People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University of Skikda
Faculty of Science
Department of Computer Science

Branch: Mathematics and Computer Science

Major: Computer Science

Option: Artificial Intelligence

Master Thesis
Title :

Explainable AI for Medical Image Analysis: A Comparative Study of Post Hoc and Model-Based Explainability Techniques

Presented by: LATRACH Mohamed Ali
SASSENE Abderraouf

Thesis Supervisor:

Dr. Adel Lahsasna University of Skikda Supervisor

Skikda 2025

Dedication

This work is dedicated to our families, whose unwavering support and encouragement have made this journey possible. To our parents and siblings, thank you for your patience, love, and belief in us even during the most challenging times. To our friends, who provided laughter, motivation, and a listening ear, your companionship has been invaluable. We also dedicate this thesis to our supervisor, whose guidance, constructive feedback, and mentorship shaped this research and inspired us to strive for excellence. Finally, we express our gratitude to our professors and teachers, whose knowledge and dedication instilled in us a passion for learning and research. Without you all, this work would not have been realized.

Abstract

Explainable artificial intelligence (XAI) is critical for building trust and ensuring safe deployment of deep learning models in healthcare. This thesis presents a comparative study of two XAI approaches—Grad-CAM (a post hoc method) and ProtoPNet (a model-based method)—applied to multilabel chest X-ray interpretation. Both models were trained and evaluated on the VinDr-CXR dataset under identical conditions. The Grad-CAM approach, built on an EfficientNetV2-S backbone, achieved superior predictive performance (macro ROC AUC = 0.86, macro F1 = 0.72) and generated clear, reliable heatmaps with minimal computational overhead (hit-rate = 64%, mIoU = 42%). In contrast, ProtoPNet, which learns prototypical image patches for inherently interpretable "this looks like that" explanations, produced lower classification metrics (macro ROC AUC = 0.73, macro F1 = 0.52) and weaker localization performance (hit-rate = 0.7%, mIoU = 42%) while incurring approximately 25 % more inference time. Despite these drawbacks, ProtoPNet's case-based explanations more closely align with clinical reasoning, offering tangible examples that radiologists find meaningful. Our findings indicate that, for rapid deployment and high accuracy, post hoc methods like Grad-CAM are preferable. However, the richer, example-driven explanations of ProtoPNet highlight the need to further refine prototype-based models—by optimizing prototype selection and expanding datasets—so that they can deliver both strong performance and intuitively interpretable results in real-world clinical settings.

Contents

List of Figures

List of Tables

Introduction

Artificial intelligence (AI) has transformed many areas in the past decade, and healthcare is no different. In particular, AI-driven medical image analysis has shown great potential for improving diagnostic accuracy, simplifying workflows, and eventually enhancing patient outcomes. Imaging types such as computed tomography (CT), magnetic resonance imaging (MRI), and X-rays provide detailed, high-dimensional data that can be used by deep learning models to detect, classify, and locate diseases with performance often matching or outperforming human experts. For example, AI systems for multilabel chest X-ray interpretation have shown accuracy comparable to experienced radiologists, opening new paths for fast diagnosis and treatment planning [1, 2]. However, despite these good results, the use of AI models in clinical practice has been blocked by a basic issue: the lack of transparency in many deep neural networks. Clinicians and regulators both want clear, understandable explanations to trust and check AI-driven decisions in critical care settings.

The problem of model interpretability has caused a growing interest in Explainable AI (XAI), which aims to develop methods that explain how and why AI models make certain predictions. In a medical setting, explainability is not just a theoretical concern; it directly affects clinical trust, legal compliance, and patient safety. Clear explanations can show whether a model is focusing on relevant features (e.g., chest anatomy, opacity patterns) or on irrelevant artifacts (e.g., image noise, scanner-specific markings). These insights then help refine models, encourage teamwork between experts, and support the integration of AI tools into everyday diagnostic routines [3, 4]. Broadly speaking, XAI approaches for medical image analysis fall into two categories: post hoc methods, which try to interpret

a trained "black-box" model, and model-based (or inherently interpretable) methods, where explainability is part of the design.

Post hoc explainability techniques—such as saliency maps, gradient-based attribution, and class activation mappings—create visual or textual explanations after the model is trained. Among these, Grad-CAM (Gradient-weighted Class Activation Mapping) is one of the most popular tools in medical imaging research because of its ease of use and clear visuals [2]. By highlighting areas of an input image that most influence the model's output, Grad-CAM produces heatmaps overlaid on original scans, letting clinicians check whether the model's focus matches known disease regions. Despite its usefulness, Grad-CAM and similar methods rely on certain assumptions (e.g., linear combinations of activation maps) and can sometimes give vague or misleading explanations when working with complex, high-resolution medical images.

In contrast, model-based explainability methods build interpretability into the network's structure. Prototype-based networks (ProtoPNet) are one example: these architectures learn a set of prototypical image patches during training and classify new inputs by comparing them to these prototypes. Each prototype represents an understandable concept (e.g., a typical lung opacity), and the network's decision can be traced to the similarity scores between input regions and prototypes [5, 4]. This straightforward design lets clinicians view prototypes, check their clinical relevance, and understand the reason behind each classification. However, prototype-based methods often need careful choices (e.g., number and size of prototypes) and may involve a trade-off between interpretability and raw performance.

Given the combined strengths and weaknesses of post hoc and model-based explainability approaches, a comparative study is needed to see their advantages in medical image analysis. To our knowledge, few studies have directly compared both XAI categories on the same diagnostic task. This thesis aims to address that gap by conducting two case studies focused on multilabel chest X-ray interpretation: (1) a convolutional neural network (CNN) backbone with Grad-CAM to provide post hoc saliency maps, and (2) a ProtoPNet-based architecture offering inherent interpretability through learned prototypes.

The remainder of this introduction is organized as follows. First, we lay out the main

problem that drives our work: while deep learning models achieve top accuracy in medical image tasks, their lack of transparency creates serious barriers to clinical use. We then give an overview of existing explainability methods in medical imaging, contrasting post hoc approaches like Grad-CAM with model-based strategies like ProtoPNet. Finally, we state the objectives and scope of this thesis, outlining how our comparative study is organized and what readers will find in later chapters.

