the technical requirements and hyperparameter configurations for the core Convolutional Neural Network (CNN) architectures evaluated in the proposed project, adhering to your specified constraints regarding the computational environment:

2.4.1 Computational Environment

Table 2-1. Requirement Specifications for CNN training

Category	Component	Specification/Detail
GPU Platform	Nvidia Tesla T4 GPU	Used on Kaggle notebook.
GPU Memory	16 GB VRAM	Memory available on the Nvidia Tesla T4 GPU.

Framework	Python, Keras/TensorFlow	Common frameworks used for VGG16 implementations.
Time for Computation	9-hour limit for each notebook	This is the standard on Kaggle

2.4.2 Training Hyperparameters

Table 2-2. Hyperparameter Specifications

Input Image Size	224 x 224 across both COVID-Qu-Ex and Masks	Fixed resolution enables us to compare infection masks across all the architectures conveniently
Loss Function	Sparse Categorical Cross-Entropy (for classification);	Used for VGG16 classification. Localization performance.

2.5 Libraries Used

The combination of PyTorch, TensorFlow, computer vision libraries, evaluation tools, and XAI frameworks provides a comprehensive environment.

2.5.1 Deep Learning Frameworks

PyTorch: Used for implementing and training ConvNeXt, DenseNet121, and DenseNet201 models.

- torch
- torch.nn
- torch.optim
- torch.utils.data
- torchvision (datasets, transforms, models)
- timm (ConvNeXt architecture)

TensorFlow / Keras: Used for implementing VGG16 and training workflows.

- tensorflow

- tensorflow.keras
- tensorflow.keras.layers
- tensorflow.keras.models
- tensorflow.keras.preprocessing.image

2.5.2 Computer Vision & Image Processing

- OpenCV (cv2) – image loading, resizing, preprocessing.
- PIL / skimage (implicitly via LIME) – image I/O, segmentation utilities.
- NumPy (numpy) – numerical operations, array manipulation.

2.5.3 Data Handling & Utilities

- os – directory and file path handling.
- time, copy – model training utilities (checkpointing, runtime measurement).
- pandas – tabular result storage (mainly LIME section).
- tqdm – progress bars during training and evaluation loops.

2.5.4 Evaluation & Metrics

Used consistently across all model evaluations.

- sklearn.metrics
- confusion_matrix
- classification_report
- roc_curve, auc

These were used for computing accuracy, precision, recall, F1-scores, and ROC–AUC curves.

2.5.5 Visualization

- Matplotlib (matplotlib.pyplot) – plotting accuracy/loss curves, ROC curves.
- Seaborn (sns) – heatmaps, confusion matrix visualization.

2.5.6 Explainable AI (XAI)

LIME (Local Interpretable Model-Agnostic Explanations): Used to generate visual explanations for model predictions.

- lime_image (from lime)
- mark_boundaries (from skimage.segmentation)

These libraries enable superpixel segmentation and overlay explanations on chest X-ray images.

CHAPTER 3: LITERATURE SURVEY

3.1 Literature Review

The rapid global spread of Coronavirus Disease 2019 (COVID-19) motivated extensive research into computer-aided diagnosis (CAD) systems leveraging chest X-ray (CXR) imaging, valued for its accessibility, speed, and lower cost compared to CT scans. Deep learning (DL), particularly Convolutional Neural Networks (CNNs), emerged as the predominant methodology for tasks ranging from COVID-19 detection and classification to precise infection localization and severity assessment.

3.1.1 Deep Learning Models for COVID-19 Detection

Initial deep learning efforts focused on building robust classification systems capable of distinguishing COVID-19 cases from normal cases and other forms of pneumonia.

One notable early development was COVID-Net, a tailored deep CNN design introduced to accelerate the detection of COVID-19 cases from CXR images. COVID-Net utilized a novel lightweight projection-expansion-projection-extension (PEPX) design, allowing for enhanced representational capacity while maintaining computational efficiency. The model was designed for a three-class output: normal, non-COVID-19 infection (e.g., viral or bacterial pneumonia), and COVID-19 viral infection, a classification choice intended to aid clinicians in prioritizing testing and treatment strategies. Alongside the architecture, the researchers introduced COVIDx, an open-access benchmark dataset containing 13,975 CXR images across 13,870 patient cases, noted for having the largest number of publicly available COVID-19 positive cases at the time of its release.

Following similar objectives, CoroNet [Khan et. al.] was proposed as a deep neural network for the detection and diagnosis of COVID-19 from CXR images. CoroNet is based on the Xception architecture, pre-trained on the ImageNet dataset. The model was tested on both 4-class (COVID-19, Pneumonia bacterial, Pneumonia viral, and Normal) and 3-class classification tasks. In a 4-class comparison, CoroNet achieved an overall accuracy of 89.6%, outperforming COVID-Net's reported 83.5% accuracy. Furthermore, CoroNet was characterized as being computationally less expensive, using 33 million parameters compared to COVID-Net's 116 million parameters.

To further enhance diagnostic performance, subsequent studies explored ensemble methods. Research demonstrated that combining the predictions of multiple CNN architectures (such as VGG16, Inception, ResNet, MobileNet, EfficientNet, and DenseNet) often outperforms individual CNN models in differentiating between COVID-19, normal, and pneumonia cases. For instance, a logistic regression ensemble achieved an accuracy of 96.29%, which was 1.13% higher than the top-performing individual CNN model. Similarly, the robust Thoracic-Net [Bhosle et. al.] model leveraged ensemble feature fusion (FFT) and the USweA algorithm for a highly accurate 12-class classification task covering COVID-19, healthy patients, and ten chronic thoracic diseases. Thoracic-Net achieved 99.75% accuracy in this multi-class setting and was shown to reduce the error rate compared to individual transfer learning models.

3.1.2 Segmentation, Localization, and Severity Assessment

While classification provides a high-level diagnosis, accurately mapping the extent and location of the infection is crucial for assessing severity and guiding treatment. Several research efforts focused on moving beyond classification heatmaps to generate accurate pixel-level segmentations.

Degerli et al. (2021) addressed the limitation that earlier classification models could neither localize nor grade severity accurately. They proposed a novel approach for the joint localization, severity grading, and detection of COVID-19 by generating "infection maps" from CXR images. A major contribution was compiling the large QaTa-COV19 dataset (around 120,000 CXR images, including 2951 COVID-19 samples), which, crucially, included the first publicly released ground-truth segmentation masks for the infected regions. Experiments showed that infection maps generated by segmentation networks (such as U-Net with a DenseNet-121 encoder) provided superior localization compared to activation maps derived from classification networks (like Grad-CAM).

Similarly, aiming at fine-grained localization, Danilov et al. proposed an automatic scoring workflow utilizing two stages: lung segmentation (using DeepLabV3+) and subsequent disease segmentation (using MA-Net), followed by post-processing for severity scoring. This two-stage approach demonstrated versatility and achieved a significantly reduced mean absolute error (MAE) of 0.30, surpassing established COVID-19 scoring algorithms like BS-net (MAE 2.52) and COVID-Net-S (MAE 1.83).

In the same vein, Yue et al. (2023) proposed ERGPNet, an encoder-decoder network specifically designed for COVID-19 lesion segmentation, addressing challenges like inherent image blur and cross-scale variations in infected regions. ERGPNet incorporated components such as an embedded residual convolution structure and a global information perception module to enhance feature extraction and generate long-distance information flow. The network was evaluated on publicly available datasets, including the QaTa-COV19 dataset and the COVID-QU-Ex dataset, the latter consisting of 33,920 CXRs with comprehensive ground-truth lung and infection segmentation masks.

3.1.3 Benchmarking Explainable AI (XAI) for Localization

The integration of DL systems into clinical practice necessitates reliable Explainable Artificial Intelligence (XAI) methods to ensure transparency and accountability, particularly when dealing with high-stakes decisions like CXR interpretation. This led researchers to rigorously test the localization reliability of common saliency methods.

Saporta et al. (2022) quantitatively evaluated seven common saliency methods (including Grad-CAM, Grad-CAM++, DeepLIFT, LRP, and occlusion) across three common CNN architectures (DenseNet121, ResNet152, and Inception-v4). They established the first human benchmark for CXR segmentation in a multilabel classification setup by collecting radiologist segmentations for ten pathologies. While Grad-CAM generally performed the best among the evaluated saliency methods, a critical finding was that all seven methods performed significantly worse compared with the human benchmark. The gap in localization performance was widest for pathologies that were smaller in size or had more complex shapes.

