# CAPSTONE PROJECT:

## The Explainability Value in CNN-Based Medical Imaging Diagnostics: A Comparative Study of Grad-CAM, LIME, and SHAP

Student: Nino Miljkovic

Hawaii Pacific University

DSCI 7000: Capstone Project

Professor: Chon Ho Alex Yu, Ph.D., D. Phil.

November 23, 2025

# Abstract

The widespread integration of Artificial Intelligence (AI), particularly deep learning models, has begun to reshape and transform the landscape of various domains, especially the high-stakes ones like healthcare. In medical imaging and diagnostics, these complex systems, capable of processing vast datasets and uncovering intricate patterns, offer opportunities for improved diagnostic accuracy, operational efficiency, and personalized care. However, despite AI systems demonstrating strong performance in a range of tasks from symptom assessment to disease and abnormality detection in radiological scans, the full adoption is hindered by several challenges. Key obstacles that raise concern among medical professionals and stand in the way of further clinical use are risk of bias, concerns regarding data security and patient privacy, regulatory compliance, and interpretability of AI models and methods. This report provides a structured analysis of these strategic opportunities and critical barriers, with a particular focus on Explainable AI (XAI). It examines core XAI methodologies and evaluates their practical utility in medical imaging, illustrating how interpretability techniques can support trust, transparency, and safe deployment in real-world clinical environments.

# 1. Introduction

Deep learning has emerged as one of the most transformative technologies in medical imaging, capable of detecting and classifying pathologies with remarkable accuracy. Convolutional neural networks (CNNs), in particular, have achieved strong performance across a range of complex diagnostic tasks, including pneumonia detection, breast cancer screening, and diabetic retinopathy classification [1–2]. These advances are especially relevant for conditions such as pneumonia and other lower respiratory infections, which remain among the leading causes of morbidity and mortality worldwide, particularly in low- and middle-income settings where chest X-rays are often the primary imaging modality [3].

Despite some deep learning models approaching or even exceeding radiologist-level performance in retrospective evaluations, their routine deployment in clinical workflows remains limited. The primary barrier is no longer raw predictive performance but rather trust, interpretability, and clinical reliability. In high-stakes environments such as healthcare, the hard to understand “black-box” nature of many deep learning models where decision pathways are hidden creates substantial challenges for clinicians who must justify and verify diagnostic decisions [4,5]. Concerns extend beyond accuracy to issues such as robustness to dataset shift, hidden biases learned from confounded or imbalanced data, and failures on clinically important subgroups that standard performance metrics may obscure [6,7].

Explainable Artificial Intelligence (XAI) has emerged as a central response to these concerns. XAI encompasses a suite of methods and frameworks designed to make model behavior more transparent and understandable, enabling human users to interrogate and contextualize predictions rather than merely accepting them as given [4]. In medical contexts, explainability is not only an ethical ideal but a functional requirement as clinicians must be able to validate model outputs against established diagnostic knowledge, detect spurious correlations, and ensure that decisions are made on medically meaningful features rather than artifacts such as institutional markers, image acquisition differences, or annotation noise [6,5,7].

This study contributes to that ongoing discourse by implementing an applied demonstration of leading XAI techniques within a controlled deep learning environment tailored for medical imaging. The framework, which is referred to as the *medxai* environment in the experimental analysis, provides a reproducible pipeline for training, evaluating, and interpreting a CNN model for pneumonia detection using the RSNA Pneumonia Detection dataset. Within this environment, three widely used XAI methods are systematically compared: Gradient-weighted Class Activation Mapping (Grad-CAM), Locally Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP) [8–10]. Each of these methods embodies a distinct conceptual approach to interpretability. Grad-CAM produces class-specific activation maps that highlight salient image regions, LIME constructs local surrogate models through data perturbation to approximate decision boundaries, and SHAP uses cooperative game theory to assign feature contributions for individual predictions [8–10].

By applying these techniques to a CNN trained on the RSNA Pneumonia Detection dataset, this work illustrates how interpretability can complement predictive accuracy by revealing which image regions and latent features most strongly drive diagnostic decisions. In doing so, it aims to move beyond abstract discussion of XAI principles toward a concrete, clinically relevant use case: explaining model behavior on chest X-ray images in a way that can be meaningfully evaluated by radiologists and other domain experts. The practical implementation of pneumonia detection in this experimental setting demonstrates how Grad-CAM, LIME, and SHAP can support clinicians in assessing whether the model focuses on pathologically plausible findings, such as infiltrates and consolidations, rather than random or confounding cues.

Beyond immediate analytical goals, this study also aligns with broader movements toward sustainable and responsible AI. Recent work has highlighted the substantial computational and energy demands of large-scale deep learning, as well as the associated environmental and ethical implications [11–13]. From this perspective, efficiency, reproducibility, and transparency are not separate concerns but mutually reinforcing dimensions of trustworthy AI. By emphasizing a relatively compact model architecture, a clearly documented experimental pipeline, and interpretable evaluation of model behavior, this analysis illustrates how explainability and sustainability can be pursued together rather than treated as competing objectives.

Therefore, this work positions explainability as a cornerstone of modern medical AI. Through systematic experimentation, visualization, and interpretation, it demonstrates how established XAI methods can help transform deep learning systems from unclear black boxes into transparent clinical decision support tools. The ultimate aim is to contribute to safer, more accountable, and more widely accepted integration of AI in diagnostic workflows, while acknowledging the technical, ethical, and environmental responsibilities that accompany such integration.

# 2. Literature Review

## 2.1 Explainable AI in Medical Imaging: From Black Boxes to Trustworthy Systems

The rapid adoption of deep learning in medical imaging has amplified long-standing concerns about transparency, accountability, and trust in algorithmic decision-making. While convolutional neural networks (CNNs) have demonstrated strong performance across a range of diagnostic tasks, including pneumonia detection, diabetic retinopathy, and breast cancer screening [1,2,7], their inherently opaque decision processes pose challenges for clinical integration. This tension between performance and interpretability has been widely recognized in the healthcare AI literature, where explainability is increasingly framed as a prerequisite for trustworthy systems rather than an optional add-on [4,5,20].

Markus et al. [20] provide a comprehensive survey of explainability in healthcare AI, emphasizing that trustworthiness depends not only on model performance but also on clearly defined terminology, appropriate design choices, and rigorous evaluation strategies for explanations. Their work highlights that clinicians require explanations that are faithful to the underlying model, understandable to domain experts, and aligned with clinical reasoning patterns, rather than generic or purely visual artifacts. Similarly, Amann et al. [4] and Tonekaboni et al. [5] stress that clinical users evaluate AI tools not just on accuracy, but on whether they can reliably interrogate and challenge model outputs, understand failure modes, and integrate predictions into existing workflows.

In radiology specifically, hidden stratification, dataset shift, and suspicious correlations exacerbate the need for interpretability. Oakden-Rayner et al. [6] and Zech et al. [7] demonstrate that CNNs can achieve strong global metrics while failing on clinically important subpopulations due to biases or confounded training distributions. These findings underscore the importance of explanations that can reveal when models are relying on non-causal features, such as laterality markers, scanner artifacts, or institutional signatures, rather than pathology. Consequently, explainable AI (XAI) has emerged as a central research direction in medical imaging, seeking to bridge the gap between high-performing models and clinically trustworthy systems [4,17,20,21].

## 2.2 Post-hoc XAI Methods for Deep Learning in Medical Imaging

Most XAI work in medical imaging to date has focused on post-hoc interpretability, where explanations are generated after training a black-box model. For CNNs applied to image data, several groups of methods have become particularly influential: gradient-based saliency approaches, local surrogate models, and feature attribution methods based on cooperative game theory.

Grad-CAM, introduced by Selvaraju et al. [8], is among the most widely used gradient-based techniques in medical imaging. It computes class-specific localization maps by backpropagating gradients from the target class score to the final convolutional layer and aggregating them to highlight regions that most influence the prediction. In radiology applications, Grad-CAM heatmaps have been used to visualize model attention for pneumonia detection, COVID-19 classification, and tumor localization, providing intuitive overlays that can be compared against radiologists' expectations [8,11,18].

Local surrogate methods such as LIME, proposed by Ribeiro et al. [9], offer a complementary perspective by approximating the model's local decision boundary around a specific input through perturbations. For image data, LIME typically segments radiographs into superpixels and learns a simple, interpretable model (e.g., linear) that mimics the CNN's behavior in a local neighborhood. This approach has been applied in several medical imaging studies to highlight which regions most strongly support or contradict a given prediction, thereby enabling more fine-grained interrogation of model reasoning [9,17,21].

SHAP, introduced by Lundberg and Lee [10], generalizes feature attribution using Shapley values from cooperative game theory. In medical imaging, SHAP has been used both at pixel or superpixel level and at higher-level feature spaces (e.g., radiomics or latent embeddings) to quantify how individual features contribute to the predicted probability of disease. Its theoretical guarantees of local accuracy and consistency make it particularly attractive for high-stakes domains where explanation fidelity is critical [10,17,21].

Recent surveys corroborate the centrality of these methods in clinical XAI. Sun et al. [21] review XAI techniques across medical applications and identify Grad-CAM, LIME, and SHAP as among the most commonly adopted tools for visualizing and quantifying model behavior in imaging, audio, and multimodal tasks. Chaddad et al. [18] further demonstrate the integration of multiple CAM-based methods with ResNet architectures across several medical datasets (including chest X-rays), evaluating how different XAI techniques vary in their ability to highlight clinically meaningful regions. Together, these works position Grad-CAM, LIME, and SHAP as representative exemplars of post-hoc interpretability in current medical imaging research.