Problematique and Motivation

Deep neural networks are great at extracting complex, layered features from imaging data, but they give little insight into how specific predictions are made. In medical diagnosis scenarios, this lack of transparency is more than an academic issue: it can hide model mistakes, make quality checks harder, and slow regulatory approval. For instance, imagine a CNN that predicts multiple conditions from a chest X-ray. If the model's heatmap highlights a normal rib area instead of a lung lesion, clinicians may doubt the system's reliability. Therefore, explainability is not just a desirable feature but a necessary requirement for trust, accountability, and safe use in healthcare [3].

Different stakeholders have different explainability needs. Radiologists often want localized, pixel-level explanations that justify a segmentation or classification; hospital managers and regulators may need higher-level reasoning to ensure models meet ethical and legal standards; and patients might seek clear explanations to feel confident about AI-assisted diagnoses. Thus, XAI methods must meet a range of interpretability needs without sacrificing too much accuracy or speed.

Thesis Objectives and Contributions

This thesis offers a direct comparison of post hoc and model-based explainability methods applied to multilabel chest X-ray interpretation. The two case studies are:

1. **CNN + Grad-CAM (Post Hoc Explainability):** We build a standard CNN to perform multilabel classification on chest X-rays, then use Grad-CAM to create

heatmaps showing regions that influenced each label prediction. These maps are checked for clinical relevance and localization accuracy.

2. **ProtoPNet (Model-Based Explainability):** We design a prototype-based network that learns a set of prototypical lung patterns during training. Each prototype can be displayed as a representative image patch, and the classification is explained by showing similarity scores between input regions and prototypes.

Key contributions of this thesis include:

- A clear comparison of post hoc and model-based explainability techniques on a shared multilabel chest X-ray benchmark, evaluating both prediction performance and interpretability.
- An analysis of the balance between transparency and predictive performance, showing when one approach may be better than the other.
- Practical advice for integrating explainable AI into medical image workflows, focusing on ease of use for clinicians and meeting healthcare guidelines.

By presenting a direct comparison of two common XAI approaches in a clinically relevant setting, this work aims to guide future research and help practitioners choose the right explainability methods for medical imaging tasks. The rest of the thesis is organized as follows:

Chapter 1 Background on AI in Medical Imaging and the Need for Explainability

Chapter 2 Literature review and State Of The Art

Chapter 3 Methodology

Chapter 4 Results and Evaluation

Chapter 5 Discussion: Insights, Limitations, and Recommendations

Conclusion

Chapter 1

Towards Transparent AI in Healthcare

1.1 Historical Background of AI in Healthcare

Artificial intelligence (AI) has long been recognized as a powerful tool for medicine. Early work in AI dates back to the 1940s–1960s, when pioneers formulated theoretical models of neural networks and basic learning machines [6]. In the 1970s and 1980s, AI in healthcare primarily took the form of rule-based and knowledge-based expert systems. For example, systems like MYCIN demonstrated the potential for computer programs to aid in diagnosis of infections by encoding medical rules [6]. These early applications, though limited by the computing power of the time, showed that AI could support clinicians' decision-making by providing computer-guided advice. Over time, however, interest in AI in medicine waxed and waned. AI first boomed, then suffered an "AI winter" when excitement outpaced technical capability, and finally revived with better algorithms and hardware [6].

By the late 20th century, statistical machine learning methods (such as *Decision Trees* [7], *Support Vector Machines* (SVMs) [8], *Naïve Bayes Classifiers* [9], and *K-Nearest Neighbors* (KNN) [10]) began to supplement expert systems.

- **Decision Trees:** A non-parametric supervised learning algorithm, which is used in classification by partitioning data based on features.
- **Support Vector Machines (SVMs):** Introduced in the 1960s, SVMs are used for classification tasks by finding the hyperplane that best separates the data and

are widely used in medical imaging.[11]

- **Other ML Methods:** Like Naïve Bayes classifiers, which rely on probability theory, and KNN that assigns classifications based on the nearest labeled samples.[11]

Researchers also applied early neural network models to medical data, for instance in radiographic image analysis. However, these approaches often required careful feature engineering and still offered only incremental gains. It was not until the 21st century, with the advent of high-speed graphics processors (GPUs) and large digital datasets, that AI's potential in medicine truly began to accelerate [6]. As one review notes, the refinement of learning algorithms, development of deep convolutional networks, and availability of powerful computing revived interest in AI. AI soon transitioned from theoretical models to practical medical tools [6].

1.2 Emergence and Evolution of Machine Learning and Deep Learning

Machine learning (ML) and its subfield deep learning (DL) have driven the recent resurgence of AI in medicine. ML encompasses a range of algorithms that learn from data, while DL refers to multi-layer neural networks that can automatically extract features from raw inputs. The foundations of neural networks date back to work by McCulloch and Pitts in the 1940s and Rosenblatt's perceptron in 1958, but training deep networks was initially hampered by lack of data and computing power. In 1986, backpropagation made multi-layer networks more practical, but it was not until the 2010s that DL began to dominate AI. In 2012, a landmark event known as the "ImageNet moment" occurred: a deep convolutional neural network won the ImageNet object-recognition competition by a large margin. This achievement proved the power of DL for image tasks [6].

Since then, advances in ML and DL have been rapid. Researchers developed many new network architectures (e.g. ResNet, DenseNet) and techniques (dropout, batch normalization, data augmentation) that steadily improved performance. The widespread use of GPUs and cloud computing enabled researchers to train very large models on massive datasets [6]. Today's DL models range from convolutional neural networks (CNNs) for

images, to recurrent networks and transformers for sequential data, to generative models (GANs, diffusion models) for creating new data. In healthcare, these technical advances have meant that models which previously struggled now achieve breakthrough results. For example, CNNs once confined to computer vision labs are now routinely applied to medical imaging tasks. As Zhou and Greenspan observe, modern DL "has been widely used in various medical imaging tasks and has achieved remarkable success" [12].

Table 1.1: Key Milestones in AI and Healthcare

Year	Event
1961	**ELIZA**: Early chatbot, precursor to healthcare chatbots
1972	**MYCIN**: Expert system for diagnosing bacterial infections
1976	**EMYCIN**: Framework for medical diagnostic systems
1986	**Deep Learning Introduced**: Key for modern healthcare applications like imaging
1996	**DXplain**: Clinical decision support system for diagnosis
2007	**CAD in Endoscopy**: AI-enhanced diagnostic imaging in gastroenterology
2011	**Alexa Released**: Voice assistant with potential for patient monitoring
2014	**Siri Introduced**: Voice assistant with potential healthcare uses
2015	**ARTERYS Approved**: AI-powered cloud platform for medical imaging
2017	**MANDY Chatbot**: Healthcare-specific chatbot for patient interaction
2018–2022	**AI in Gastroenterology**: Broad application of AI in diagnostics and treatment

1.3 Deep Learning Applications in Medical Imaging

Medical imaging has been one of the most impactful areas for AI. Imaging modalities (X-ray, CT, MRI, ultrasound, pathology slides, etc.) generate about 90% of healthcare data [12], making them a rich domain for AI analysis.