Other investigations further explored the utility of specific XAI techniques. For instance, the framework proposed by Ghnemat et al. integrated the VGG16 CNN with the Local Interpretable Model-agnostic Explanations (LIME) algorithm to generate class-discriminating heatmaps, thus overcoming the "black-box" problem of DL classifiers. This approach aimed to enhance model transparency, demonstrating competitive accuracy (90.6%) and computational efficiency compared to other large CNN models.

However, the efficacy of generating visual explanations through classification networks remains a contested area. Research demonstrated that visualization derived from segmentation masks is generally of higher quality and more precise than localization heatmaps obtained by classification-based approaches like Grad-CAM, although Grad-CAM requires less effort for data labeling. Addressing the low resolution and imprecision typically associated with standard XAI methods like Grad-CAM, the CXRNet framework proposed an Encoder-Decoder-Encoder multitask architecture for accurate classification and enhanced pixel-level visual explanation. CXRNet successfully produced sharper and more detailed visualizations compared to both the saliency map technique and Grad-CAM, demonstrating improved F1-score and classification accuracy when using lung-segmented images.

Collectively, these works underscore the evolution of AI for COVID-19 CXR analysis, moving from foundational detection models (like COVID-Net and CoroNet) toward sophisticated segmentation and severity assessment techniques (like infection maps and the two-stage segmentation workflow), all while facing the persistent challenge of establishing highly reliable and clinically validated explainability methods.

3.2 Literature Gap

Research into deep learning solutions for Chest X-ray (CXR) interpretation, particularly for COVID-19 detection and localization, reveals several areas where novel architectural choices and rigorous explainability benchmarking remain scarce. The identified gaps highlight opportunities for methodological advancements focused on model efficiency, transparency, and precision in localization.

3.2.1 Limited Exploration of Modern Encoder Architectures (ConvNeXt)

A significant gap exists in the adoption and systematic testing of modern Convolutional Neural Network (CNN) architectures, such as ConvNeXt, for medical imaging localization tasks.

While established architectures like DenseNet, ResNet152, Inception-v4, and VGG are widely used and benchmarked in classification and saliency studies, the use of ConvNeXt represents a key differentiating factor and novelty driver for research.

ConvNeXt, introduced recently in 2022, was not commonly used in existing papers, especially for localization tasks in the medical imaging domain.

This lack of use means that most prior work does not incorporate the potential benefits of this architecture for balancing performance and computational requirements.

Addressing this gap allows for one of the first explicit performance comparisons between Local Interpretable Model-agnostic Explanations (LIME) and Gradient-weighted Class Activation Mapping (Grad-CAM) tailored to the ConvNeXt architecture in a clinical setting.

3.2.2 Infrequent Utilization of LIME for Interpretability

Although Explainable Artificial Intelligence (XAI) is deemed critical for clinical translation due to the "black-box" nature of deep learning, the use of LIME, a powerful perturbation-based method, is less prevalent compared to gradient-based methods like Grad-CAM.

Grad-CAM is frequently employed and often considered the conventional choice for generating saliency maps, largely due to its accessibility and efficiency, being used for approximate localization and requiring less effort for data labeling compared to segmentation approaches.

In contrast, LIME is a perturbation-based technique that generates local segments ("superpixels") to explain predictions. The emphasis on LIME is often a direct response to the known limitations of traditional saliency methods like Grad-CAM, which produce coarse localization heatmaps and struggle to capture the geometric nuances of complex pathologies.

Despite findings demonstrating LIME's superiority over Grad-CAM in generating precise and boundary-respecting infection masks, the field generally leans towards simpler, gradient-based solutions. This suggests a gap in widespread adoption and thorough investigation of perturbation-based XAI techniques in routine deep learning pipelines for medical diagnostics.

3.2.3 Bias in Saliency Method Benchmarking

Prior seminal benchmarking studies on XAI methods in medical imaging have focused predominantly on gradient-based and specific post-hoc techniques, often omitting a quantitative analysis of LIME's localization capabilities relative to human experts or explicit segmentation methods.

One rigorous investigation systematically evaluated seven common saliency methods on CXR interpretation, including Grad-CAM, Grad-CAM++, integrated gradients, Eigen-CAM, DeepLIFT, LRP, and occlusion, across three common CNN architectures (DenseNet121, ResNet152, and Inception-v4). LIME was not included in this list of seven rigorously evaluated methods.

Other localization efforts compared against Grad-CAM or specialized proprietary methods like GSInquire, further indicating a research focus skewed away from LIME when seeking systematic quantification against Ground Truth segmentation data.

This benchmarking gap leaves open the question of whether LIME, being perturbation-based and generating highly resolved segments, consistently outperforms the wide array of gradient-based or deep feature propagation methods on standard localization metrics (mIoU and DICE Score) when dealing with complex or geometrically challenging pathologies.

3.2.4 Potential of LIME for Generating Unsupervised Infection Masks

The literature demonstrates a clear distinction between highly accurate, supervised segmentation methods (like U-Net, which require expensive pixel-level ground-truth annotations) and unsupervised saliency methods (like Grad-CAM, which are computationally cheap but provide coarse results). This dichotomy presents a gap where unsupervised methods capable of achieving high-resolution localization are critically needed.

The usage of LIME and Grad-CAM for creating infection masks is inherently unsupervised because these methods are applied post-hoc to classification models that were never exposed to pixel-level segmentations during training.

Crucially, research has shown that LIME possesses a strong potential to bridge the gap between coarse localization and expensive supervised segmentation. Quantitatively, LIME showed superior fidelity to boundary detection compared to Grad-CAM, achieving a mIoU of 0.3336 versus Grad-CAM's 0.1257, and a DICE Score of 0.4659 versus 0.1960.

The fundamental capability of LIME to generate segments that are more faithful to ground-truth infection boundaries suggests a significant opportunity to optimize LIME to produce clinically reliable infection masks without requiring explicit, fully supervised segmentation training. Further investigation into modifying and tuning LIME (e.g., using clinically meaningful neighbors like in KP-LIME) confirms this trajectory toward higher quality, unsupervised localization.

In essence, while gradient-based methods dominated early explainability research due to convenience, the field now faces the necessity of adopting sophisticated architectures (like ConvNeXt) and superior post-hoc methods (like LIME) to produce precise and trustworthy localization insights that are essential for reliable clinical decision support.

CHAPTER 4: METHODOLOGY

The methodology detailed herein describes the rigorous development and evaluation of Deep Learning (DL) workflows for Chest X-ray (CXR) interpretation, primarily focusing on disease classification and localization, while addressing the critical need for Explainable Artificial Intelligence (XAI) in high-stakes clinical settings. The overall approach employs a multi-stage strategy, integrating image preprocessing steps such as lung segmentation and contrast enhancement, frequently using CLAHE, with advanced Convolutional Neural Network (CNN) architectures. Diverse models—including VGG16, DenseNet, and novel architectures like ConvNeXt—are trained for classification on large benchmark datasets, notably COVID-Qu-Ex. A core component is the rigorous benchmarking of localization performance using two distinct XAI methods: LIME (Local Interpretable Model-agnostic Explanations), a perturbation-based technique, and Grad-CAM (Gradient-based Visualization Method), a saliency method. This comparison seeks to enhance transparency by evaluating whether LIME provides explanations superior to Grad-CAM's coarse heatmaps for precisely localizing complex pathologies, thereby improving the trustworthiness and clinical utility of DL systems.

4.1 Dataset (COVID-Qu-Ex)

The **COVID-QU-Ex** dataset is a large, publicly available resource used as a critical benchmark for localization tasks in CXR interpretation due to its scale and detailed annotations.

4.1.1 Content and Scale:

- The dataset was compiled by researchers at Qatar University.
- The total dataset consists of 33,920 Chest X-ray (CXR) images. The distribution of images in the full dataset includes 11,956 COVID-19 cases, 11,263 Non-COVID infections (Viral or Bacterial Pneumonia), and 10,701 Normal cases. For the entire dataset, ground-truth lung segmentation masks are provided.

4.1.2 Role in Research:

- COVID-Qu-Ex provides comprehensive ground-truth lung and infection segmentation masks, which are crucial for conducting quantitative localization evaluations. It is specifically used for the localization and detection.

4.1.3 Specific Subset for Infection Segmentation:

A specific subset of the COVID-QU-Ex dataset was utilized for studies related to infection segmentation and scoring, which contains the necessary ground-truth infection information:

- 2,913 COVID-19 CXRs
- 1,456 Normal CXRs
- 1,457 Non-COVID-19 CXRs

The COVID-19 CXRs in this subset had corresponding infection masks, which were sourced from the QaTaCov19 dataset in conjunction with COVID-QU-Ex data.