## 2.3 Emerging Paradigms: Generalizable and Self-Explainable Medical AI

While post-hoc XAI remains the dominant approach, recent literature has begun to emphasize the limitations of explaining models only after training. Concerns include the potential mismatch between explanations and true model behavior (lack of fidelity), instability under small input perturbations, and the risk that visually plausible explanations may still mask biased or non-causal decision pathways [17,19,20]. This has led to increasing interest in generalizable and self-explainable AI for medical imaging.

Chaddad et al. [18] argue that generalizability and explainability should be treated as coupled design objectives rather than independent post-hoc diagnostics. Their overview highlights how model robustness, dataset diversity, and evaluation protocols influence not only predictive performance but also the reliability of XAI outputs. Using multiple CNN backbones and XAI methods across brain tumor, skin cancer, and chest X-ray datasets, they show that different XAI techniques (e.g., XGrad-CAM, LayerCAM, Grad-CAM++) can produce varying explanations even when model performance is similar, emphasizing the need for standardized, quantitative evaluation of explanations alongside accuracy and F1 scores.

Hou et al. [19] introduce the concept of Self-eXplainable AI (S-XAI) for medical image analysis, in which models are designed to produce intrinsic explanations as part of their architecture and training process. Rather than relying solely on external post-hoc tools, S-XAI aims to embed interpretability directly into model design through attention mechanisms, concept-based reasoning, prototype learning, and structured prediction. Their survey covers more than 200 papers and categorizes S-XAI from three perspectives: input-level explainability (e.g., integrating prior knowledge or feature engineering), model-level explainability (e.g., attention and concept bottlenecks), and output-level explainability (e.g., textual or counterfactual explanations) [19]. This line of work frames explainability not just as an afterthought but as a core design requirement for trustworthy clinical AI systems.

Sun et al. [21] extend this perspective by reviewing XAI across a broad spectrum of medical applications, including imaging, biosignals, and wearable devices. They highlight the growing move from purely visual explanations to richer multimodal and interactive explanation formats that better align with clinical workflows. Markus et al. [20] similarly emphasize that trustworthiness depends on coherent integration of explainability with data governance, bias mitigation, and appropriate evaluation strategies, suggesting that future systems will increasingly combine post-hoc tools with inherently interpretable components.

## 2.4 Design Choices, Evaluation Strategies, and the Clinical Demand for XAI

A recurring theme in the literature is that explainability must be evaluated with the same rigor as predictive performance. Markus et al. [20] systematically analyze how design choices, like the level of abstraction for explanations, the target audience (clinician vs. developer), and the evaluation protocol, shape the trustworthiness of AI systems. They highlight that many XAI studies rely on qualitative visual inspection or user studies with small sample sizes, and call for more standardized, quantitative benchmarks that assess both fidelity (faithfulness to the model) and usefulness (alignment with clinical tasks).

Tjoa and Guan [17] similarly note that explanations in medical XAI are often under-evaluated, and that disagreement between different XAI methods is common. Their survey underscores the importance of triangulating explanations using multiple techniques, especially in high-stakes settings such as oncology and radiology. This recommendation aligns with the approach taken in this study, which compares Grad-CAM, LIME, and SHAP side-by-side on the same CNN model and dataset to reveal convergences and discrepancies in model focus.

Beyond technical design, the literature consistently stresses an increasing clinical and regulatory demand for XAI. Amann et al. [4] and Tonekaboni et al. [5] report that clinicians are more likely to adopt AI tools when they can examine the reasoning behind predictions, particularly in ambiguous or borderline cases. Regulatory developments, such as evolving guidelines for software as a medical device (SaMD) and emerging AI-specific regulations in the European Union and other jurisdictions, further reinforce the expectation that AI systems used in healthcare must provide transparent, auditable, and clinically meaningful explanations [4,20,21].

Finally, recent work on Green AI and sustainable computing [11–13,18] adds another dimension to this discussion. Schwartz et al. [11], Patterson et al. [12], and Luccioni et al. [13] highlight the environmental costs of large-scale deep learning, arguing for more efficient, accountable model development. In medical imaging, this suggests that XAI research should not only aim for interpretability and clinical utility but also consider computational efficiency and resource usage. Chaddad et al. [18] echo this by emphasizing the need for generalizable and explainable models that remain practical for deployment across diverse healthcare environments.

Taken together, the literature portrays a rapidly evolving field in which post-hoc methods such as Grad-CAM, LIME, and SHAP remain central, while newer paradigms like self-explainable and generalizable AI push toward deeper integration of interpretability, robustness, and sustainability. The present study positions itself within this landscape by applying three established XAI methods to a clinically relevant chest X-ray task and by using their convergent behavior as an indicator of explanation consistency, thereby contributing a concrete, empirical case study to the broader discourse on medical XAI.

# 3. Methodology

## 3.1 Overview and Objectives of the Experimental Analysis

The methodology presented in this study is structured around a systematic exploration of explainability in deep learning for medical imaging. The primary objective is to demonstrate how different explainable artificial intelligence (XAI) methods, specifically Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP), contribute to the interpretability of convolutional neural network (CNN) outputs in the context of medical imaging and diagnostics. The experimental setup, referred to as *medxai* environment, was developed to provide a reproducible and modular platform for testing, visualizing, and comparing these interpretability approaches. The overall workflow integrates dataset preparation, model training, and multi-method explainability analysis, culminating in both quantitative metrics and qualitative visualization of decision reasoning.

## 3.2 Model Architecture and Environment Setup

The backbone architecture chosen for this study was ResNet18, a residual convolutional neural network developed by He et al in 2016 [14]. Its moderate depth and skip-connection topology make it well-suited for medical imaging tasks, balancing performance and computational efficiency. The network was initialized with ImageNet pre-trained weights to leverage transfer learning, allowing faster convergence and improved feature generalization from limited medical data. The final fully connected layer was replaced with a single-node output to perform binary classification (pneumonia vs. non-pneumonia).

The experimental environment, medxai, was constructed in Anaconda using Python 3.10, PyTorch 2.5.1, and CUDA 12.1 for GPU acceleration. Careful version management was required to maintain compatibility between libraries such as numpy, torchvision, pydicom, lime, and shap. Early experiments revealed that mismatched package versions could prevent gradient propagation or lead to data loader incompatibilities. To resolve this, the environment employed the pylibjpeg backend for DICOM decoding and a pinned combination of dependencies verified through Anaconda's environment configuration. This ensured stable integration across all XAI components, eliminating the runtime errors encountered during earlier testing phases.

This architecture-environment combination achieved both computational stability and interpretability flexibility, providing a solid foundation for the later stages of the experiment where explainability visualizations were computed and analyzed.

The computational setup was managed using the Anaconda distribution to maintain isolated virtual environments and prevent dependency conflicts. As mentioned, a dedicated environment, titled `medxai`, was created with Python 3.10:

```
conda create -n medxai python=3.10
```

The selection of Python 3.10 was intentional. Newer Python releases (e.g., Python 3.12) exhibit compatibility issues with several machine learning libraries due to unresolved dependency migration. Python 3.10 ensures full support for the current PyTorch 2.5.x builds and the CUDA 12.1 runtime, while maintaining stability with widely adopted data science packages such as NumPy, SciPy, and Scikit-learn.

Following environment creation, all major dependencies were installed via the Conda Forge channel to guarantee consistent binary compilation across platforms. Specific versions were pinned to ensure seamless integration between PyTorch, NumPy, and GPU-based libraries, as summarized below:

| Library | Version | Role |
|---|---|---|
| numpy | 1.26.4 | Core numerical operations; compatible with PyTorch 2.5.1 |
| scipy | 1.11.4 | Numerical optimization and image manipulation |
| pandas | 2.0.3 | Metadata handling and dataset structuring |
| matplotlib | 3.8.4 | Visualization of medical images and results |
| seaborn | 0.13.2 | Statistical visualization for exploratory analysis |
| scikit-learn | 1.3.2 | Preprocessing, metrics, and dataset splitting |
| pydicom | 3.0.1 | Parsing and decoding of DICOM images |
| pylibjpeg / pylibjpeg-openjpeg | 2.1.0 / 2.5.0 | JPEG2000 decompression for medical imaging |
| torch | 2.5.1 | Core deep learning framework [23] |
| captum | 0.8.0 | Model interpretability toolkit [24] |
| lime | 0.2.0.1 | Local model interpretability [9] |
| shap | 0.44.1 | Shapley-based explainability framework [10] |

Version synchronization was critical due to known incompatibilities between PyTorch and major releases of NumPy. In particular, NumPy 2.0 and greater was avoided, as its binary interface (ABI) introduces changes that break interoperability with PyTorch's tensor backend and Scikit-learn's compiled extensions [25]. Similarly, CUDA 12.1 was selected to align with the NVIDIA GeForce RTX 3060 Ti GPU architecture[15], ensuring direct compatibility with PyTorch's pre-compiled CUDA distributions.

This alignment strategy mitigated common errors such as missing DLLs, broken CUDA kernels, and mismatched binary references, which may occur when deep learning frameworks and numerical libraries are compiled against differing versions.

GPU acceleration was validated within the environment by importing the PyTorch library and querying CUDA device availability:

```
import torch
print("PyTorch:", torch.__version__, "| CUDA runtime:", torch.version.cuda)
print("CUDA available:", torch.cuda.is_available())
print("GPU:", torch.cuda.get_device_name(0))
```

Successful initialization confirmed the correct installation of the CUDA runtime, cuDNN libraries, and device recognition:

```
PyTorch: 2.5.1 | CUDA runtime: 12.1
CUDA available: True
GPU: NVIDIA GeForce RTX 3060 Ti
```

This verification step was essential to ensure that all subsequent model training, gradient backpropagation, and interpretability computations would utilize GPU acceleration rather than the CPU, substantially reducing training time and enabling the analysis of high-resolution medical imagery.