Modern DL approaches have achieved success on many imaging tasks: detection, classification, segmentation, and quantification. For example, in chest radiology a deep network (CheXNeXt) was trained on tens of thousands of X-rays and demonstrated performance comparable to expert radiologists in identifying 14 pathology types [13]. In digital pathology, CNNs have matched pathologists in detecting tumor metastases in lymph-node slides [12]. In ophthalmology, a deep learning algorithm for diabetic retinopathy achieved expert-level sensitivity and specificity on retinal fundus photos [12]. In dermatology, models have classified skin lesion images at a level similar to dermatologists (e.g. melanoma detection).

Overall, deep learning has "propelled us into the AI era" for medical imaging [12]. Segmentation tasks have seen dramatic improvements thanks to networks like U-Net (2015) designed for biomedical images. These models automatically learn hierarchical features, from edges to textures to anatomical structures, enabling accurate delineation of organs or lesions. Classification tasks, such as diagnosing pneumonia or fractures on X-rays, have also benefited from DL's feature learning. The combination of large annotated datasets and powerful networks has yielded impressive results on many challenges. As one review summarizes, deep learning excels at medical image interpretation and its success comes largely from big data and high-performance computing [12].

1.4 Human–AI Collaboration and Trust in Medical Contexts

As AI models become more powerful, their role in clinical decision-making is evolving towards collaboration with human clinicians. Instead of replacing doctors, AI is often envisioned as a diagnostic aid. Trust and effective interaction are crucial: clinicians must understand when and how to rely on AI output. Studies have shown that when

AI predictions are accompanied by explanations, human decision-makers perform better. For instance, Senoner et al. demonstrated that providing heatmap visual explanations alongside AI recommendations significantly improved diagnostic accuracy [14].

However, achieving such collaboration requires transparency. Clinicians need human-understandable explanations of AI decisions to trust and effectively use the system [15]. Explainable outputs like annotated images, feature highlights, or example cases can make AI reasoning clearer.

1.5 Regulatory and Ethical Motivations for Explainability

In high-stakes fields like healthcare, there's a strong imperative for AI transparency. The European Union's AI Act will classify medical AI as high-risk, mandating that systems provide clear information on their logic and limitations [15]. Similarly, the FDA and other regulators emphasize transparency of machine-learning-enabled medical devices [16].

From an ethical standpoint, biased or opaque AI can harm patients. Reviews highlight concerns about AI-driven disparities and call for fairness measures [17]. Ethical frameworks stress accountability and the "right to explanation."

1.6 Roadmap to Explainable AI

Explainable AI (XAI) refers to methods that make a model's behavior or decisions understandable to humans. Broadly, XAI approaches are categorized into intrinsic methods (interpretable-by-design models) and post-hoc methods (tools explaining black-box models).

Common XAI techniques include:

- Visual explanations (e.g. Grad-CAM heatmaps) [14]
- Surrogate models (LIME, SHAP feature importance)
- Example-based explanations (similar-case retrieval)

- Concept-based methods (TCAV concept influence)

In the following chapters, we will compare these XAI approaches in medical imaging tasks, evaluating their clarity, reliability, and clinical utility.

1.7 Conclusion

The evolution of artificial intelligence in healthcare has been marked by significant milestones, from early rule-based systems like MYCIN to the advent of deep learning and its transformative impact on medical imaging. As AI technologies continue to advance, their integration into clinical practice necessitates a focus on transparency, explainability, and ethical considerations. Human–AI collaboration, underpinned by trust and clear communication, is essential for the effective and equitable application of AI in medicine. Future developments should prioritize explainable AI to ensure that healthcare professionals can confidently leverage these tools to enhance patient care.

Chapter 2

Literature review

2.1 Introduction to XAI in Medical Imaging

2.1.1 Applications of AI in Medical Imaging

Advancements in medical imaging and artificial intelligence (AI) have transformed healthcare, enabling early disease detection, precise diagnosis, personalized treatment planning, and improved patient outcomes. Imaging modalities such as X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) generate vast datasets, necessitating efficient analysis where AI excels [3]. Deep learning algorithms, trained on large datasets, identify complex patterns imperceptible to the human eye, enhancing diagnostic accuracy and efficiency [18]. Recent developments have further advanced AI capabilities, with algorithms predicting diseases like breast cancer years before manifestation and reducing diagnosis time by up to 30% [19, 20].

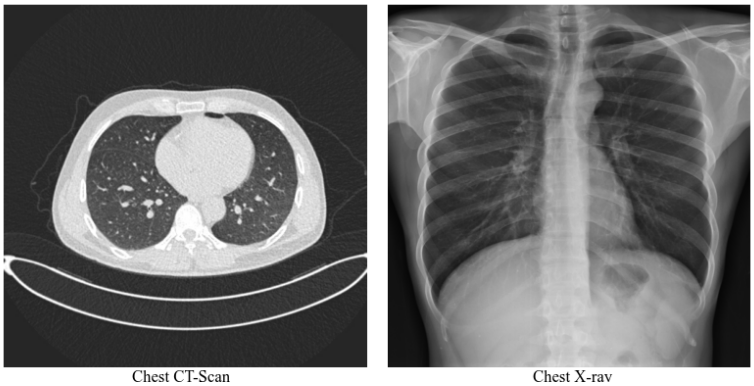

Figure 2.1: Examples of Imaging Modalities

2.1.2 Challenges of Black-Box Models

Deep learning models, while powerful, are complex due to intricate architectures (e.g., convolutional neural networks), numerous parameters, and optimization processes. This complexity enables capturing intricate patterns but poses challenges in high-stakes applications like healthcare, where interpretability is crucial [21]. The "black-box" nature of these models, where internal processes are opaque, raises concerns about bias, transparency, and accountability, limiting clinical adoption. For instance, biased training data can lead to misdiagnoses in underrepresented groups, underscoring the need for explainability [22].