For one experimental validation, the 2,913 COVID-19 samples with ground-truth segmentation masks were divided into a training set of 1864, a validation set of 466, and a test set of 583. The images in this context had a pixel size of 256 x 256 and an 8-bit depth.

Table 4-1. Train Data Split

Class	**No. of Images**
COVID	**1864**
Healthy	**932**
Non-COVID	**932**

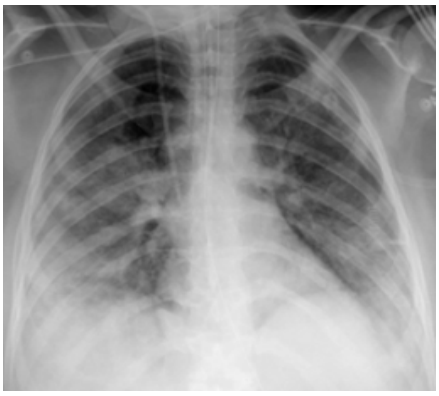

Figure 4-1. CXR image (COVID)

Table 4-2. Validation Data Split

Class	**No. of Images**
COVID	**466**
Healthy	**233**
Non-COVID	**233**

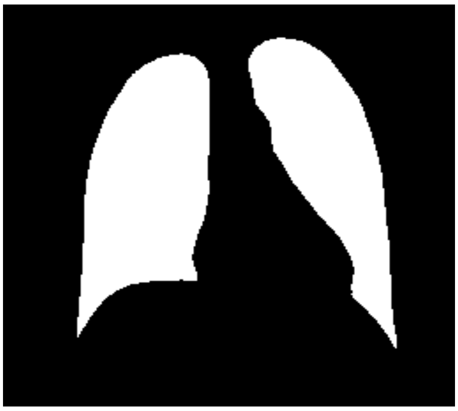

Figure 4-2. Lung Segmentation Mask

Table 4-3. Test Data Split

Class	**No. of Images**
COVID	**583**
Healthy	**291**
Non-COVID	**292**

4.2 Data Preprocessing

The primary goal of preprocessing is to prepare the CXR images for training deep convolutional neural networks (CNNs), ensuring consistent input size, enhanced feature focus, and sufficient data variation to improve model generalization.

4.2.1 Lung Segmentation: Applying Masks Present in the Dataset

This initial step utilizes ground-truth binary masks to isolate the lung fields, aligning with a common strategy in two-stage deep learning workflows for thoracic image analysis.

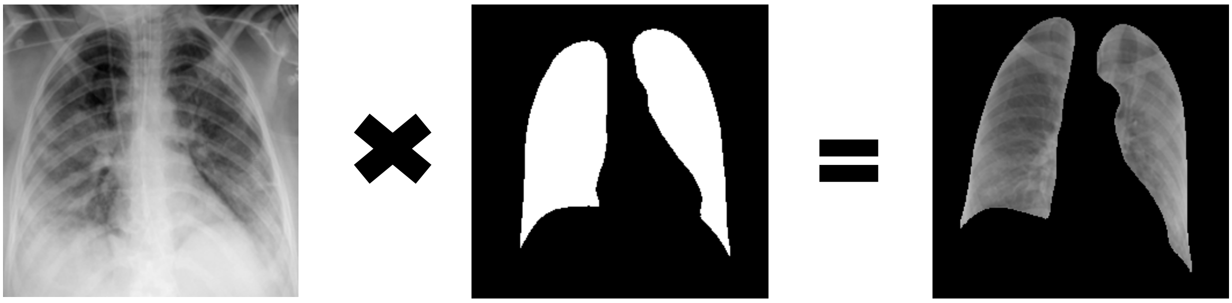

Figure 4-3. Lung Segmentation (Masking Process)

Table 4-4. Data preprocessing explanation

Step	Detail & Rationale
Action	Applying pre-existing pixel-level ground-truth segmentation masks.
Purpose	To achieve pixel-level localization of the lungs and perform the removal of unnecessary areas outside the lung region.
Technical Benefit	This preprocessing is crucial for enhancing model performance. Removing non-informative regions helps circumvent issues like the gradient vanishing problem and increases the gradient magnitude during back-propagation. The use of images with different pathologies for training segmentation models also helps improve the model's generalization ability.
Contextual Note	In advanced systems, this step often serves as Stage I of a workflow, mimicking the radiological procedure where a specialist first assesses the region of interest.

4.2.2 TorchVision Transforms (Augmentation and Normalization)

These steps are applied to the masked images to prepare them for ingestion by the neural network, managing data distribution and artificially increasing the dataset size through augmentation.

I. transforms.Resize((224, 224)): This step resizes all input images to a standard dimension of 224 x 224 pixels. This size is frequently used because it corresponds to the input dimensions expected by many popular CNN architectures, such as VGG16, InceptionV3, and Xception.

II. transforms.RandomHorizontalFlip(): This is a data augmentation technique applied during the training step. Its function is to artificially increase the size and diversity of the dataset. Horizontal flipping specifically helps the model generalize by ensuring it is invariant to the left/right orientation of the patient.

III. transforms.RandomRotation(10): Random rotation, typically applied, is another geometric data augmentation method. Data augmentation is applied during both hyperparameter tuning and training to act as a regularizer to help reduce overfitting.

IV. transforms.ToTensor(): This converts the images from the Python Imaging Library (PIL) format or NumPy arrays into PyTorch Tensors, which is the required format for deep learning model input in the PyTorch framework.

V. transforms.Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]): This normalization step standardizes the input data distribution across all three color channels. The specific mean and standard deviation values used are the ImageNet statistics. This is critical for models utilizing Transfer Learning (TL), where the CNN layers (e.g., DenseNet, ResNet) are initialized with weights pre-trained on the vast ImageNet dataset, ensuring the input distribution matches the expected distribution of the pre-trained weights.

The methodology for the Convolutional Neural Network (CNN) Model Training focuses on training diverse CNN architectures for classification, which subsequently enables the generation of infection masks using explainable AI (XAI) techniques.

4.3 Model Selection and Architectural Approach

The CNN methodology involves training and evaluating classification performance using four diverse CNN architectures: VGG16, DenseNet121, DenseNet201, and ConvNeXt.

- VGG16 serves as a deep convolutional baseline architecture for the classification task. VGG16 uses 3x3 filters throughout its 16 layers and has a uniform architecture.
- DenseNet architectures (DenseNet121 and DenseNet201) are included for evaluation due to their popularity, strengthening feature propagation, and ability to encourage feature reuse by connecting each layer to every subsequent layer in a feed-forward fashion. DenseNet121 is a deeply layered network (121 layers) designed to maximize information flow.
- ConvNeXt is included as a recent model (2022) to drive novelty and expand the understanding of its explainability within the medical imaging domain, as it was not commonly used in existing papers for localization tasks.

4.3.1 Image Preparation

I. Input Image Size: All input Chest X-ray (CXR) images are standardized to a fixed resolution of 224 x 224 pixels across the dataset (COVID-Qu-Ex) to ensure consistent comparison of the resulting infection masks. In general, CNN methods, CXRs are often processed by resizing them, converting them to grayscale, and normalizing the pixel values.

II. Transfer Learning and Pretraining: Transfer learning is employed for initialization, as models like DenseNet variants and VGG16 were frequently pre-trained on large datasets like ImageNet. Transfer learning helps overcome the limitation of a limited dataset size and reduces training time.

4.3.2 Training Configuration and Hyperparameters

These masks can be generated by U-Net consistently with supervised learning for up to 99 percent accuracy, making supervised segmentation methods appropriate for this task.

The model training relies on specific hyperparameter settings and optimization techniques:

Table 4-5. Model Hyperparameters for CNN model training

Model Architecture	**Batch Size**	**Epochs**	**Optimizer**	**Learning Rate (LR)**
VGG16	32	20	Adam	1e-5
DenseNet121	32	Head Classifier: 5 Partial Unfreeze: 10 Full Unfreeze: 10	Adam	1e-5 to 1e-3 (Scheduler: ReduceLROnPlateau)
DenseNet201	32	Head Classifier: 5 Partial Unfreeze: 10 Full Unfreeze: 10	Adam	1e-5 to 1e-3 (Scheduler: ReduceLROnPlateau)
ConvNeXt	16 (due to memory constraints)	Head Classifier: 5 Partial Unfreeze: 10 Full Unfreeze: 10	AdamW	Head Classifier: 1e-3 Partial Unfreeze: 5e-4 Full Unfreeze: 1e-5

I. Selection of Architectures: The system design is driven by the goal of benchmarking XAI methods across diverse architectures. VGG16 serves as a deep convolutional baseline, while DenseNet (DenseNet121 and DenseNet201) is included for its feature reuse capability and wide use. ConvNeXt is included as a recent model (2022) to drive novelty and expand the understanding of explainability in the medical domain.