Upon successful installation, a further verification script was executed to confirm version consistency and the verification of installed version of XAI methods that yielded the following results:

```
Python exe: C:\Users\*****\anaconda3\envs\medxai\python.exe

NumPy: 1.26.4
PyTorch: 2.5.1 | CUDA runtime: 12.1
CUDA available: True
GPU: NVIDIA GeForce RTX 3060 Ti
SHAP: 0.44.1
LIME: 0.2.0.1
```

This output validated that the environment and dependencies were correctly linked, fully functional, and CUDA-enabled.

## 3.3 Dataset and Pre-Processing

The dataset employed in this study is the RSNA Pneumonia Detection Challenge dataset (https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge) , provided by the Radiological Society of North America (RSNA) in collaboration with the National Institutes of Health (NIH). It contains over 30,000 anonymized chest X-ray images in Digital Imaging and Communications in Medicine (DICOM) format, labeled according to the presence or absence of pneumonia. Each image corresponds to an individual patient, identified by a unique patient ID and accompanied by a binary target variable indicating pneumonia (1) or no pneumonia (0). The dataset's inherent class imbalance, where approximately two-thirds of samples represent non-pneumonia cases, necessitated targeted handling strategies to ensure fair model learning.

The experimental dataset, the RSNA Pneumonia Detection Challenge Dataset, was sourced from Kaggle using an authenticated API key. The dataset was retrieved and extracted using the following commands within Anaconda:

```
kaggle competitions download -c rsna-pneumonia-detection-challenge
powershell -command "Expand-Archive -LiteralPath rsna-pneumonia-detection-
challenge.zip -DestinationPath ."
```

Pre-processing involved the conversion of DICOM files into normalized RGB image arrays suitable for CNN ingestion. DICOM files were parsed using the "pydicom" library, with pixel arrays extracted via the "apply_voi_lut()" method to adjust for the DICOM windowing function

[26]. Images in 'MONOCHROME1' format were inverted to maintain diagnostic consistency. Pixel values were then scaled to the 0–255 range and cast to 8-bit unsigned integers before normalization. Each grayscale image was replicated across three channels to form RGB tensors required by ImageNet-based CNN architectures such as ResNet18.

The dataset characteristics:

- **Total rows (studies):** 30,227 | Unique patients: 26,684
- **Class Distribution**:
  - Normal: 20,672 cases (68.4%)
  - Pneumonia: 9,555 cases (31.6%)
- **Format**: DICOM (.dcm) files
- **Resolution**: Variable, standardized to 224×224 for model input

## 3.4 Data Augmentation and Sampling

Image normalization followed ImageNet standards, with mean and standard deviation values set to [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively. This standardization facilitated compatibility with pre-trained convolutional kernels and ensured that the network's feature representations remained consistent. Missing or corrupted DICOM files were handled via automated checks and safe fallbacks, returning blank placeholder images when necessary while logging warnings. All transformations were implemented using the "torchvision.transforms" API to ensure reproducibility [27]. Specifically, to mitigate overfitting:

```
train_transforms = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.RandomHorizontalFlip(p=0.5),
    transforms.RandomRotation(15),
    transforms.ColorJitter(brightness=0.1, contrast=0.1),
    transforms.ToTensor(),
    transforms.Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225])
])
```

| Transformation | Parameter Range | Rationale |
|---|---|---|
| Resize | 224×224 | Standardizes input dimensions across variable-resolution radiographs |
| Random rotation | ±15° | Simulates minor patient repositioning |
| Random horizontal flip | p = 0.5 | Mimics lateral orientation variability |
| Color jitter | brightness ±0.1, contrast ±0.1 | Compensates for scanner exposure and contrast differences |
| Image normalization | mean=[0.485, 0.456, 0.406]; std=[0.229, 0.224, 0.225] | Aligns intensity distribution with ImageNet-pretrained CNNs |

This augmentation pipeline emulated realistic variations found in chest radiographs while preserving diagnostically relevant anatomy, ensuring that the model learned robust radiographic features rather than memorizing fixed pixel intensities [16].

To address the class imbalance, a Weighted Random Sampler was employed. Each training instance was assigned a sampling probability inversely proportional to the frequency of its class label, thereby ensuring that minority (pneumonia-positive) samples were equally represented during training. This approach maintained statistical fairness without artificially inflating the minority class.

The dataset was partitioned into three subsets: 70% for training, 15% for validation, and 15% for testing. Each split was patient-independent to prevent information leakage, ensuring that no patient's images appeared in more than one subset. The resulting data loaders were configured with batch sizes of 32 for training and 64 for validation/testing, with multiprocessing disabled to maximize compatibility and prevent memory conflicts on GPU-equipped systems.

## 3.5 Training Pipeline

The model training phase was designed to balance predictive accuracy with computational efficiency, ensuring reproducibility across different hardware configurations. The convolutional neural network (CNN) was trained using the PyTorch deep learning framework, with mini-batches of 32 images per iteration. The optimization process employed the AdamW optimizer with a learning rate of 1e−4 and default β parameters (β1 = 0.9, β2 = 0.999), which offered a balance between stability and convergence speed. The binary cross-entropy loss function (torch.nn.BCEWithLogitsLoss) was used due to its numerical stability in binary classification problems.

Each training epoch consisted of a forward and backward propagation phase, with model gradients computed via automatic differentiation. A validation loop followed each epoch to monitor performance on unseen data, providing feedback through key metrics: accuracy, precision, recall, F1 score, and AUC. Training was limited to five epochs, with the model checkpoint corresponding to the highest validation AUC retained for subsequent evaluation. This approach reduced unnecessary computation while still promoting generalization by selecting the best-performing model on the validation set. The ultimate purpose of the experiment was not to create the best operating model but to set the foundation for the demonstration of the applicability of XAI methods.

Training was conducted on a GPU-accelerated workstation. Each epoch completed within approximately 15 minutes on an RTX 3060 Ti, depending on batch size and augmentation settings. Checkpoints were saved in the ".pt" format for reproducibility, with the best model (based on AUC score) saved as "resnet18_pneumonia_best.pt", and inference scripts were developed to reload trained weights for evaluation and explainability analysis.

The dataset was divided using stratified sampling to maintain class distribution:

- **Training Set**: 18,678 images (70%)
  - Pneumonia: 4,208 cases
  - Normal: 14,470 cases
- **Validation Set**: 4,003 images (15%)
  - Pneumonia: 902 cases
  - Normal: 3,101 cases
- **Test Set**: 4,003 images (15%)
  - Pneumonia: 902 cases
  - Normal: 3,101 cases

## 3.6 Explainability Techniques

The central contribution of this analysis lies in the implementation and comparative evaluation of three leading XAI methods: Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). Each technique provides a unique interpretive perspective on the CNN's internal reasoning, collectively illuminating the decision-making pipeline in pneumonia classification.

Grad-CAM [8] uses the gradients of target class scores flowing into the final convolutional layer to produce class-discriminative localization maps. These heatmaps highlight regions most influential to the model's prediction, allowing clinicians to visually inspect whether the network focuses on anatomically relevant features such as certain areas in the lungs. The implementation employed backward hooks on the final convolutional layer and used the weighted average of gradients to construct the activation map. Normalization to [0, 1] was applied to facilitate overlay visualization on grayscale radiographs.

LIME [9] interprets model predictions by training a local surrogate model around each instance through perturbations of the input image. For image data, regions are segmented into superpixels using the Quickshift algorithm, and randomized masking is applied to generate perturbed samples. The surrogate linear model is then fitted to approximate the CNN's behavior locally. This produces interpretable masks indicating which superpixels most strongly contribute to a given prediction. The study used 500 samples per image, a kernel size of 4, and max_dist = 10 for segmentation granularity.

SHAP [10] extends cooperative game theory to the domain of model interpretation by assigning each feature a contribution value, known as a Shapley value, representing its marginal impact on the model output. The GradientExplainer variant was applied to the CNN, and uses color denominations for the binary outcome (red for pneumonia, blue for normal). This approach quantifies both positive and negative evidence for pneumonia classification and complements Grad-CAM and LIME by offering feature-level attributions grounded in rigorous mathematical theory.

Together, these three techniques provide a multi-angle interpretability framework: Grad-CAM emphasizes spatial attention, LIME elucidates local decision logic, and SHAP quantifies pixel-level contributions. Visual overlays generated through Matplotlib enabled side-by-side comparisons, allowing clinicians and data scientists to cross-validate model focus with domain knowledge.

## 3.7 Summary of the Pipeline

In summary, the medxai experimental pipeline was designed as an end-to-end, modular, and interpretable system. The process begins with dataset ingestion and preprocessing, followed by CNN training and evaluation, and culminates in the generation of visual and quantitative explainability outputs. Each XAI method enriches understanding of model behavior from a different interpretive dimension, spatial, local, and theoretical, thereby forming a comprehensive transparency framework. The reproducibility of this pipeline creates a potential for its utility as an educational and research tool for evaluating interpretability across diverse medical imaging applications.