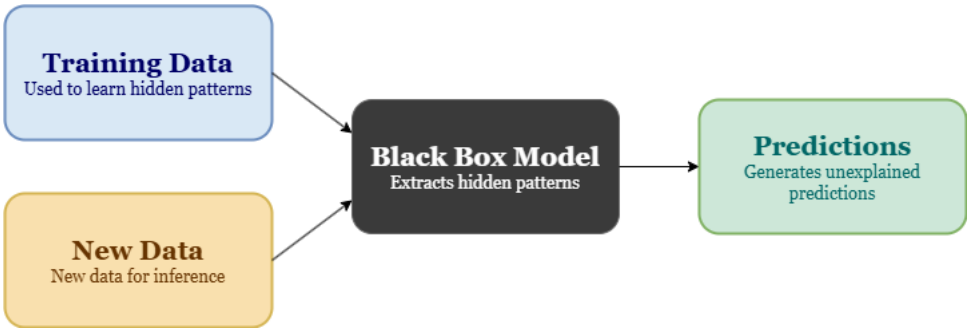

Figure 2.2: Black-Box Model Workflow

2.1.3 Transparency and Trust

AI aims to assist clinicians, not replace them, requiring trust built through transparent decision-making. XAI provides interpretable explanations, such as identifying influential image regions, fostering clinician confidence [23]. Regulatory frameworks, like the EU's GDPR (Article 15), mandate transparency, and ongoing efforts to standardize XAI evaluation ensure clinical relevance [24]. Compliance enhances patient trust and supports ethical AI integration.

2.2 Classification of Explainability Techniques

2.2.1 Model-Based vs. Post Hoc Explanations

- **Model-Based**: These methods integrate explainability into the model, using simpler models (e.g., decision trees) or techniques like attention mechanisms in deep learning to enhance transparency [23]. Example: Attention-based CNNs highlight relevant image regions.

- **Post Hoc**: Post-hoc explainability methods analyze and interpret the decision-making process of a trained machine learning model after it has made predictions, providing insights into how the model arrived at its outputs. An important distinction between post hoc explanation and model-based explanation is that the former trains a neural network and subsequently attempts to explain the behavior of the ensuing black box network, whereas the latter forces the neural network to be explainable.[25]

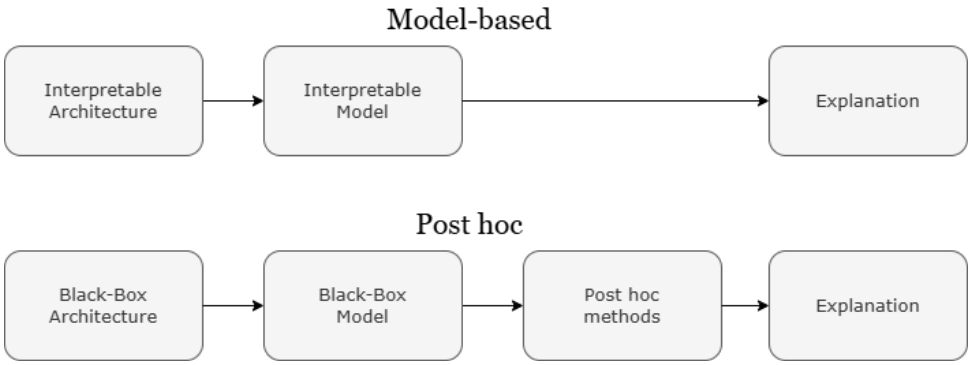

Figure 2.3: Model-Based vs. Post Hoc Explanations

2.2.2 Model-Specific vs. Model-Agnostic Approaches

- **Model-Specific**: Tailored to specific model architectures (e.g., CNNs). For instance, such a method might rely on attributes unique to a specific type of neural network. A disadvantage is that by focusing on model-specific explanation, we constrain our selection of neural networks, potentially omitting one that could more effectively align the output with the input data[23]. Example: Layer-wise Relevance Propagation for CNNs.

- **Model-Agnostic**: Model-agnostic explanation does not depend on the type of neural network, functioning only based on the input and output of the neural network. By modifying the input, the user can observe how the output of the neural network changes. This can thus reveal which regions influence the output. Model-agnostic explanation is inherently post hoc[23]. Example: SHAP values for any model.

2.2.3 Global vs. Local Explanations

- **Global Explanations**: Also known as dataset-level explanation, reveals general patterns learned by the neural network. For instance, global explanation can provide feature importance scores at the dataset level, indicating how much each feature contributes to the output across the entire dataset. As an example, one might observe from a neural network that – or even to what extent – high blood pressure increases the likelihood of a cardiac event. Another example of global explanation is the visualization of learned filters, showing which features are extracted by the neural network and how relevant they are to the given task[23]. Example: SHAP feature importance scores.

- **Local Explanations**: Provides clarification for a single input. In the context of lung disease, an input would correspond to an individual patient. Local explanation would therefore explain why a particular feature in an X-ray is crucial for diagnosing lung disease for that specific patient, whereas global explanation would illustrate the connection between that feature and lung disease across the entire dataset. Another instance of a local explanation is a saliency map emphasizing an affected region on a chest X-ray to indicate which part of the image predominantly influenced the classifier output 'disease.' Since this determines the specific section of the image that guided the classifier to its conclusion for that individual case, it constitutes a local explanation[23]. Example: Grad-CAM for a single X-ray.

2.3 XAI Methods for Medical Image Analysis

2.3.1 Concept Learning Models

These models predict human-interpretable concepts before final labels, enhancing transparency. Concept Bottleneck Models allow clinicians to modify concepts but may misalign with image features [26]. Capsule networks, like X-Caps, encode visual attributes for lung nodule malignancy, outperforming some CNNs [26]. Recent applications include predicting tumor characteristics in MRI scans [27].

2.3.2 Case-Based Models

ProtoPNet introduced prototype-based reasoning by learning class-discriminative prototypes—image patches representing key features (e.g., joint-space narrowing in arthritis)—and comparing them with input regions to generate explanations [28, 29]. XProtoNet extends this paradigm to chest radiography by dynamically predicting occurrence maps for each pathology and learning prototypes within those maps; it matches or exceeds performance on the NIH CXR-14 dataset while providing clear heatmaps of prototypical evidence [30]. NP-ProtoPNet further incorporates noise-robust similarity scoring to improve lesion localization, although it can be vulnerable to compression artifacts and dataset noise [31].