II. Optimizer and Learning Rate: The Adam optimizer or its variants generally proved to be the optimal solution for network training and effective convergence in related segmentation and classification tasks referenced in the sources. Optimal learning rates typically fall into the range of 1e-3 to 1e-5.

III. Visualization Method (LIME): The proposed system relies on LIME for unsupervised mask generation, hypothesized to be superior to Grad-CAM for creating precise and boundary-respecting infection masks. LIME is a perturbation-based technique that generates explanations that are locally faithful to the complex black-box model, typically represented by linear models over interpretable components like superpixels. The result is visualized using heat maps as a mask for the classified images to mark boundaries.

IV. Evaluation Datasets: The use of COVID-QU-Ex (providing ground-truth lung and infection segmentation masks) is critical for quantitative localization evaluation and comparison against the gold standard annotations.

4.3.3 Computational Environment

The models are implemented using standard frameworks like Keras/TensorFlow and executed on hardware environments equipped with high-performance GPUs, such as the Nvidia Tesla T4 (with 13 GB GPU memory) or Nvidia Tesla P100 GPUs, which are capable of supporting the specified memory-intensive training configurations.

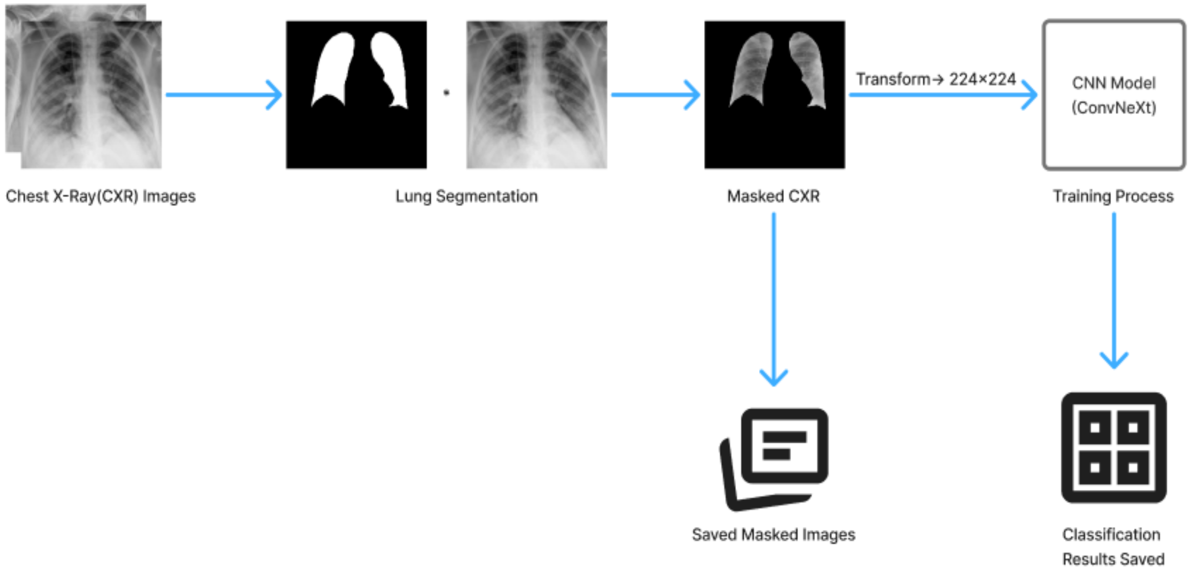

Figure 4-4. Workflow Diagram for CNN training Pipeline

4.4 Generating Infection Maps using Unsupervised XAI

The localization step is critical because, alongside detection, COVID-19-related infection localization is a crucial problem for accurate assessment of the infection location and disease severity. Utilization of LIME and Grad-CAM is categorized as an unsupervised method because these saliency techniques provide post-hoc interpretability of models that were never exposed to pixel-level segmentations (ground truth) during training.

The methodology for generating these infection maps involves two distinct approaches: the perturbation-based method (LIME) and the gradient-based method (Grad-CAM).

4.4.1 Rationale for Choosing Unsupervised XAI Methods

This methodology focuses on rigorously benchmarking the localization fidelity of LIME against Grad-CAM across diverse Convolutional Neural Network (CNN) architectures.

I. Local Interpretable Model-agnostic Explanations (LIME)

LIME is a modular and extensible algorithm designed to explain the predictions of any classifier by approximating it locally with a simple, interpretable model.

(a) Interpretable Representation: For image classification, the input image is transformed into an interpretable data representation, typically a binary vector indicating the presence or absence of a contiguous patch of similar pixels, often called a superpixel.

(b) Generating Perturbations: To explain a prediction for a given image, LIME segments the input into these "superpixels". Simulated data points (neighbors) are generated in the neighborhood of the original image by systematically turning these superpixels off and on. Turning segments off usually means substituting the segment with a default color like grey.

(c) Local Fidelity: LIME seeks to identify an interpretable model (e.g., a sparse linear model) that is locally faithful to the complex black-box classifier. This is achieved by weighting the sampled neighbors based on their proximity to the original image and then fitting the simple model locally to approximate the true model.

(d) Output Generation: The LIME process generates a set of interpretations defining each feature's input to a prediction for a specific sample, resulting in a local understanding. LIME Visualization (Infection Maps): The visualization stage uses the LIME output to create interpretable infection maps

LIME generates visualizations, often presented as heat maps, which act as a mask for the classified images to mark boundaries in the classification decision. These heat maps highlight the specific image segments that contributed most to the model's classification decision. For three-class explanations (e.g., COVID-19, Normal, Pneumonia), LIME can show the regions that contributed *toward* the predicted class (often highlighted in green) and those that contributed *against* it (often highlighted in red).

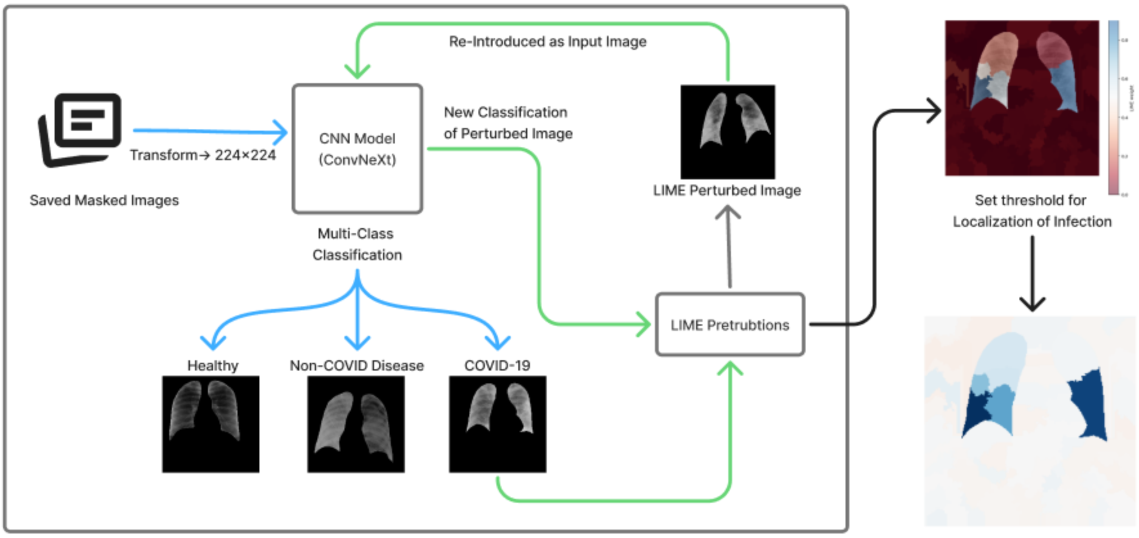
Re-Introduced as Input Image
Saved Masked Images
Transform→ 224×224
CNN Model (ConvNeXt)
New Classification of Perturbed Image
LIME Perturbed Image
Multi-Class Classification
LIME Pretrubtions
Healthy
Non-COVID Disease
COVID-19
Set threshold for Localization of Infection

Figure 4-5.Workflow Diagram of Generating Infection Map Pipeline

II. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is a popular visualization technique and a gradient-based localization method that can be applied to a wide variety of CNN model families, including those with fully-connected layers (like VGG). Grad-CAM is a generalization of the Class Activation Map (CAM) method.

(a) Mechanism of Grad-CAM: Grad-CAM uses the gradient information flowing into the convolutional layers to assign importance values to each neuron for a particular classification decision.