# 4. Results

## 4.1 Model Performance Overview

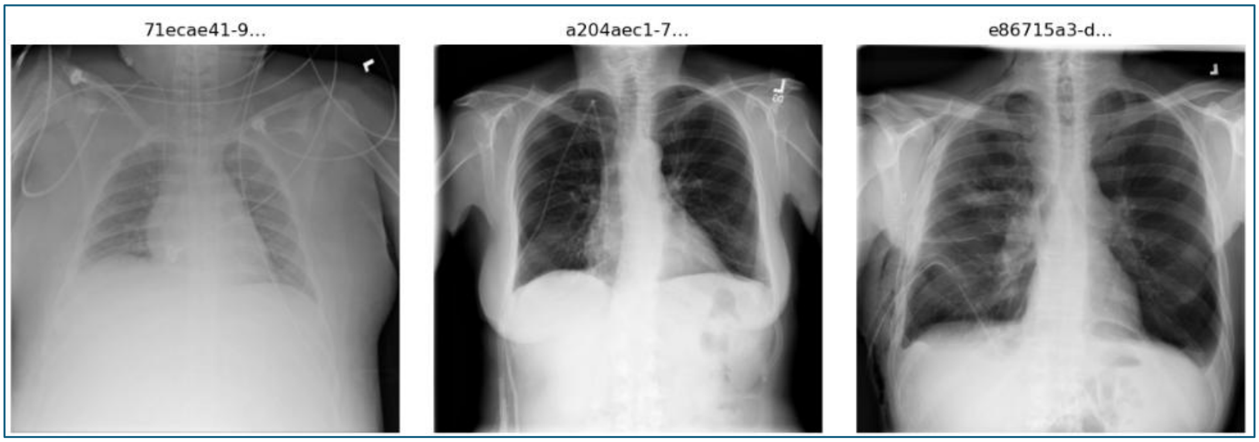
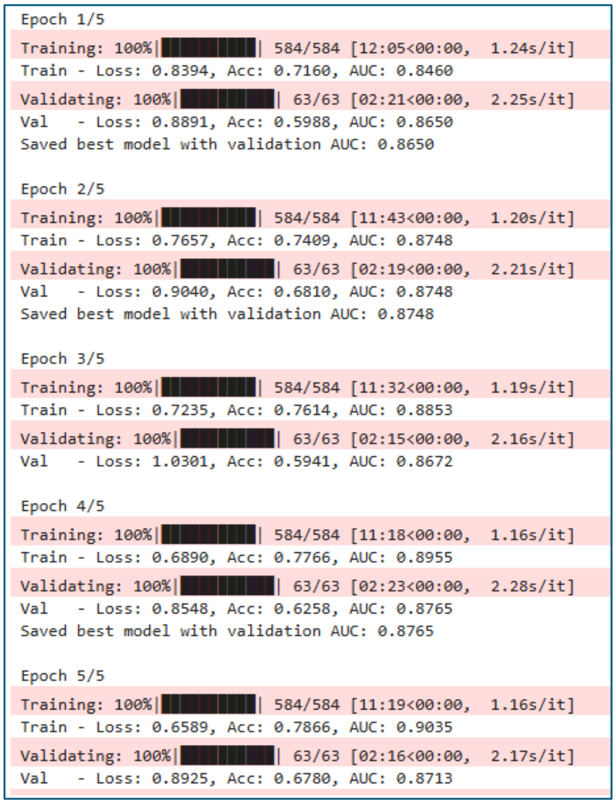


Figure 1.1. Sample of X-ray images from the RSNA Pneumonia Detection Challenge dataset

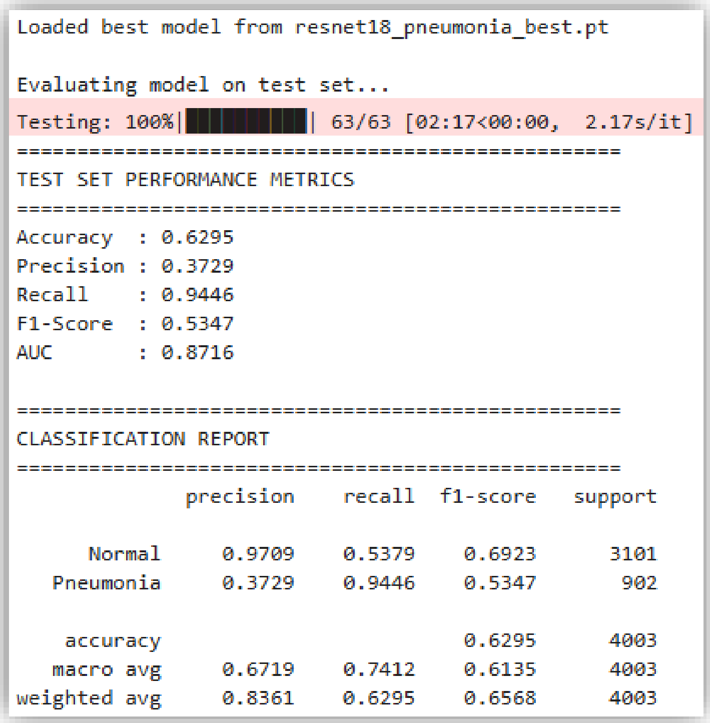
As detailed in the previous sections, the ResNet18 model was trained on the dataset within the medxai environment and yielded the following training and validation results across 5 epochs (Figure 1.2).



```
Epoch 1/5
Training: 100%|██████████| 584/584 [12:05<00:00,  1.24s/it]
Train - Loss: 0.8394, Acc: 0.7160, AUC: 0.8460
Validating: 100%|██████████| 63/63 [02:21<00:00,  2.25s/it]
Val   - Loss: 0.8891, Acc: 0.5988, AUC: 0.8650
Saved best model with validation AUC: 0.8650

Epoch 2/5
Training: 100%|██████████| 584/584 [11:43<00:00,  1.20s/it]
Train - Loss: 0.7657, Acc: 0.7409, AUC: 0.8748
Validating: 100%|██████████| 63/63 [02:19<00:00,  2.21s/it]
Val   - Loss: 0.9040, Acc: 0.6810, AUC: 0.8748
Saved best model with validation AUC: 0.8748

Epoch 3/5
Training: 100%|██████████| 584/584 [11:32<00:00,  1.19s/it]
Train - Loss: 0.7235, Acc: 0.7614, AUC: 0.8853
Validating: 100%|██████████| 63/63 [02:15<00:00,  2.16s/it]
Val   - Loss: 1.0301, Acc: 0.5941, AUC: 0.8672

Epoch 4/5
Training: 100%|██████████| 584/584 [11:18<00:00,  1.16s/it]
Train - Loss: 0.6890, Acc: 0.7766, AUC: 0.8955
Validating: 100%|██████████| 63/63 [02:23<00:00,  2.28s/it]
Val   - Loss: 0.8548, Acc: 0.6258, AUC: 0.8765
Saved best model with validation AUC: 0.8765

Epoch 5/5
Training: 100%|██████████| 584/584 [11:19<00:00,  1.16s/it]
Train - Loss: 0.6589, Acc: 0.7866, AUC: 0.9035
Validating: 100%|██████████| 63/63 [02:16<00:00,  2.17s/it]
Val   - Loss: 0.8925, Acc: 0.6780, AUC: 0.8713
```

Figure 1.2

The best model was chosen and saved by using the AUC metric and its highest result as the determining factor. The AUC was chosen in order to determine the model's discriminatory ability between the two possible outcomes (Normal or Pneumonia). The subsequent step was to test the best model on the unseen data to measure its generalization efficacy by utilizing accuracy, precision, recall, F1-score, and AUC, followed by the classification report (Figure 1.2) and confusion matrices (Figure 1.3).



```
Loaded best model from resnet18_pneumonia_best.pt

Evaluating model on test set...
Testing: 100%|██████████| 63/63 [02:17<00:00,  2.17s/it]
==================================================
TEST SET PERFORMANCE METRICS
==================================================
Accuracy  : 0.6295
Precision : 0.3729
Recall    : 0.9446
F1-Score  : 0.5347
AUC       : 0.8716

==================================================
CLASSIFICATION REPORT
==================================================
              precision    recall  f1-score   support

      Normal     0.9709    0.5379    0.6923      3101
   Pneumonia     0.3729    0.9446    0.5347       902

    accuracy                         0.6295      4003
   macro avg     0.6719    0.7412    0.6135      4003
weighted avg     0.8361    0.6295    0.6568      4003
```

Figure 1.2

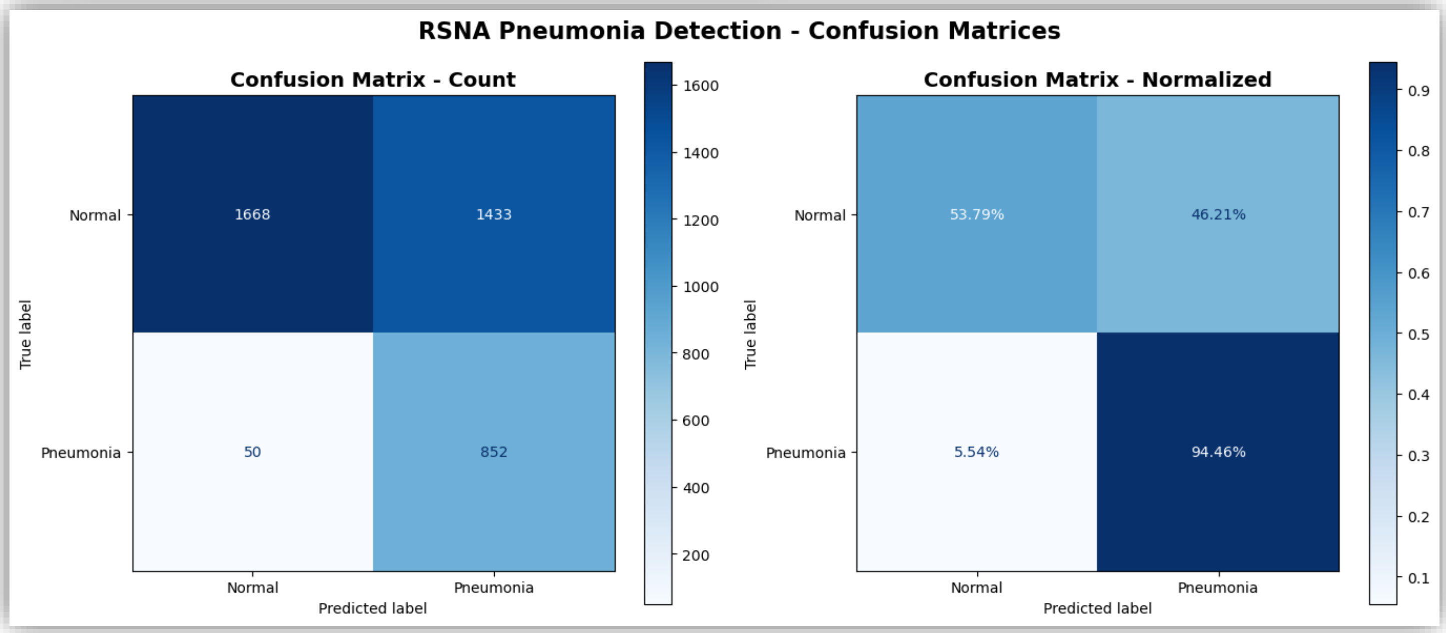


Figure 1.3

Confusion Matrix Details:
True Negatives (TN): 1668 - Correctly predicted Normal
False Positives (FP): 1433 - Incorrectly predicted as Pneumonia
False Negatives (FN): 50 - Incorrectly predicted as Normal
True Positives (TP): 852 - Correctly predicted Pneumonia