2.3.3 Concept Attribution

Concept Relevance Propagation (CRP) conditions traditional Layerwise Relevance Propagation (LRP) on specific clinical concepts encoded in hidden layers, producing concept-conditional relevance maps that reveal both the "what" (concept identity) and "where" (spatial localization) of model reasoning [32, 33]. In digital pathology, CRP has been applied to whole-slide images to verify histological features such as mitoses and glandular structures, aligning model insights with pathologist annotations [33]. Toolkits like Zennit-CRP automate this process for arbitrary architectures, enabling construction of "concept atlases" for 3D microscopy and radiological classification tasks [33].

2.3.4 Attribution Maps

Attribution maps are visual explanations that show which pixels or regions of an image mattered most for a neural network's decision. In other words, they answer "where did the model look?" by producing a heatmap over the input image. Below we describe four main types of attribution maps, each with a simple mathematical formulation.

- **Layerwise Relevance Propagation (LRP)** redistributes a model's final output score $f(\mathbf{x})$ back through the layers to the input pixels, such that the total relevance is conserved at each layer. For neuron i in layer l and neuron j in layer $l+1$, one common rule is:

$$R_i^{(l)} = \sum_j \frac{a_i^{(l)} w_{ij}}{\sum_{i'} a_{i'}^{(l)} w_{i'j} + \epsilon} R_j^{(l+1)},$$

  where $a_i^{(l)}$ is the activation of neuron i, w_{ij} the weight connecting $i \to j$, and ϵ a stabilizer. At the input layer, the $R_i^{(0)}$ form a pixel-wise relevance map [34].

- **Class Activation Maps (CAM)** use the fact that, after global average pooling (GAP), the score for class c is

$$S_c = \sum_k w_k^c \, \bar{f}_k, \quad \text{with} \quad \bar{f}_k = \frac{1}{Z} \sum_{x,y} f_k(x,y),$$

  where $f_k(x,y)$ is the k-th feature map and w_k^c its weight for class c. The CAM heatmap is

$$M_c(x,y) = \sum_k w_k^c \, f_k(x,y).$$

  This highlights regions that push the score S_c high [35]. Extensions like Score-CAM replace weights by score differences, and Grad-CAM++ refines them using higher-order gradients [36].

- **Grad-CAM** simplifies CAM by approximating w_k^c with the average gradient of the class score S_c w.r.t. each feature map:

$$\alpha_k^c = \frac{1}{Z} \sum_{x,y} \frac{\partial S_c}{\partial f_k(x,y)}, \quad M_c(x,y) = \text{ReLU}\Big(\sum_k \alpha_k^c \, f_k(x,y)\Big).$$

  The ReLU ensures we only visualize positive influences. Grad-CAM produces coarser maps but works for any convolutional net without retraining [2].

- **Perturbation-Based Methods** fit a simple, interpretable model (e.g. linear regression) locally around a single input. For example, LIME approximates f by

$$g(\mathbf{z}) = b + \sum_{i=1}^{d} \phi_i \, z_i,$$

where $\mathbf{z}$ indicates which superpixels are kept. The coefficients ϕ_i serve as importances. SHAP uses a weighted Shapley-value formula to assign each pixel i an attribution

$$\phi_i = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{i\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \Big[f(S \cup \{i\}) - f(S) \Big].$$

These methods are model-agnostic but can be slow, since they need many perturbations [37, 38].

2.3.5 Natural Language Explanations

Natural language explanations generate textual justifications aligned with medical reports.

- **Supervised NLEs** train encoder–decoder models on paired image–report datasets to produce clinically coherent rationales, e.g., "confluent consolidation in the right upper lung concerning for pneumonia" [39].
- **Multimodal Frameworks** (e.g., ExAID) combine Concept Activation Vectors with localization maps to produce fine-grained text alongside visual evidence for dermatology or pathology tasks [40].
- **Emergent Language Models** investigate dual-agent LSTM systems that learn to encode diagnostic features in discrete symbols without explicit text supervision, though clinical validation remains ongoing [41].

2.3.6 Latent Space Interpretation

Interpreting latent representations reveals what networks learn:

- **Activation Maximization** synthesizes inputs maximizing neuron activations, uncovering patterns like tumor texture preferences, albeit with limited diagnostic realism [42].
- **Network Dissection** quantifies alignment between hidden units and semantic concepts (e.g., calcifications, anatomical structures) using labeled concept datasets [43].
- **Disentangled Representations** via VAEs and concept whitening separate latent axes into clinically meaningful dimensions (e.g., lesion size vs. shape), enabling controlled perturbations for brain ventricle segmentation [44].
- **Emergent Symbolic Representations** map image patches to discrete symbols corresponding to immune cell markers or radiographic patterns, producing human-readable latent codes [41].

2.4 Comparative Analysis of XAI Methods

2.4.1 Categorization of XAI Methods

Table 2.1 summarizes XAI techniques and their applications.

Category	Technique	Example Methods
Model-Based	Attention Mechanisms	ViTs, Attention-Based CNNs
	Self-Explaining Neural Networks	ProtoPNet, SENN
Post Hoc	Saliency Maps	Grad-CAM, Integrated Gradients
	Perturbation Methods	LIME
	Feature Attribution	SHAP
Model-Specific	Feature Visualization	Filter Visualization, DeepDream
	Layerwise Relevance Propagation	LRP
Model-Agnostic	Surrogate Models	LIME, SHAP
	Example-Based Explanations	Counterfactual Explanations
Global Explanation	Feature Importance Analysis	SHAP
Local Explanation	Counterfactual Explanations	Counterfactual Explanations

Table 2.1: Categorization of XAI Methods for Deep Learning in Medical Imaging

2.4.2 State-of-the-Art Comparison

Post-hoc saliency methods (e.g., Grad-CAM) are easy to deploy but produce coarse heatmaps, with mIoU ranging from 0.07 to 0.26 [45]. Prototype-based methods (e.g., ProtoPNet) offer semantic clarity, with XProtoNet outperforming saliency baselines in localization [31]. Recent studies confirm ProtoPNet's clinical usability in mammography [46].

2.4.3 Localization Fidelity

Grad-CAM's coarse heatmaps limit precision in subtle pathologies, while ProtoPNet achieves tighter alignment with expert segmentations [47]. XProtoNet's localization precision is notable in breast cancer detection [31].

2.4.4 Semantic Clarity

Grad-CAM indicates "where" but not "what," lacking feature semantics, whereas ProtoPNet's prototype-based explanations align with clinician intuition [**?**]. Clinicians can inspect prototype galleries for clarity [31].