(b) Gradient Flow: The technique utilizes the gradients of any target concept (e.g., the score for a specific class before the softmax layer) flowing into the final convolutional layer.

(c) Neuron Importance Weight : These gradients are spatially global-average-pooled over the width and height dimensions of the feature maps to obtain the neuron importance weights . This weight represents a partial linearization of the deep network downstream from and captures the 'importance' of feature map k for the target class c.

(d) Localization Map Generation : The Grad-CAM localization map is produced by performing a weighted combination of the forward activation maps using the importance weights. A Rectified Linear Unit (ReLU) is then applied to this linear combination:

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} \quad \alpha^{c}_{k} . A^{k}\right) \qquad —— \qquad Equation\ 1$$

(e) The ReLU ensures that only the features that have a positive influence on the class of interest are highlighted.

Grad-CAM produces a coarse localization heatmap highlighting the important regions in the image for predicting the target concept. The visualization typically results in a coarse heat map of the same size as the convolutional feature maps, which is subsequently interpolated to the original image resolution. This reliance on coarse resolution features limits Grad-CAM's ability to capture the geometric nuances and boundaries of complex pathologies. The coarse heatmap represents where the model has to look to make the particular decision.

4.4.2 Infection Map Generation and Visualization

The final infection maps are created by transforming the raw output of the XAI methods into a visual representation superimposed on the original CXR image. This step is conceptually derived from infection map generation using trained segmentation networks:

(a) Mask Generation: The output from the XAI methods serves as a prediction mask (or probability map), where pixel values reflect the model's confidence or activation in that region for the predicted class.

(b) Color Transformation: An RGB-based color transform, such as the jet color scale, is applied to the prediction mask to obtain a color-coded probability measure.

(c) Superimposition: The infection map is generated by reflecting the network prediction ranslucently onto the Chest X-ray image. For visualization clarity, regions with zero probabilities are not shown. Areas with high weights or positive influence (often highlighted in yellow/red on the map) indicate critical locations where the deep learning algorithm detected COVID-19 signals or abnormalities in the lungs.

CHAPTER 5: RESULTS

5.1 COVID-Qu-Ex

The COVID-Qu-Ex dataset, compiled by researchers at Qatar University and publicly available on Kaggle, is recognized for its scale and detailed annotations, serving as a critical benchmark for localization tasks in Chest X-ray (CXR) interpretation. While the full dataset encompasses 33,920 CXR images for lung segmentation studies, the subset utilized for COVID-19 infection segmentation and scoring contains key ground-truth information necessary for this comparative analysis.

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, and 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). The presence of these gold-standard infection masks is vital, as it facilitates the quantitative evaluation of the predicted saliency maps generated unsupervised by LIME and Grad-CAM for localization performance.

5.1.1 Tertiary and Binary - Class COVID Classification

Utilizing large CXR datasets like COVID-Qu-Ex, which contains samples across three categories—COVID-19, Non-COVID infections (Viral or Bacterial Pneumonia), and Normal patients—often conduct multiclass classification to distinguish these three groups. For evaluating the performance of models in this 3-class classification task, Accuracy, Precision, and AUC are critical metrics.

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

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \qquad \text{——} \qquad Equation\ 2$$

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

$$Precision = \frac{TP}{TP+FP} \qquad \text{——} \qquad Equation\ 3$$

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.

$$AUC = \sum_{i+1}^{n+1}(FPR_{i+1} - FPR_{i})\frac{TPR_{i+1} + TPR_{i}}{2} \qquad \text{——} \qquad \textit{Equation 4}$$

Table 5-1. Class Classification Results

Model	**CoroNet** [Khan et al.]	**COVID-Net** [Wang et al.]	**VGG16**	**DenseNet121**	**DenseNet201**	**ConvNeXt** **[Base]**
Parameters (millions)	33	11.75	130	8	20	89
Dataset Split	500: Healthy, 500: Pneumonia 157: COVID-19	13,975 CXR images from 13,870 patients	1,456: Healthy 1457: Non-COVID 2,913: COVID-19			
Accuracy	0.902	0.93	0.85	0.89	0.90	**0.93**
Precision (COVID)	0.97	0.97	0.94	0.92	0.94	**0.97**
Precision (Non-COVID)	**0.92**	0.913	0.75	0.81	0.85	0.87
Precision (Healthy)	0.87	0.905	0.82	0.91	0.88	**0.91**
Precision (Average)	0.92	0.93	0.83	0.88	0.89	**0.92**
Macro-AUC	-	-	0.94	0.97	0.98	**0.99**

Table 5-2. Class Classification Results

Model	**ConvNeXt**	**Swin Transformer** [Ma 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	0.983
Macro-AUC	1.00	-	-

Note: The dataset used in CoroNet is relatively small i.e. 29 COVID cases and 72 Normal Cases, which increases the risk of overfitting and may inflate performance metrics. The near-perfect precision values should therefore be interpreted cautiously, as they may not generalize to larger, more diverse clinical populations. Additional validation on larger, multi-center datasets is required to confirm model robustness.

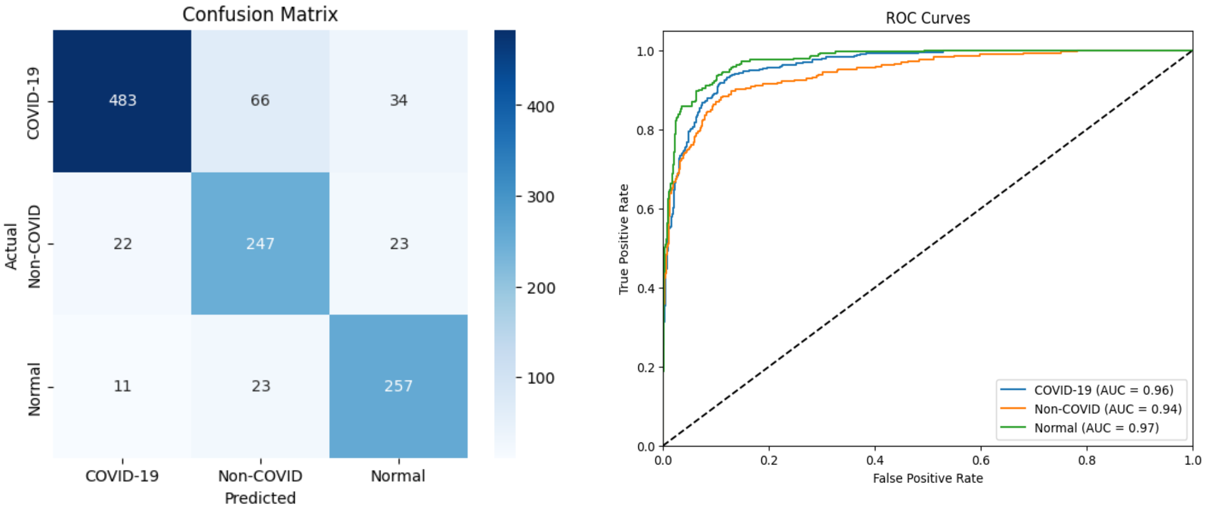

Figure 5-1. VGG-16 Result Graphs

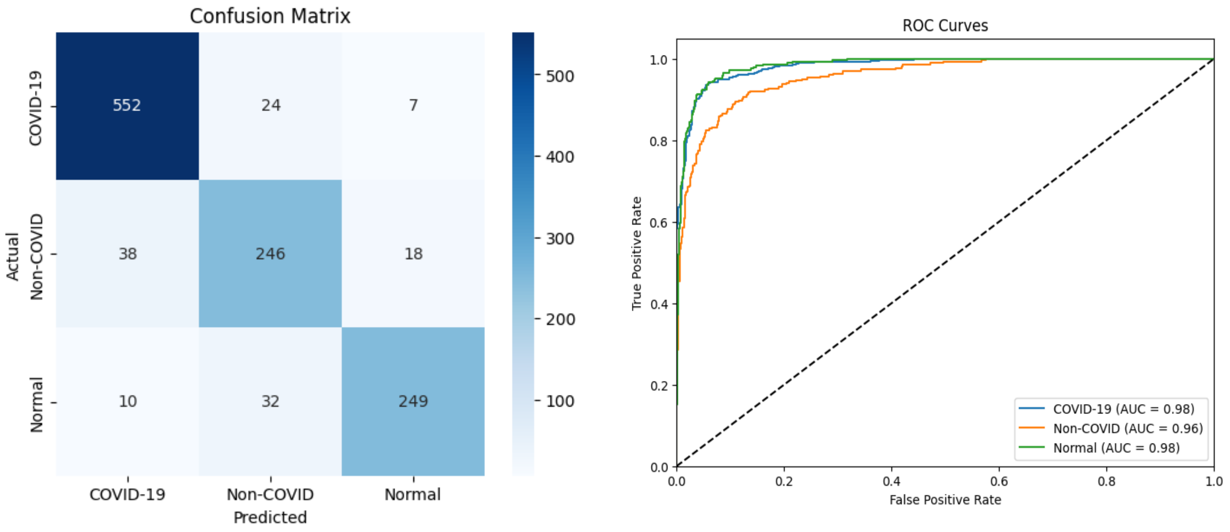

Figure 5-2. DenseNet-121 Result Graphs

DenseNet-121 generally performs better than VGG because it uses dense connectivity, where each layer receives feature maps from all previous layers. This leads to stronger feature reuse, reduces overfitting, and makes the network far more parameter efficient. As a result, DenseNet learns richer representations with fewer weights, giving higher accuracy (0.89 vs 0.85) while being faster and lighter than VGG.