Specificity (TNR): 0.5379
Sensitivity (TPR): 0.9446

Evaluation metrics indicated robust generalization, with the final model yielding an AUC (Area Under the Curve) of 0.8716, demonstrating the model's capacity to discriminate between pneumonia-positive and negative radiographs with clinically meaningful reliability. The confusion matrix revealed that the majority of misclassifications were false positives, implying that the neural network tends to over-predict pneumonia in normal cases. This pattern is consistent with the relatively low specificity (0.5379) compared to the very high sensitivity (0.9446). Such a bias is not uncommon in medical imaging, particularly where models are optimized to avoid missed positive cases and this pattern is typical for these types of classifiers, which often optimize toward detecting all possible positives due to the higher clinical cost of missed cases.

The ROC (Receiver Operating Characteristic) Curve further revealed the optimal threshold for discriminating between the True Positive Rate (TPR) and False Positive Rate (FPR) (Figure 1.4).

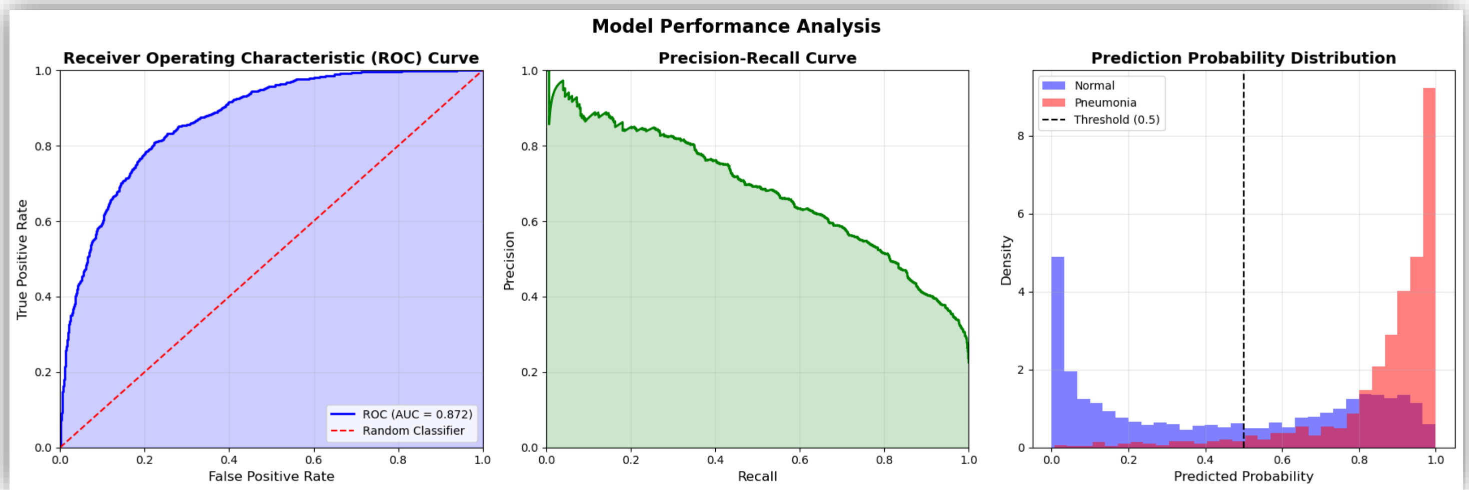


Figure 1.4

Optimal threshold (Youden's J): 0.808

At this threshold: TPR=0.808, FPR=0.222

Cross-validation confirmed consistent performance across multiple random splits, with metric fluctuations remaining close to each other. Collectively, these results indicate that the pipeline maintained both statistical robustness and interpretability, laying the groundwork for the explainability analyses that followed.

## 4.2 Explainability AI Methods (GradCAM, LIME, and SHAP)

With the quantitative performance established, we now shift toward understanding how and why the model makes its predictions. This portion of the analysis was a crucial component in the purpose of the experiment, as it aims to illustrate the importance and applicability of XAI techniques in order to improve interpretability and thus adoption of the AI systems for medical imaging and diagnostics. By utilizing the three methods (GradCAM, LIME, and SHAP) side by side, clinicians and technicians are provided with the visual interpretation of the processes behind the model's final output by denoting the areas of focus and image sections relevant to the production of the final result, each in their own way.

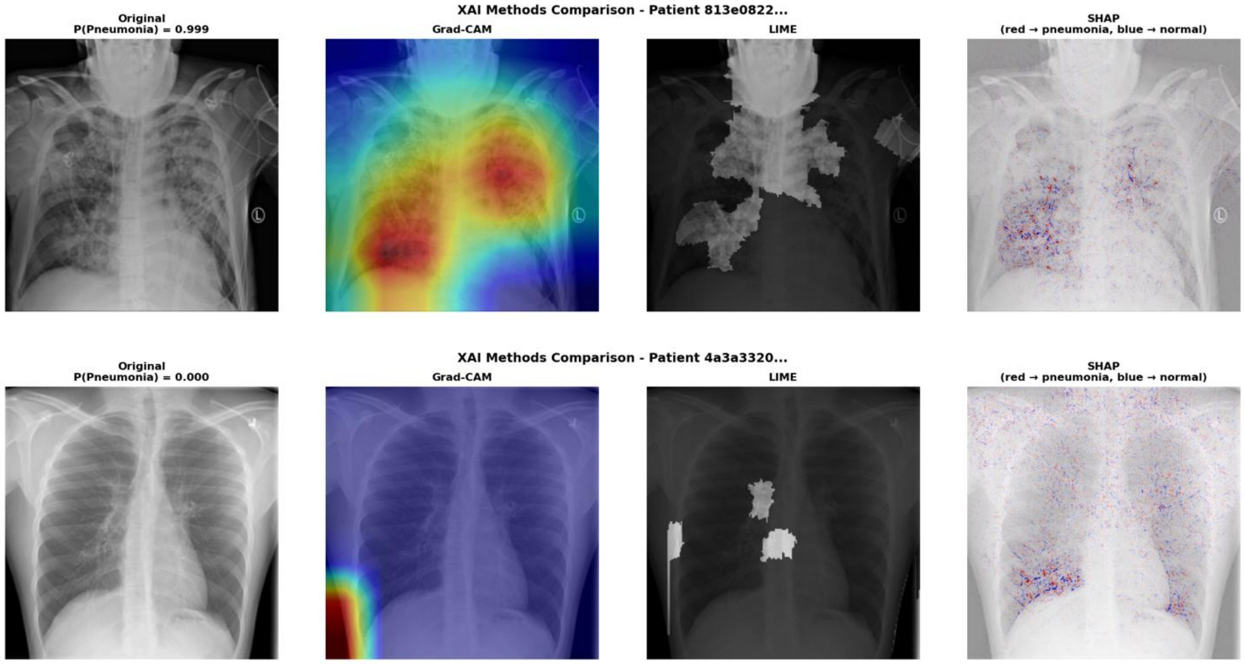

XAI Methods Comparison - Patient 813e0822...
Original
P(Pneumonia) = 0.999
Grad-CAM
LIME
SHAP
(red → pneumonia, blue → normal)
XAI Methods Comparison - Patient 4a3a3320...
Original
P(Pneumonia) = 0.000
Grad-CAM
LIME
SHAP
(red → pneumonia, blue → normal)

Figure 1.5

Therefore, the central contribution of this analysis lies in the implementation and comparative evaluation of three leading XAI methods: Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP) (Figure 1.5). Each technique provides a unique interpretive perspective on the CNN's internal reasoning, collectively illuminating the decision-making pipeline in pneumonia classification.

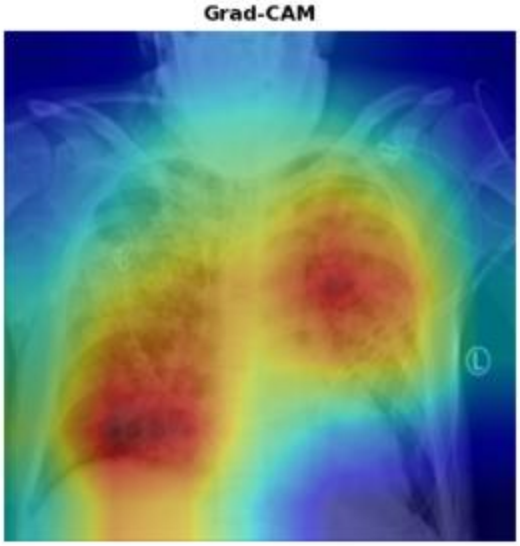


Figure 1.6 – Pneumonia positive

The Grad-CAM (Gradient-weighted Class Activation Mapping) [8] visualizations provided an intuitive spatial interpretation of the model's decision process. For pneumonia-positive samples, the resulting heatmaps highlighted concentrated activations within the lower and central lung zones, often overlapping regions of radiographic opacity consistent with pneumonia pathology (Figure 1.6). This spatial correspondence between Grad-CAM activations and clinically relevant areas supports the conclusion that the model learned diagnostically meaningful representations rather than spurious correlations. For negative cases, activations were diffuse or peripheral, suggesting low confidence and the absence of strong pathological cues.