2.4.5 Robustness and Faithfulness

Saliency methods are sensitive to perturbations, while prototype networks anchor explanations to learned patterns, though spurious correlations remain a risk [48]. NP-ProtoPNet improves robustness [30].

2.4.6 Clinical Usability and Deployment

Grad-CAM's minimal overhead suits existing pipelines, while ProtoPNet requires architectural integration but offers ready explanations [49]. XProtoNet's integration into PACS viewers enhances usability [31].

2.5 Evaluation Metrics for XAI in Medical AI

2.5.1 Faithfulness

Faithfulness ensures explanations reflect the model's reasoning, measured by consistency and sufficiency [50]. SHAP maintains perfect consistency, unlike LIME [50].

2.5.2 Consistency and Sufficiency

Consistency ensures similar explanations yield similar predictions, while sufficiency guarantees explanations justify decisions. Decision trees and Anchors enhance interpretability [50].

2.5.3 Robustness

Robustness measures explanation stability under input changes, critical for medical AI. Regularization and stability tests enhance reliability [50]. Standardized metrics are needed [24].

2.5.4 Practical Implications

Balancing accuracy and explainability requires domain-specific validation. Clinical XAI Guidelines propose criteria like understandability and truthfulness [24].

2.6 Ethical Considerations in XAI

Ethical challenges in XAI include data privacy, bias mitigation, and accountability. During data collection, respecting patient rights and ensuring data quality are crucial [51]. Model development must prioritize safety, efficacy, and fairness, while deployment requires clinicians to retain responsibility [51]. XAI enhances transparency, addressing these concerns by explaining decisions [27].

2.7 Conclusion

This chapter explores XAI's role in medical image analysis, addressing the opacity of deep learning models. AI's transformative impact on diagnostics is tempered by black-box challenges, necessitating XAI for transparency and trust. We categorized XAI techniques into model-based vs. post hoc, model-specific vs. model-agnostic, and global vs. local, detailing methods like concept learning, case-based models, and attribution maps. Comparative analysis highlighted trade-offs between Grad-CAM and ProtoPNet, while evaluation metrics emphasized faithfulness and robustness. Ethical considerations underscored XAI's role in ensuring fairness and accountability. Future directions include customized XAI techniques, standardized metrics, and biological explanations, aligning with clinicians' needs to enhance clinical adoption

Chapter 3

Methodology

3.1 Introduction

Explainable Artificial Intelligence (XAI) is critical in medical imaging, where opaque models risk undermining clinician trust and may obscure diagnostic errors. In this chapter, we present a *comparative case-study* framework to investigate transparency in multi-label chest X-ray classification using the VinDr-CXR dataset. Specifically, we contrast:

1. a **post hoc** saliency-map method, applying Grad-CAM to an EfficientNetV2-S backbone fine-tuned under an asymmetric loss, and
2. an **intrinsic** prototype-based approach, using ProtoPNet to embed explanations directly via learned class prototypes.

While Grad-CAM offers flexible, model-agnostic visualizations with minimal additional training, questions remain about its fidelity and spatial precision. By contrast, ProtoPNet provides built-in interpretability through prototype matching, at the cost of added architectural steps (e.g. prototype pushing) and potential prototype selection risks. Our design holds dataset, splits, and evaluation criteria constant, enabling a focused assessment of whether a lightweight, post hoc explainer can achieve comparable transparency and clinical relevance to an interpretable-by-design model in high-stakes diagnosis tasks.

3.1.1 Objectives

To structure our comparative analysis, we define the following high-level objectives for each case study:

1. **Grad-CAM on EfficientNetV2-S**
    - Adapt and train EfficientNetV2-S on VinDr-CXR with an asymmetric loss, then generate Grad-CAM heatmaps to visualize decision regions.
    - Quantitatively evaluate explanation fidelity and consistency across disease labels.
2. **ProtoPNet Model-Based Explanation**
    - Train ProtoPNet on the same data splits, perform prototype pushing to align prototypes with real lesions, and inspect prototype relevance.
    - Assess inherent interpretability by examining prototype–image matches.
3. **Comparative Evaluation**
    - Compare both approaches on fidelity, granularity, and clinical utility using consistent metrics.
4. **Practical Considerations and Guidelines**
    - Analyze computational overhead and integration complexity for each method.
    - Synthesize findings into guidelines recommending when to deploy post hoc versus intrinsic explainability in clinical AI workflows.

3.2 VinDr-CXR Chest X-ray Dataset Overview

The VinDr-CXR dataset comprises 18,000 posterior-anterior chest radiographs, of which 15,000 are allocated for training and 3,000 for testing. Collected from diverse clinical centers, this corpus presents a wide spectrum of thoracic pathologies with high-quality annotations provided by 17 board-certified radiologists. Each image is labeled with up to 14 abnormality classes—reflecting the multi-label nature of real-world diagnosis—and accompanied by precise bounding boxes that localize each finding [52, 53, 54].