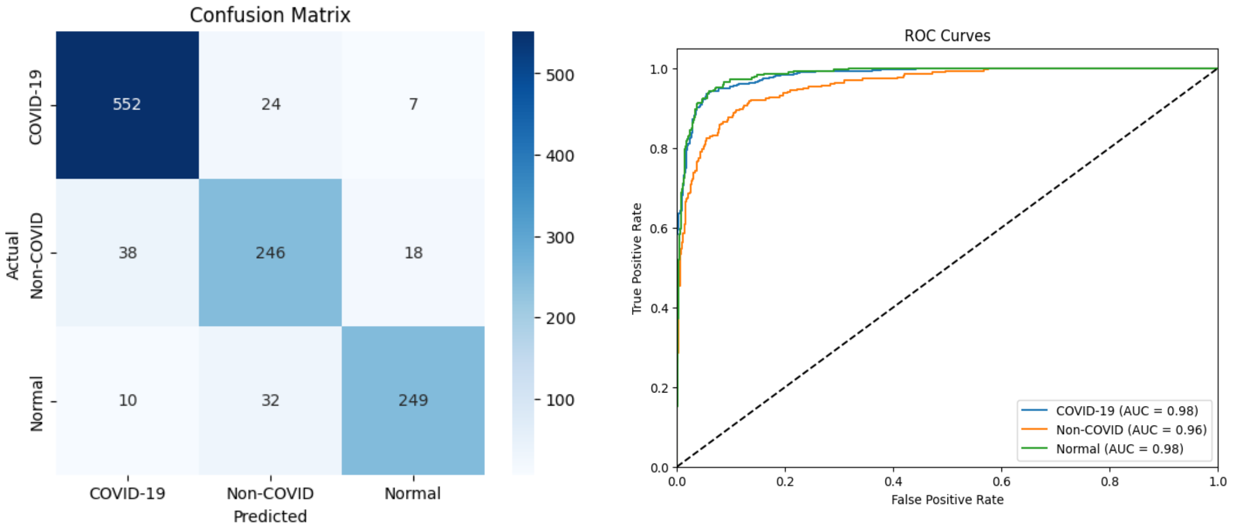

Figure 5-3. DenseNet-201 Result Graphs

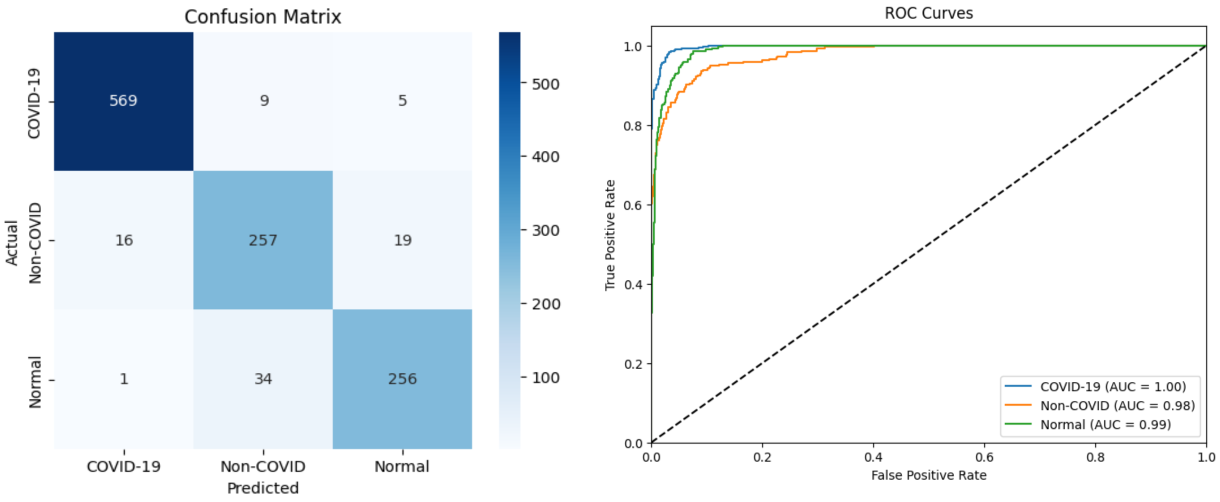

Figure 5-4. ConvNeXt-Base Result Graphs

ConvNeXt has a more modern architecture inspired by Transformers, with improved normalization, depthwise convolutions, and better scaling. It captures larger spatial patterns and finer details compared to DenseNet. With strong ImageNet pretraining and optimized design, ConvNeXt often achieves higher AUC and better generalization on medical imaging tasks.

5.1.2 Localization\Saliency Map Comparison

The standard 3-class classification task involves assigning a CXR image to one of three distinct diagnostic groups present in datasets like COVID-Qu-Ex:

I. COVID-19: Patients confirmed to have the SARS-CoV-2 infection.

II. Non-COVID Infections: Patients suffering from other respiratory illnesses, such as Viral or Bacterial Pneumonia.

III. Normal: Healthy patients showing no disease or pathology.

This multiclass approach aims to provide valuable insights into the detection and differentiation of COVID-19 from other similar thoracic diseases and healthy cases. It represents a significantly more challenging task compared to simple binary classification, as Non-COVID infections often exhibit high visual similarity to COVID-19 infection in the lungs.

Utilizing large CXR datasets like COVID-Qu-Ex, which contains samples across three categories—COVID-19, Non-COVID infections (Viral or Bacterial Pneumonia), and Normal patients—often conduct multiclass classification to distinguish these three groups. For evaluating the performance of models in this infection mask generation task, MIOU, DICE, and Hausdorff Distance are critical metrics.

I. **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:

$$IoU = \frac{TP}{FN+FP+TP} \qquad \text{———} \qquad \textit{Equation 5}$$

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

$$DICE\ Score = \frac{2 \times IoU}{IoU + 1} \qquad \text{———} \qquad \textit{Equation 6}$$

III. **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:

$$HD\ (Gt, Pd) = mean_{p_{pd}\ \epsilon\ Pd} min_{p_{gt}\ \epsilon\ Gt} ||p_{gt} - p_{pd}||^2 \qquad \text{———} \qquad \textit{Equation 7}$$

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).

Table 5-3. Class Localization Results

Technique	**LIME**			**Grad-CAM**		
Model	DenseNet121	DenseNet201	**ConvNeXt Base**	DenseNet121	DenseNet201	**ConvNeXt Base**
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

Table 5-4. Class Localization Results

Technique	LIME	Grad-CAM
Model	**ConvNeXt Base**	**ConvNeXt Base**
MIOU	**0.3950**	0.2355
DICE Score	**0.5312**	0.3449
Hausdorff Distance	**81.9963**	92.5856

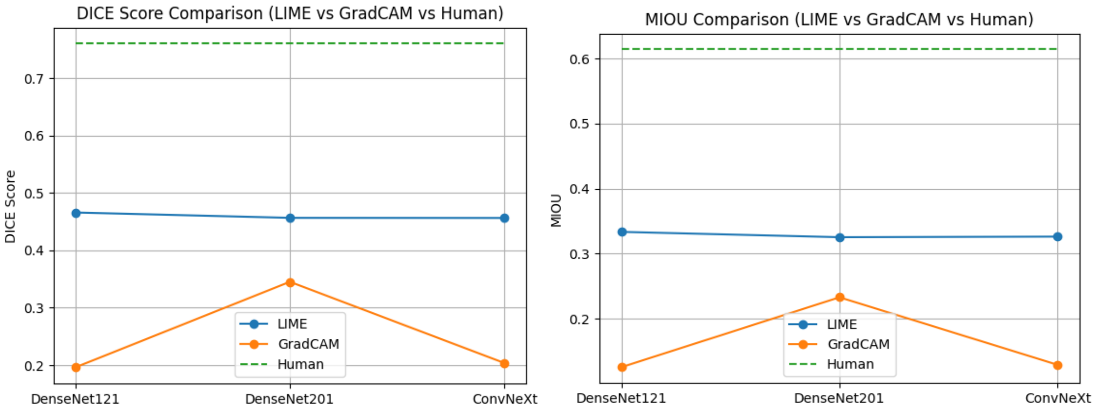

Figure 5-5. Comparison Graphs between LIME and Grad-CAM

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.