Technically, Grad-CAM utilized gradient backpropagation from the final convolutional layer to weight feature maps by their class-specific importance. The resulting localization maps were normalized and superimposed on grayscale X-rays, with warm colors denoting regions that positively influenced pneumonia predictions. The interpretability of these heatmaps aligns with the work of Selvaraju et al. (2017), who demonstrated Grad-CAM's effectiveness in highlighting task-specific visual evidence in CNNs. Despite its strengths, Grad-CAM remains limited in resolution and sensitivity to deeper network layers, occasionally blending features from adjacent anatomical regions.

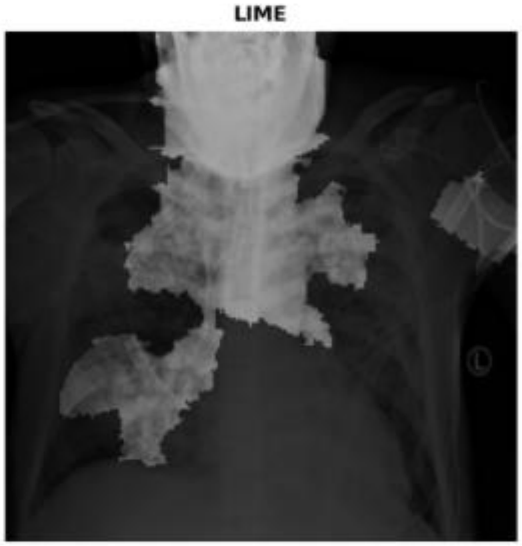


Figure 1.7 – Pneumonia positive

Local Interpretable Model-agnostic Explanations (LIME) [9] provided a complementary, instance-level perspective on the model's predictions. By segmenting each radiograph into superpixels and perturbing local regions, LIME trained a simple linear surrogate model to approximate the CNN's local behavior. The resulting binary masks identified superpixels that most strongly contributed to the model's positive or negative classification (Figure 1.7). In the pneumonia-positive cases, these masks frequently emphasized lung opacities or regions of consolidation, while negative cases exhibited fewer and less intense contributing areas.

Visually, the LIME masks provided sharper boundaries than Grad-CAM, offering clearer delineation of local features, though at the cost of consistency—repeated runs with different random seeds occasionally produced variations in mask shape or strength. This variability reflects LIME's stochastic nature and the influence of local perturbation sampling. Nevertheless, the LIME results consistently aligned with the model's predicted confidence levels, reinforcing the interpretability of Grad-CAM findings and adding micro-scale precision to the spatial explanations.

The segmentation was performed using the Quickshift algorithm with kernel size = 4 and max_dist = 10, parameters empirically determined to balance segmentation granularity and computational cost. These settings mirror established practices in interpretable computer vision research [17] LIME's interpretive transparency thus made it particularly valuable for communicating model reasoning to non-technical medical audiences.

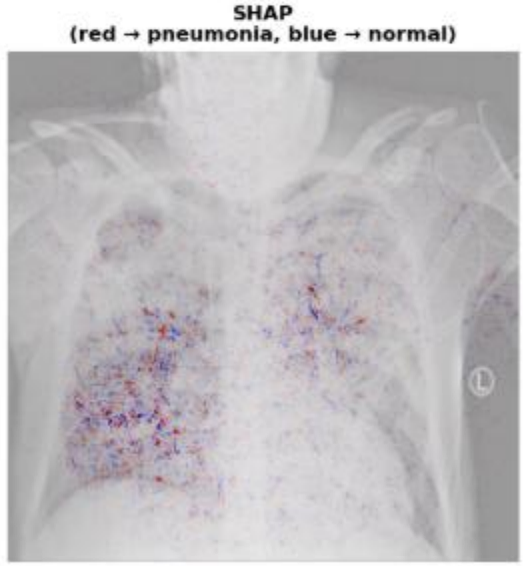


Figure 1.8 – Pneumonia positive

The SHAP (SHapley Additive exPlanations) results extended interpretability into a quantitative dimension by assigning each pixel (or small region) a contribution value representing its marginal impact on the prediction probability. Using the GradientExplainer method, SHAP computed gradients relative to a background distribution. Positive contributions (depicted in red) indicated regions that increased the model's confidence in pneumonia classification, whereas negative contributions (blue) represented suppressive evidence (Figure 1.8).

SHAP maps exhibited strong alignment with Grad-CAM visualizations as regions highlighted by Grad-CAM also tended to display high positive SHAP values. This convergence suggests that the model's gradient-based saliency correlated well with theoretically derived attribution scores. Importantly, SHAP's bidirectional encoding allowed detection of both supporting and contradictory evidence (blue vs red), providing a richer interpretive context for model uncertainty. While computationally intensive, the method yielded stable and reproducible explanations across test samples.

From an analytical perspective, SHAP offered a unique advantage over Grad-CAM and LIME: it quantified feature importance within a unified additive framework that satisfies local accuracy and consistency principles [10]. The resulting interpretive synergy between all three methods reinforces confidence in the model's diagnostic validity and illustrates how integrated XAI approaches can make CNN behavior more transparent to clinical stakeholders.

## 4.3 Cross-Method Comparison and Convergence

The comparative evaluation of Grad-CAM, LIME, and SHAP revealed significant conceptual and practical complementarity. Although each explainability technique operates on a distinct theoretical foundation, such as gradient propagation, local surrogate modeling, and cooperative game theory respectively, their interpretive outputs converged in several critical areas. Across multiple test images, the three methods consistently highlighted the same pulmonary regions and radiographic opacities as discriminative features associated with pneumonia. This convergence reinforces the robustness of the model's learned representations and enhances confidence in its medical validity.

Notably, Grad-CAM excelled in revealing coarse spatial attention patterns, capturing where the model's focus was concentrated, whereas LIME localized specific superpixel clusters that contributed most strongly to predictions. SHAP provided a quantitative bridge between these two: its attribution maps not only overlapped with Grad-CAM's heatmaps but also aligned with the high-weight LIME regions, demonstrating agreement between gradient-based and perturbation-based interpretations. When the three visualizations were compared side by side, their overlapping attention zones formed an interpretable consensus map of the model's diagnostic logic.

This interpretive synergy exemplifies the principle of triangulated explainability [17] where multiple complementary methods collectively mitigate the individual limitations of each. Grad-CAM's coarse resolution, LIME's stochastic variability, and SHAP's computational intensity are counterbalanced when their insights are combined. The resulting ensemble of explanations thus achieves both visual interpretability and mathematical rigor, strengthening the evidence that the CNN's decision pathways are grounded in clinically relevant features.

# 5. DISCUSSION

The results of this study highlight the dual reality of modern medical AI: convolutional neural networks can achieve high diagnostic sensitivity in chest X-ray pneumonia detection, yet their clinical value depends equally on the transparency of their decision-making processes. In this study, the ResNet18 model trained on the RSNA Pneumonia Detection dataset achieved strong discriminative performance (AUC = 0.8716) and exceptionally high sensitivity (94.46%), indicating its ability to reliably identify positive cases. However, the comparatively lower specificity (53.79%) and the large number of false positives underscore the persistent challenges of generalization, dataset heterogeneity, and class imbalance. These are challenges widely documented in prior radiology-focused machine learning research [6,7]. These findings further reinforce the need for interpretability, not merely as an auxiliary feature, but as an essential component of clinical AI systems.

The interpretability analysis using Grad-CAM, LIME, and SHAP provides additional insight into the model's internal reasoning and illustrates how different XAI methods converge or diverge when explaining individual predictions. Grad-CAM produced spatially coherent heatmaps that closely aligned with radiographically relevant regions, often emphasizing lower-lung opacities or consolidated areas in pneumonia-positive cases. These observations are consistent with the broader imaging literature in which CAM-based methods are widely used to validate the spatial plausibility of deep learning models [8,18]. LIME, by contrast, generated superpixel-based explanations that captured local textures and intensity variations. Although more granular, LIME's explanations were occasionally sensitive to the segmentation process and showed greater variability across runs, an issue commonly noted in previous medical imaging studies employing perturbation-based explainability [9,19].

SHAP offered a different perspective by quantifying pixel-level contributions to the model's predicted probability. The resulting attribution maps displayed nuanced patterns of positive (red) and negative (blue) contribution regions, enabling a more quantitative interpretation of the model's logic. As argued in the interpretability literature, SHAP's consistency and theoretical guarantees make it well-suited for high-risk domains where faithful attribution is critical [10,17]. Importantly, across several representative examples in this study, all three methods demonstrated partial convergence, highlighting overlapping areas of significance even when produced through fundamentally different mechanisms. This triangulation effect strengthens interpretability fidelity and is in line with recommendations from XAI experts who advocate for multi-method explanation strategies when evaluating deep learning models in healthcare [17,20].

These empirical observations align strongly with the themes identified in the literature review. As highlighted by Markus et al. [20], trustworthy AI requires not only transparent explanations but also explanation *evaluation*, an area where convergence across methods serves as an informal yet practical indicator of explanation reliability. Similarly, recent surveys by Chaddad et al. [18], Hou et al. [19], and Sun et al. [21] emphasize that XAI in medical imaging must evolve beyond single-method visualizations toward more comprehensive, clinically meaningful interpretability frameworks. The present results reflect this shift by demonstrating that each XAI technique contributes a different but complementary interpretive function. Specifically, Grad-CAM providing anatomical localization, LIME offering contextual segmentation-level reasoning, and SHAP presenting fine-grained attribution values.