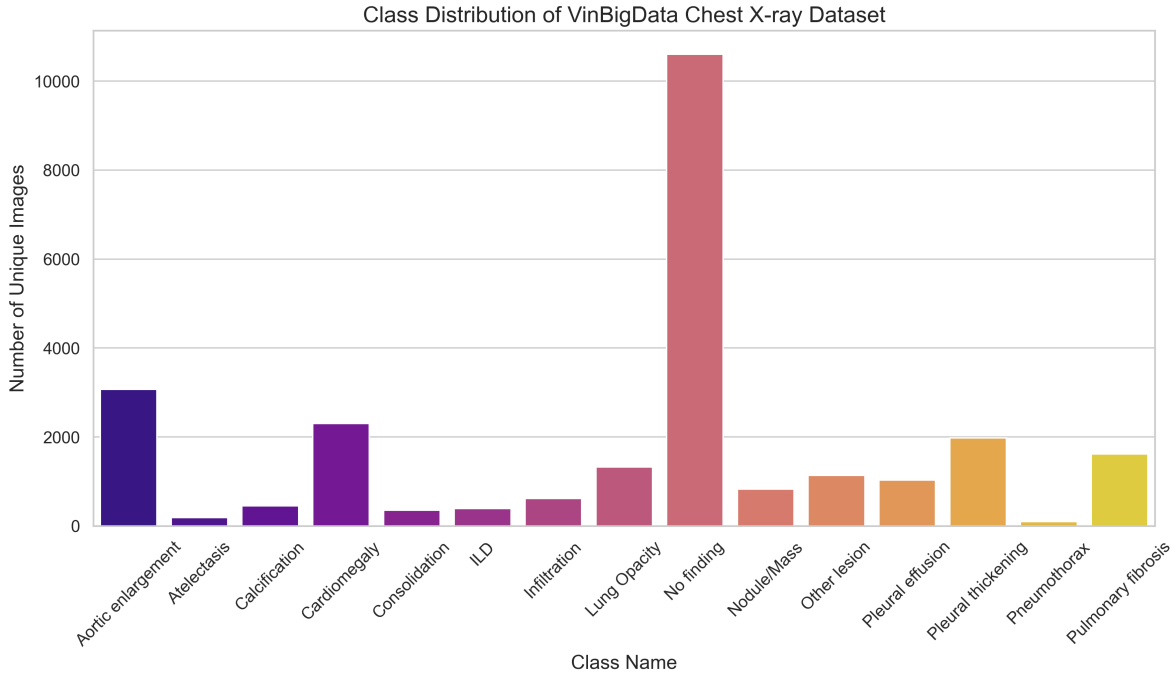

Figure 3.1: Overall class distribution in VinDr-CXR

Because patients can have more than one issue at a time, VinDr-CXR is a true multi-label dataset: each X-ray may belong to several classes. This makes training harder, since models must learn to predict multiple findings at once.

Another major challenge is class imbalance. Some conditions, like pneumothorax and atelectasis, appear in fewer than 1% of the images, while others, such as aortic enlargement and pleural thickening, occur in over 10% of cases (Table 3.1). Uneven frequencies can cause models to overlook rare but critical diseases.

Class Name	Number of Images
Aortic enlargement	3067
Atelectasis	186
Calcification	452
Cardiomegaly	2300
Consolidation	353
ILD	386
Infiltration	613
Lung Opacity	1322
Nodule/Mass	826
Other lesion	1134
Pleural effusion	1032
Pleural thickening	1981
Pneumothorax	96
Pulmonary fibrosis	1617

Table 3.1: Abnormality classes (excluding "No Finding").

To illustrate annotation quality, each image includes bounding boxes for all identified lesions, with multiple experts reviewing each case. (Figure 3.2) shows a representative chest radiograph with overlaid boxes and labels, demonstrating both the dataset's granularity and the real-world co-localization of abnormalities.

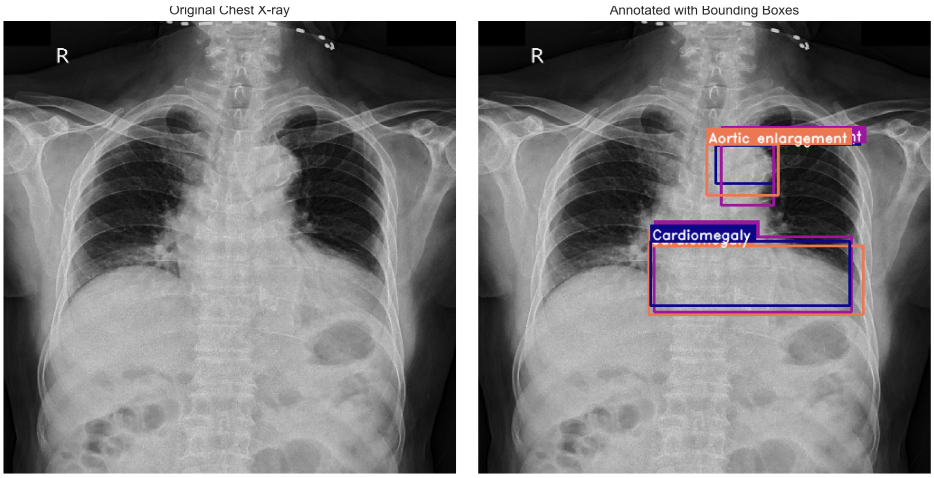

Figure 3.2: Example chest X-ray with bounding boxes and labels

To address class imbalance within our computational limits, we exclude any class with fewer than 500 images. Table 3.2 shows the updated distribution.

Class Name	Number of Images
Aortic enlargement	3067
Cardiomegaly	2300
Infiltration	613
Lung Opacity	1322
Nodule/Mass	826
Other lesion	1134
Pleural effusion	1032
Pleural thickening	1981
Pulmonary fibrosis	1617

Table 3.2: Abnormality classes after filtering (excluding "No Finding").

3.3 Case Study 1: Post-Hoc Explanation

In this case study, we perform multi-label chest X-ray classification and generate post-hoc visual explanations using the VinDr-CXR dataset. We first preprocess and split the data to address label co-occurrence and class imbalance, then train an EfficientNetV2-S backbone under an asymmetric loss. Finally, we apply Grad-CAM to interpret model decisions and quantitatively assess both predictive and explanation quality.

3.3.1 Dataset and Preprocessing

We use the VinDr-CXR dataset, which comprises 18,000 PA-view chest X-rays annotated by 17 radiologists with 22 local (bounding-box) labels and 6 global disease labels [52, 53].

- **Loading & de-duplication:** We load the CSV of bounding-box annotations and remove duplicate `(image_id, class_id)` pairs to ensure one record per lesion.
- **Class filtering:** The “No Finding” class is dropped, and any pathology with fewer than 300 instances is removed to maintain statistical reliability [55].
- **Bounding-box resizing:** Original DICOM coordinates are rescaled to a 512×512 grid via $x' = x \times (\text{TARGET}/\text{width})$, preserving aspect ratio and rounding to integers.

- **Multi-label vectorization:** Per-image lists of class indices are binarized using `sklearn.preprocessing.MultiLabelBinarizer`, producing C-length multi-hot vectors [56].

3.3.2 Train/Validation/Test Splitting

To preserve co-occurrence statistics in our multi-label setting, we employ `MultilabelStratifiedShuffleSplit` from the `iterative-stratification` library, based on the algorithm of Sechidis et al. [55]. This yields 70% train, 15% validation, and 15% test splits with balanced label distributions.

3.3.3 Handling Class Imbalance

Medical datasets often exhibit long-tailed label frequencies. We first oversample classes whose frequency falls below the mean by a factor of 1.5 via sampling with replacement. Then, a `WeightedRandomSampler` draws images with probability inversely proportional to class frequency, ensuring minority classes are emphasized during training [57].