I. Resolution and Precision of Explanations:

- LIME is a perturbation-based technique that generates segments (or superpixels) to explain predictions. By integrating neighbor images (in advanced versions like KP-LIME), segments in the generated data can be set to genuine portions of X-rays rather than masked patches of grey, introducing clinical variability while respecting locality.
- Grad-CAM is a gradient-based visualization method that produces a localization heatmap by utilizing gradients flowing into the final convolutional layer. This process results in coarse localization heatmaps and is usually used for approximate localization.
- The visualization and decision-making based on segmentation masks (as produced by LIME) are inherently of higher quality than the coarse localization heatmaps obtained by Grad-CAM.

II. Inability to Capture Geometric Nuance:

- Grad-CAM's reliance on coarse resolution features limits its ability to capture the geometric nuances of complex pathologies. Saliency methods generally fail to capture the geometric nuances of a given pathology, instead producing coarse, low-resolution heat maps.
- The relatively small heat maps produced by Grad-CAM (e.g., 14×14), which are interpolated to the original image dimensions, result in coarse resolutions.
- In contrast, pixel-level segmentation, which is conceptually closer to LIME's output segments, offers greater precision. LIME was found to consistently outperform Grad-CAM in generating precise and boundary-respecting infection masks.

III. Fundamental Approach Distinction:

The findings confirm that segmentation-based local explanations (LIME) are superior to gradient-based localization heatmaps (Grad-CAM). Grad-CAM, despite generally performing better than *other* saliency methods, struggles to handle pathologies that are smaller in size and have shapes that are more complex. LIME's superior performance reflects its ability to produce segments that align better with the ground-truth segmentation masks provided by datasets like COVID-Qu-Ex.

CHAPTER 6: CONCLUSION AND FUTURE SCOPE

6.1 LIME vs GradCAM

The project findings conclusively demonstrate the superiority of Local Interpretable Model-agnostic Explanations (LIME) over Gradient-weighted Class Activation Mapping (Grad-CAM) for generating precise pathology localization masks in medical imaging. This conclusion is robustly supported by quantitative evaluation against gold-standard infection masks from datasets like COVID-Qu-Ex.

6.1.1 Superiority of LIME in Localization

- Quantitative Performance: LIME consistently achieved significantly higher localization scores compared to Grad-CAM across all evaluated architectures. For instance, when interpreting the DenseNet121 model on the localization task, LIME achieved a Mean Intersection over Union (mIoU) of 0.3336 , which is notably higher than Grad-CAM's mIoU of 0.1257. Similarly, LIME attained a DICE Score (F1-score for segmentation) of 0.4659, more than double Grad-CAM's score of 0.1960.
- Precision and Boundary Respecting: LIME produces explanations that are significantly more faithful to the ground-truth infection boundaries. The Hausdorff Distance, where a lower value indicates better performance, reinforced this, with LIME achieving 93.0492 compared to Grad-CAM's 126.3200.
- Mechanism Advantage: LIME is a perturbation-based technique that generates segments (or superpixels) to explain predictions. This approach yields predicted infection masks that are precise and boundary-respecting.

6.1.2 Limitations of Grad-CAM

- Coarse Localization Problem: Grad-CAM is a gradient-based visualization method that typically produces coarse localization heatmaps. This coarseness fundamentally restricts its ability to capture the geometric nuances of complex pathologies.
- Approximate Use: Grad-CAM is generally used for approximate localization and requires fewer resources for data labeling and training compared to higher-resolution segmentation techniques. However, visualization and decision-making based on segmentation masks (like LIME's output) are considered inherently of higher quality.

- Benchmark Performance: While Grad-CAM typically performs better than other saliency methods evaluated in external benchmarks, all seven saliency methods tested were found to perform significantly worse compared with the human expert benchmark.
- In conclusion, for medical diagnostic purposes where fine-resolution explanations and high localization fidelity are critical, segmentation-based local explanations (LIME) proved to be the superior methodology compared to gradient-based localization heatmaps (Grad-CAM).

6.2 ConvNeXt vs other models

The project successfully benchmarked the modern ConvNeXt architecture against widely adopted models like VGG16 and DenseNet, offering key insights into its performance and explainability within the context of Chest X-ray (CXR) interpretation.

6.2.1 ConvNeXt Classification and Novelty

- Superior Classification Performance: On the 3-class classification task using the COVID-Qu-Ex dataset, the ConvNeXt model achieved the highest overall performance metrics among the benchmarked classification models. ConvNeXt attained an Accuracy of 0.93 and a Macro-AUC of 0.99. In comparison, DenseNet201 achieved an Accuracy of 0.90 (AUC 0.98), DenseNet121 achieved an Accuracy of 0.89 (AUC 0.97), and VGG16 achieved an Accuracy of 0.85 (AUC 0.94).
- Architectural Novelty: The inclusion of ConvNeXt (a recent model from 2022) served as a key differentiating factor and novelty driver for the research. The project executed one of the first explicit performance comparisons between LIME and Grad-CAM specifically on the ConvNeXt architecture within the medical imaging domain.

6.2.2 Comparison with Other Models

- VGG16 Performance and Efficiency: The VGG16 model, serving as the deep convolutional baseline architecture, demonstrated competitive performance when coupled with LIME. On the largest dataset used for performance measurement (Dataset 3), VGG16 with LIME achieved 90.6% accuracy. Critically, this model combination presented the smallest computational cost (10,936 seconds) when compared to other high-performing competitors such as Xception, Inception V4, ResNet50, XNet, and AlexNet on that same dataset.

- DenseNet Architecture: DenseNet architectures (DenseNet121 and DenseNet201) were evaluated due to their widespread use and characteristic design that strengthens feature propagation and encourages feature reuse by connecting layers in a feed-forward fashion.

6.2.3 Explainability Challenges with ConvNeXt

- LIME Stability: LIME demonstrated stability across all tested architectures, including ConvNeXt, yielding consistent mIoU (0.3263) and DICE scores (0.4564) that were substantially higher than Grad-CAM's scores on the same model.
- Grad-CAM Deterioration: For the ConvNeXt architecture specifically, Grad-CAM performance dropped. This decline is attributed to the fact that ConvNeXt uses deeper, more complex hierarchical feature representations, which makes its final activation maps harder to interpret. As a result, ConvNeXt's architecture produces more diffuse and less spatially localized gradients, reducing the localization accuracy metrics for Grad-CAM compared to its performance on other models like DenseNet201.
- This comprehensive evaluation confirms that while modern architectures like ConvNeXt provide superior classification accuracy, the choice of the appropriate Explainable AI technique (LIME over Grad-CAM) is paramount for ensuring reliable localization results, especially when dealing with complex or novel CNN designs.\

6.3 Future Scope

6.3.1 Advancements in Explainable AI (XAI) Methodology

Future work should prioritize improving the resolution, consistency, and reliability of interpretable localization techniques, particularly extending the gains achieved with LIME:

- Improving LIME for Clinical Fidelity: Further investigation should explore combining auxiliary medical report information with learned Variational Autoencoder (VAE) embeddings to identify higher-quality neighbors for the Local Interpretable Model-agnostic Explanations (LIME) process.
- Scaling LIME Training: To provide a more refined neighborhood understanding, the VAE training should be scaled up beyond the current roughly 10k images available in small subsets. This effort could also investigate the theoretical properties and computational optimizations of LIME, such as using parallelization and GPU processing, to deliver accurate, real-time explanations suitable for a clinical workflow.

- Systematic XAI Evaluation: Given the development of modified LIME architectures (e.g., K*LIME and KP-LIME) which introduce clinical variability into segment selection, a systematic evaluation strategy is required to quantitatively assess the quality of these explanations against default LIME.
- Edge Identification: Since segmenting infected regions with complex contours remains a challenge, future research should focus on further improving the network's ability to accurately identify edge information in infected areas.
- Saliency Method Reliability: Continue investigating ways to ensure saliency methods like Grad-CAM consistently guarantee reliability against all possible input transformations, which is a required research agenda for deep neural networks in domains like medicine.