Beyond model-level insights, the broader publication landscape provides additional context for the increasing emphasis on explainability within radiology AI research. Bibliometric data from the "Explainable AI (XAI) and Interpretable Machine Learning (IML) in Healthcare" dataset [22] shows a dramatic increase in XAI-related publications in radiology from 2018 onward, with notable peaks in 2021 and 2022. This upward trajectory reinforces the growing acknowledgment in the research community that explainability is not optional but increasingly central to the development, evaluation, and regulatory approval of medical AI systems. The publication trend visualizations included in this report (Figures 2.1 and 2.2) offer a concise illustration of this accelerating interest, placing this research paper within a broader movement toward transparency-centered AI development.

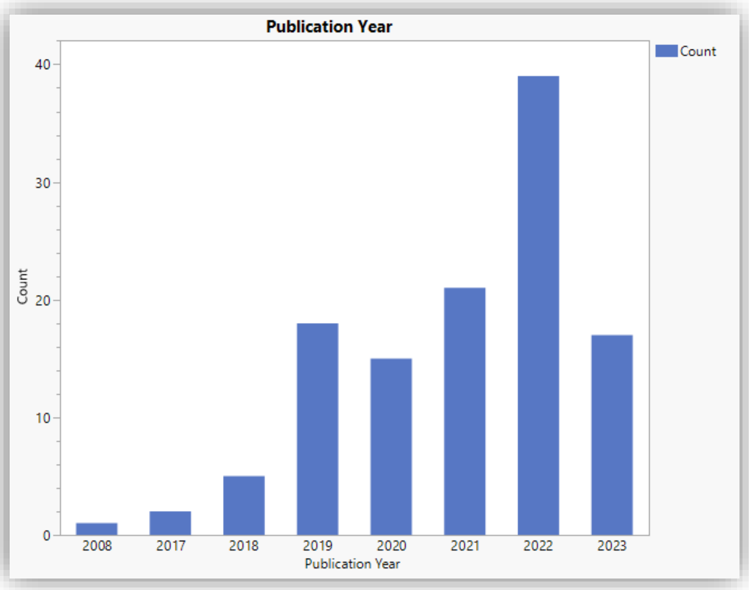
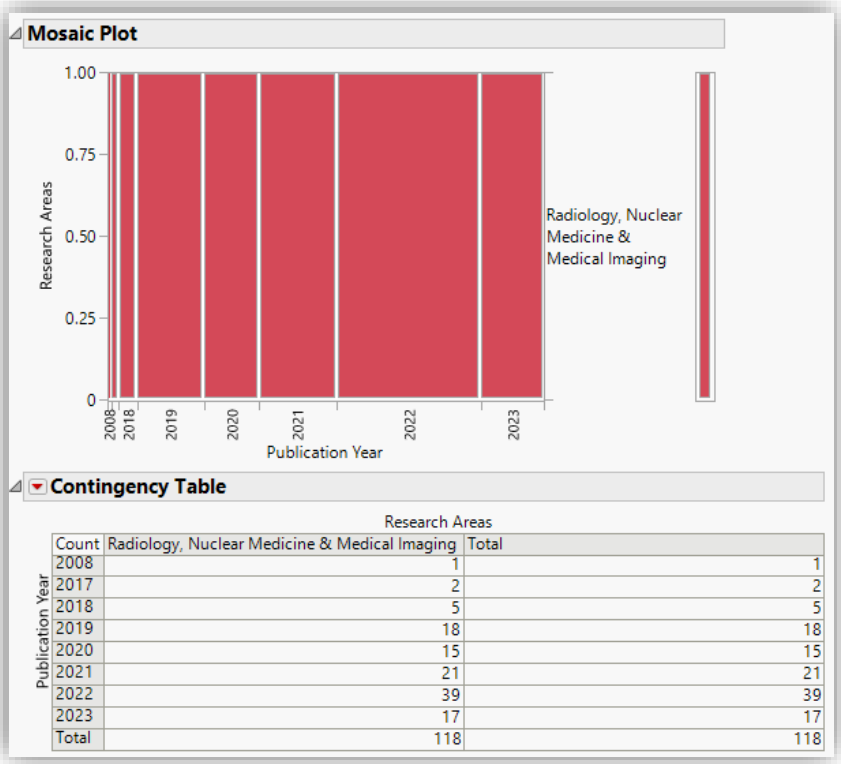


Figure 2.1 – Research articles focused on AI explainability in healthcare where area of research is Radiology, Nuclear Medicine and Medical Imaging



Publication Year

| | Research Areas | |
|---|---|---|
| Count | Radiology, Nuclear Medicine & Medical Imaging | Total |
| 2008 | 1 | 1 |
| 2017 | 2 | 2 |
| 2018 | 5 | 5 |
| 2019 | 18 | 18 |
| 2020 | 15 | 15 |
| 2021 | 21 | 21 |
| 2022 | 39 | 39 |
| 2023 | 17 | 17 |
| Total | 118 | 118 |

Figure 2.2 – Research articles focused on AI explainability in healthcare where area of research is Radiology, Nuclear Medicine and Medical Imaging

The results of that dataset coincide with the information from Google Trends pertaining to XAI in Healthcare (Figure 2.3) and Explainable AI in Healthcare (Figure 2.4) further solidifying the notion that this is a growing and expanding field.

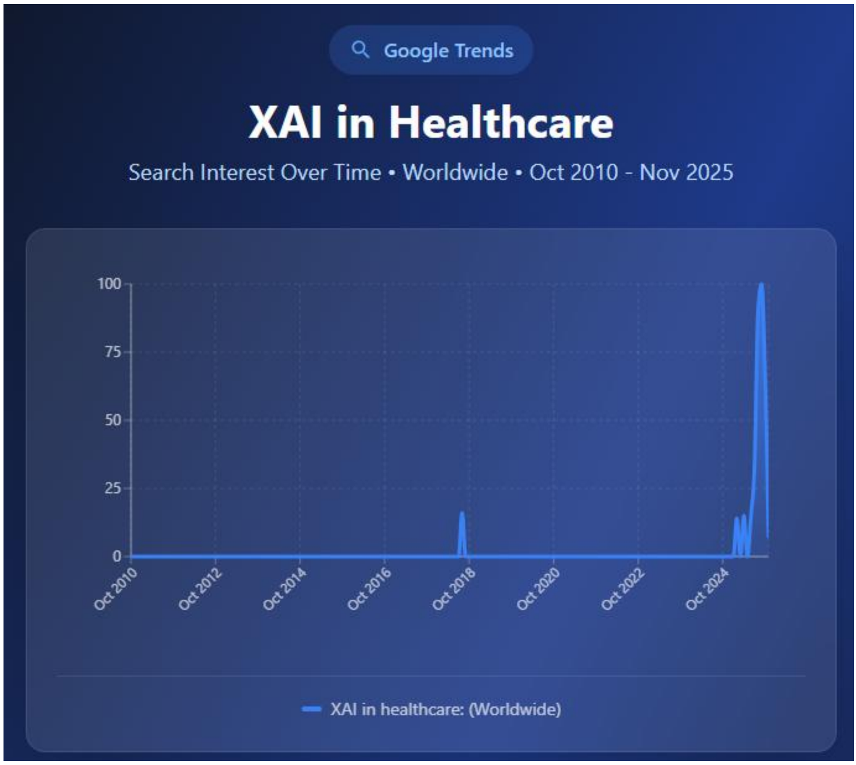


Figure 2.3

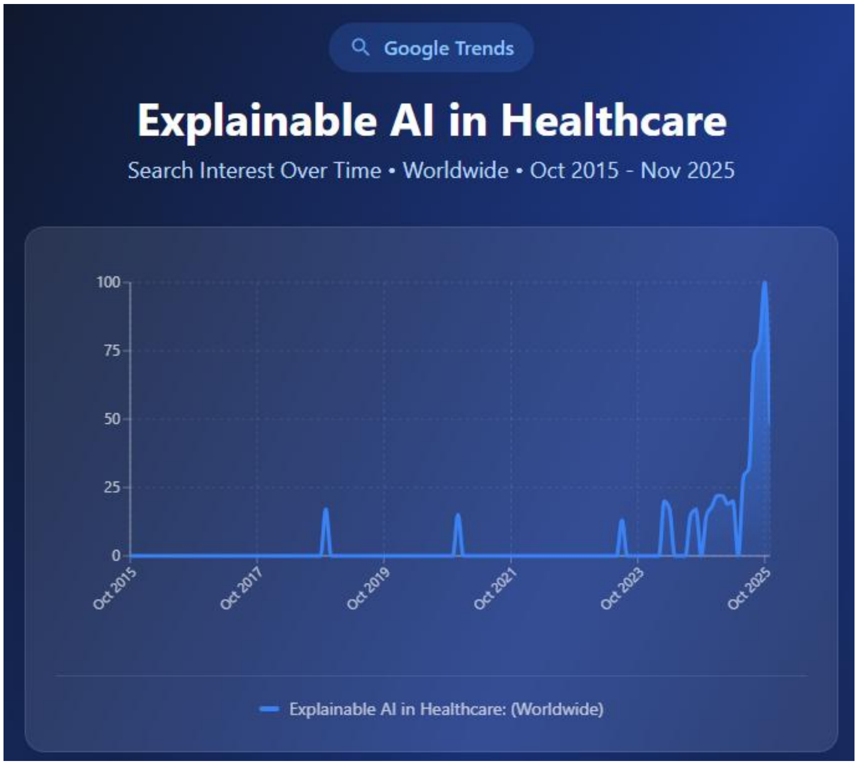


Figure 2.4

However, despite the strengths of the current study, including a well-controlled experimental pipeline and the comparative application of multiple XAI techniques, there are several limitations, due to limited resources, that must be acknowledged. First, the use of a single CNN architecture (ResNet18) provides a consistent evaluation framework but limits generalizability to other model families. Second, the binary classification task does not fully capture the complexity of multilabel or multi-pathology chest X-ray interpretation encountered in clinical practice. Third, although the RSNA dataset is large and diverse, variability across institutions, imaging devices, and patient populations may still introduce hidden biases affecting explainability outcomes. Finally, while Grad-CAM, LIME, and SHAP offer valuable insight, they remain post-hoc techniques subject to issues of stability, fidelity, and potential for misleading interpretations, which are concerns emphasized repeatedly in the XAI literature [17,19,20].