6.3.2 Architectural and Performance Enhancements

Building upon the successful benchmarking of modern CNNs, future research should focus on consolidating performance through architectural optimization:

- Advanced Ensemble Modeling: The success of combining multiple CNNs into an ensemble model suggests that future designs should leverage different sets of top-performing models to create even more accurate and resilient overall performance. This approach mitigates the weaknesses inherent in individual models.
- Validating Lightweight Models: Further work is needed to validate that lightweight networks, previously observed to not generalize well for COVID-19 and pneumonia cases, can achieve stability and robust generalization across diverse patient groups and diseases.

6.3.3 Expansion of Clinical Utility and Prognosis

Moving beyond basic classification and localization, the models should be extended to support complex clinical decision-making:

- Risk Stratification and Survival Analysis: The proposed detection pipelines can be extended to perform risk stratification for survival analysis, predicting a patient's risk status, and forecasting hospitalization duration. This capability would be highly useful for triaging, managing patient populations, and individualizing care plans.
- Multi-Model XAI Integration: The explicit comparison of LIME (perturbation-based) and Grad-CAM (gradient-based) should be extended to incorporate additional categories of XAI, such as Case-based Explanations (e.g., GANterfactuals for contrastive reasoning or TCAV for concept detection) or Textual Explanations (e.g., Visual Question

Answering/Image Captioning) to provide clinicians with a more robust, multi-modal understanding of the model's reasoning.

6.3.4 Data Diversity and Generalization

To ensure the deployability and fairness of AI systems, addressing data limitations is paramount:

- Dataset Balancing and Aggregation: Future work should address the limitation of models being trained on imbalanced datasets (e.g., where normal/pneumonia cases vastly outnumber COVID-19 cases). Aggregating vetted open-source datasets, potentially including the COVID-QU-Ex dataset, can help achieve a more representative ratio of target cases.
- Geographic and Demographic Inclusion: To enhance generalization, future training datasets must include Chest X-rays from different countries and geographic regions to ensure the models are representative of a wider demographic variety of patients worldwide.

REFERENCES

1. A. M. Tahir, M. E. H. Chowdhury, A. Khandakar, Y. Qiblawey, U. Khurshid, S. Kiranyaz, N. Ibtehaz, M. S. Rahman, S. Al-Madeed, S. Mahmud, M. Ezeddin, K. Hameed, & T. Hamid. (2021). COVID-19 Infection Localization and Severity Grading from Chest X-ray Images,. *Computers in Biology and Medicine*, *139*, 105002. https://doi.org/10.1016/j.compbiomed.2021.105002
2. Danilov, V.V., Litmanovich, D., & Proutski, A. (2022, July 27). Automatic scoring of COVID-19 severity in X-ray imaging based on a novel deep learning workflow. *Sci Rep*, *12*(12791). https://doi.org/10.1038/s41598-022-15013-z
3. Degerli, A., Ahishali, M., Yamac, M., & et al. (2021). COVID-19 infection map generation and detection from chest X-ray images. *Health Inf Sci Syst*, *9*(15). https://doi.org/10.1007/s13755-021-00146-8
4. Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger. (2016, August 25). *[1608.06993] Densely Connected Convolutional Networks*. arXiv. Retrieved November 21, 2025, from https://arxiv.org/abs/1608.06993
5. *Indian Council of Medical Research (ICMR) Ethical Guidelines for AI in Healthcare*. (2017). Indian Council of Medical Research. Retrieved November 21, 2025, from https://www.icmr.gov.in/icmrobject/custom_data/pdf/Ethical-guidelines/Ethical_Guidelines_AI_Healthcare_2023.pdf
6. *MEDICAL DEVICES RULES, MINISTRY OF HEALTH AND FAMILY WELFARE (Department of Health and Family Welfare)*. (2017). CDSCO. Retrieved November 21, 2025, from https://cdsco.gov.in/opencms/resources/UploadCDSCOWeb/2022/m_device/Medical%20Devices%20Rules%2C%202017.pdf
7. M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam. (2020). Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access*, *8*, 132665–132676. https://doi.org/10.1109/ACCESS.2020.3010287
8. Mohan, K. (n.d.). *THE DIGITAL PERSONAL DATA PROTECTION ACT, 2023 (NO. 22 OF 2023) An Act to provide for the processing of digital personal data in*. Ministry of Electronics and Information Technology. Retrieved November 21, 2025, from https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf

9. Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, Saining Xie. (2023, January 2). *[2301.00808] ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders*. arXiv. Retrieved November 21, 2025, from https://arxiv.org/abs/2301.00808

10. Saporta, & Gui, X., Agrawal, A. et al. (2022, October 10). Benchmarking saliency methods for chest X-ray interpretation. *Nat Mach Intell*, *4*, 867–878. https://doi.org/10.1038/s42256-022-00536-x

11. Simonyan, K., & Zisserman, A. (2014, September 4). *[1409.1556] Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv. Retrieved November 21, 2025, from https://arxiv.org/abs/1409.1556

12. Socolov, A. (n.d.). *LIME-for-medical-imaging*. GitHub. Retrieved November 21, 2022, from https://github.com/alexandrusocolov/LIME-for-medical-imaging

13. T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. A. Kashem, M. Islam, S. Al-Maadeed, S. Zughaier, M. Khan, and M. Chowdhury. (2021). Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection Using Chest X-ray Images. *Computers in Biology and Medicine*, *132*, 104319. https://doi.org/10.1016/j.compbiomed.2021.104319

14. Gongtao Yue, Chen Yang, Zhengyang Zhao, Ziheng An, & Yongsheng Yang. (2023). ERGPNet: lesion segmentation network for COVID-19 chest X-ray. *Computational Physiology and Medicine*, *14*. https://doi.org/10.3389/fphys.2023.1296185

15. Wang L, Lin ZQ, Wong A (2020) COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Sci Rep 10(1):1–12*

16. Khan AI, Shah JL, Bhat MM (2020) CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest X-ray images. *Compute Methods Programs Biomed 196:105581*

17. Ma, Y., & Lv, W. (2022). Identification of pneumonia in chest X-ray image based on transformer. International Journal of Antennas and Propagation, 2022, Article 5072666. https://doi.org/10.1155/2022/5072666

18. Anas Mohammed Tahir. (n.d.). *COVIDQU* [Data set]. Kaggle. https://www.kaggle.com/datasets/anasmohammedtahir/covidqu

19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. arXiv. https://arxiv.org/abs/1602.04938

20. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

21. Zhang, X., Han, L., Sobeih, T., Han, L., Dempsey, N., Lechareas, S., Tridente, A., Chen, H., & White, S. (2023). CXR-Net: A multitask deep learning network for explainable and accurate diagnosis of COVID-19 pneumonia from chest X-ray images. *IEEE Journal of Biomedical and Health Informatics, 27*(2), 980–991. https://doi.org/10.1109/JBHI.2022.3220813

22. Lua, J. W., Socolov, A., & Szep, A. (2023). *Modifying LIME for medical imaging* [Computer software]. GitHub. https://github.com/alexandrusocolov/LIME-for-medical-imaging

23. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., … Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590

24. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3462 3471) . https://doi.org/10.1109/CVPR.2017.369

25. Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Peng, Y., … Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data, 6*,317. https://doi.org/10.1038/s41597-019-0322-0

26. Garcia Santa Cruz, B., Bossa, M. N., Sölter, J., & Husch, A. D. (2021). Public COVID-19 X-ray datasets and their impact on model bias: A systematic review of a significant problem. *Medical Image Analysis, 74*, 102225. https://doi.org/10.1016/j.media.2021.102225

27. Ait Nasser, A., & Akhloufi, M. A. (2023). A review of recent advances in deep learning models for chest disease detection using radiography. *Diagnostics, 13*(1), 159. https://doi.org/10.3390/diagnostics13010159

28. Park, S., Kim, G., Oh, Y., Seo, J. B., Lee, S. M., Kim, J. H., Moon, S., Lim, J. K., & Ye, J. C. (2022). Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. *Medical Image Analysis, 75*, 102299. https://doi.org/10.1016/j.media.2021.102299

29. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems, 31 (NeurIPS)*. https://doi.org/10.48550/arXiv.1810.03292

30. van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis, 79*, 102470. https://doi.org/10.1016/j.media.2022.102470

31. Brima, Y., & Atemkeng, M. (2024). Saliency-driven explainable deep learning in medical imaging: Bridging visual explainability and statistical quantitative analysis. *BioData Mining, 17*, 18. https://doi.org/10.1186/s13040 -024-00370-4