Nonetheless, the results of this study meaningfully contribute to the ongoing discourse on medical XAI. They demonstrate how explainability enhances the interpretive value of diagnostic models, reveal the benefits of multi-method triangulation, and situate these findings within a rapidly evolving research landscape that increasingly prioritizes transparency, trustworthiness, and clinician-centered model evaluation. These insights form a foundation for the following concluding remarks, which synthesize the study’s core contributions and identify potential directions for future work in developing clinically deployable, explainable AI systems.

# 6. Conclusion

This study explored the role of explainable artificial intelligence (XAI) in improving the transparency and clinical viability of deep learning models for medical imaging. Using a ResNet18 classifier trained on the RSNA Pneumonia Detection dataset, the experimental analysis demonstrated that convolutional neural networks can achieve strong discriminative performance, particularly in sensitivity, when tasked with identifying pneumonia from chest radiographs. Yet, as emphasized by prior research and reinforced by the present findings, predictive accuracy alone is insufficient for real-world deployment in high-stakes clinical environments. The critical barrier remains interpretability: clinicians must be able to understand and evaluate the reasoning behind model predictions before such systems can be safely integrated into diagnostic workflows.

A central contribution of this work lies in its controlled, side-by-side comparison of three leading XAI methods (Grad-CAM, LIME, and SHAP) applied to the exact same model and dataset. While these methods are frequently cited in the literature, they are not often examined together in a unified experimental pipeline. This study addresses this gap by demonstrating how each method provides a distinct interpretive lens. Grad-CAM offers spatially focused localization maps, LIME yields granular, segmentation-based insight, and SHapley-based attributions quantify the relative contribution of individual image regions. Importantly, although these methods differ conceptually and computationally, they exhibited meaningful convergence in their identification of clinically relevant pulmonary regions. This multi-method triangulation provides greater confidence in the model’s reasoning than any single XAI technique could offer on its own.

This finding reinforces an increasingly prominent theme in the medical XAI literature, which is that no single interpretability method is sufficient for clinical trustworthiness. Instead, robust interpretability emerges from the alignment of multiple, independent explanatory approaches, each revealing different aspects of the model’s decision-making process. By empirically demonstrating this complementary relationship, this study provides practical evidence for why future clinical AI systems should incorporate multi-method interpretability frameworks rather than relying solely on one explanatory tool. This insight contributes directly to ongoing discussions within radiology, regulatory guidance, and trustworthy AI research, where multi-method explanation strategies are encouraged but rarely operationalized.

The broader context in which this work is situated further underscores its relevance. An analysis of publication trends from the “Explainable AI and Interpretable Machine Learning in Healthcare” dataset shows a clear and accelerating rise in radiology-focused XAI research, reflecting a growing community-wide recognition that explainability is indispensable for safe, ethical, and clinically acceptable AI integration. This aligns with the present study’s findings, which highlight how explainability enriches not only model transparency but also model evaluation, error analysis, and clinical interpretability.

Despite the strengths of this work which include its unified experimental environment, comparative interpretability approach, and alignment with contemporary research trends, several limitations can be acknowledged. The study focuses on a single CNN architecture and a binary classification task, which limits generalizability to more complex diagnostic settings. Additionally, Grad-CAM, LIME, and SHAP remain post-hoc methods subject to known challenges in fidelity and stability. These limitations reflect active research challenges and point toward promising future directions.

Future work may explore the integration of intrinsic or self-explainable architectures, expansion to multi-label or multi-pathology detection tasks, and the development of standardized quantitative metrics for explanation quality. Furthermore, evaluating clinician interpretation of explanations, and integrating explainability into regulatory and workflow-aware design frameworks, remain essential steps toward closing the gap between algorithmic performance and clinical usability.

In conclusion, this study demonstrates that explainability is not merely a supporting feature but a foundational requirement for the trustworthy deployment of AI in medical imaging. By comparing and triangulating multiple XAI methods, the work illustrates how transparent reasoning pathways can be revealed, validated, and aligned with clinical expectations. As AI continues to evolve within healthcare, multi-method explainability will remain central to ensuring that diagnostic models are not only accurate but also interpretable, reliable, and ethically grounded, ultimately enabling their safe and meaningful adoption in real-world clinical practice.

# References

**[1]** Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., … Ng, A. Y. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning.* arXiv:1711.05225. https://arxiv.org/abs/1711.05225

**[2]** Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.* JAMA, 316(22), 2402–2410. https://pubmed.ncbi.nlm.nih.gov/27898976/

**[3]** Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R., … Reiner, R. C. (2018). *Estimates of the global, regional, and national burden of lower respiratory infections, 1990–2016.* The Lancet Infectious Diseases, 18(11), 1191–1210. https://www.thelancet.com/action/showPdf?pii=S1473-3099%2818%2930310-4

**[4]** Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). *Explainability for artificial intelligence in healthcare: A multidisciplinary perspective.* BMC Medical Informatics and Decision Making, 20(1), 310. https://pubmed.ncbi.nlm.nih.gov/33256715/

**[5]** Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). *What clinicians want: Contextualizing explainable machine learning for clinical end use.* Proceedings of the 4th Machine Learning for Healthcare Conference. https://proceedings.mlr.press/v106/tonekaboni19a/tonekaboni19a.pdf

**[6]** Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). *Hidden stratification causes clinically meaningful failures in machine learning for medical imaging.* Proceedings of the ACM Conference on Health, Inference, and Learning. https://dl.acm.org/doi/pdf/10.1145/3368555.3384468

**[7]** Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). *Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs.* PLOS Medicine, 15(11), e1002683. https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.1002683&type=printable

**[8]** Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). *Grad-CAM: Visual explanations from deep networks via gradient-based localization.* International Journal of Computer Vision, 128, 336–359. https://arxiv.org/pdf/1610.02391

**[9]** Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier.* KDD Conference. https://arxiv.org/pdf/1602.04938

**[10]** Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* NeurIPS; arXiv:1705.07874. https://arxiv.org/pdf/1705.07874

**[11]** Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). *Green AI.* Communications of the ACM, 63(12), 54–63. https://dl.acm.org/doi/pdf/10.1145/3381831

**[12]** Patterson, D. et al. (2021). *Carbon emissions and large neural network training.* arXiv:2104.10350. https://arxiv.org/pdf/2104.10350

**[13]** Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). *Counting carbon: A survey of factors influencing the emissions of machine learning.* arXiv:2302.08476. https://arxiv.org/pdf/2302.08476

**[14]** He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. arXiv:1512.03385v1. https://arxiv.org/pdf/1512.03385

**[15]** CUDA GPU Compute Capability. Nvidia Developer. https://developer.nvidia.com/cuda-gpus

**[16]** Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*. https://arxiv.org/pdf/2204.08610

**[17]** E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314. https://ieeexplore.ieee.org/document/9233366

**[18]** Chaddad, A., Hu, Y., Wu, Y., Wen, B., & Kateb, R. (2025). Generalizable and explainable deep learning for medical image computing: An overview. *Current Opinion in Biomedical Engineering*, 100567. https://arxiv.org/pdf/2503.08420

**[19]** Hou, J., Liu, S., Bie, Y., Wang, H., Tan, A., Luo, L., & Chen, H. (2024). Self-explainable ai for medical image analysis: A survey and new outlooks. *arXiv preprint arXiv:2410.02331*. https://arxiv.org/html/2410.02331v1

**[20]** Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: *A comprehensive survey of the terminology, design choices, and evaluation strategies*, *Journal of Biomedical Informatics,Volume 113*, 103655,ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103655

**[21]** Sun, Q., Akman, A., & Schuller, B. W. (2025). Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare*. https://arxiv.org/pdf/2412.01829

**[22]** Alghamdi, Shatha; Mehmood, Rashid; Alqurashi, Fahad; Alzahrani, Ali (2025), "Explainable AI (XAI) and Interpretable Machine Learning (IML) in Healthcare Dataset", Mendeley Data, V1, doi: 10.17632/5tcdzzsmx8.1. https://data.mendeley.com/datasets/5tcdzzsmx8/1

**[23]** Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*. https://arxiv.org/abs/1912.01703

**[24]** Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., ... & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*. https://arxiv.org/abs/2009.07896

**[25]** NumPy Developers. (2024). *For downstream package authors - NumPy 2.0-specific advice*. NumPy Documentation. https://numpy.org/devdocs/dev/depending_on_numpy.html

**[26]** pydicom Developers. (2024). *pydicom.pixels.apply_voi_lut: Apply a VOI LUT or windowing operation.* pydicom API reference. https://pydicom.github.io/pydicom/dev/reference/generated/pydicom.pixels.apply_voi_lut.html

**[27]** PyTorch Contributors. (2024). *Torchvision Models - Pre-trained weights & input normalization*. PyTorch Documentation. https://docs.pytorch.org/vision/stable/models.html