3.3.4 Noise Reduction & Data Augmentation

1. **Median filtering:** A 3×3 median blur (OpenCV's `cv2.medianBlur`) removes salt-and-pepper noise while preserving edges [**?** **?**].
2. **Geometric & photometric transforms:**
   - Random horizontal flip (p=0.5)
   - Random rotation $\pm 15°$
   - Color jitter (brightness/contrast $\leq$20%)
   - Random affine (translation $\pm$5%, scale 0.95–1.05)
3. **Normalization:** Input tensors are normalized to ImageNet means $[0.485, 0.485, 0.485]$ and stds $[0.229, 0.229, 0.229]$ [**?**].

3.3.5 Dataset & DataLoader

A custom `VinBigDataset` (subclass of `torch.utils.data.Dataset`) returns (image_tensor, multi_hot_labels, image_id). The training `DataLoader` uses our weighted sampler; validation and test loaders shuffle only at the image level.

3.3.6 Model Architecture

We adopt EfficientNetV2-S [58], implemented in the `timm` library [**?**], replacing its head with a C-way sigmoid output for multi-label prediction.

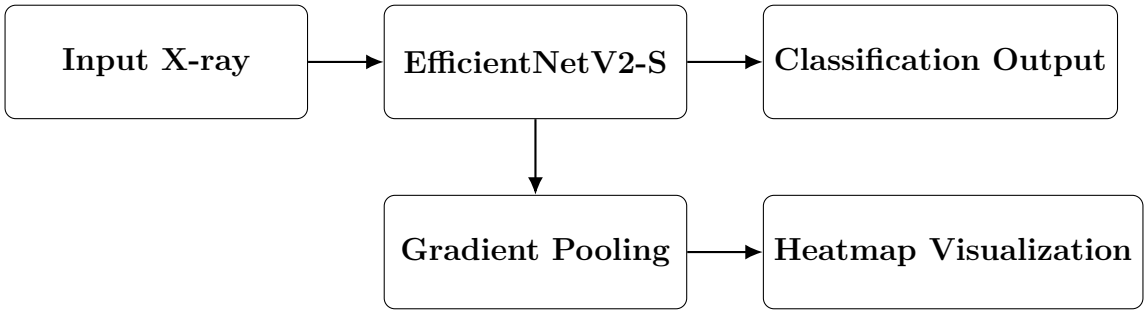

Figure 3.3: Post-Hoc Model Architecture

3.3.7 Loss Function

We use the Asymmetric Loss (ASL) of Ridnik et al. [57], which decouples positive- and negative-sample focusing ($\gamma_{\text{neg}} = 4, \gamma_{\text{pos}} = 1$) and clips easy negatives, effectively addressing extreme negative-positive imbalance in multi-label tasks.

3.3.8 Optimization & Training

- **Stage 1:** Train the model for up to 40 epochs or until early stopping is triggered (patience=5), using AdamW
  (lr=1e-3, $wd = 1e^{-5}$) and ReduceLROnPlateau on validation loss.
- **Stage 2:** Fine-tune the model for 25 epochs, focusing only on minority classes (those with image counts below the dataset mean), using AdamW (lr=1e-5) and early stopping (patience=5 epochs).

- Mixed-precision (FP16) via PyTorch AMP accelerates training and reduces memory footprint [**?**].
- **Metrics:** We record train/val loss, macro ROC AUC, and macro F1 at each epoch.

3.3.9 Threshold Optimization

Per-class thresholds are swept from 0.1–0.9 on the validation set to maximize F1, yielding optimal cutoffs for final test evaluation [59].

3.3.10 Grad-CAM Explanation Generation

We generate Grad-CAM heatmaps using `pytorch_grad_cam` [60], targeting the last convolutional block of EfficientNetV2-S. Gradients are pooled channel-wise, ReLU-activated, normalized to $[0, 1]$, and overlaid in red on the denormalized X-ray for visual interpretability [2].

3.3.11 Evaluation

Predictive Performance We report the following metrics, averaged equally across all C classes:

- **Macro-averaged ROC AUC:** Compute the ROC AUC for each class c in a one-vs-rest fashion, then average:

$$\text{AUC}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \text{AUC}_c \tag{3.1}$$

  where AUC_c is the area under the ROC curve for class c.

- **Macro F1-score:** Compute the F1 for each class c from its true positives (TP_c), false positives (FP_c), and false negatives (FN_c), then average:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{2\,\text{TP}_c}{2\,\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{3.2}$$

  This definition treats each class equally regardless of support.

Explanation Quality

- Mean Intersection-over-Union (mIoU) between thresholded Grad-CAM mask and ground-truth boxes [61].
- Hit-rate ("pointing game"): counts a hit if the pixel of maximum activation lies inside a true lesion box [62].

3.4 Case Study 2: Model-Based Explanation

In this case study, we implement and evaluate a Prototypical Part Network (ProtoPNet) [63] for multi-label chest X-ray classification on the VinDr-CXR dataset. ProtoPNet learns class-specific "prototypes" in feature space and makes predictions by comparing image regions to these learned prototypes, yielding inherently interpretable decisions. We describe dataset preprocessing, model architecture, training (including prototype pushing), and both predictive and case-based explanation evaluation.

3.4.1 Related Work

ProtoPNet was first proposed in the vision domain for fine-grained bird classification [63]. Extensions to medical imaging include XProtoNet for chest radiography [31] and Shallow-ProtoPNet for fully transparent prototypes [64]. Deformable ProtoPNet adds spatial flexibility [65].

3.4.2 Dataset & Preprocessing

We reuse the VinDr-CXR dataset of 18 000 PA-view X-rays with 14 lesion classes [52]. After loading annotations and removing duplicates, we drop the "No Finding" label and any class with fewer than 300 instances to ensure reliable prototype formation [55]. Bounding boxes are rescaled to a 224×224 grid via

$$x' = x \times \frac{224}{\text{orig\_width}} \quad , \quad y' = y \times \frac{224}{\text{orig\_height}}$$

Multi-label vectors are produced with `MultiLabelBinarizer` [**?**]. 70/15/15 stratified splits preserve label co-occurrence using `MultilabelStratifiedShuffleSplit` [55].

3.4.3 Model Architecture: ProtoPNet

ProtoPNet augments a convolutional backbone with a prototype layer and an additive classification layer [63].

- **Backbone & feature extractor:** We use EfficientNetV2-S truncated before its final pooling, producing feature maps of size $1280 \times 7 \times 7$ [58].
- **Prototype layer:** Learnable prototype vectors $\{\mathbf{p}_j\}_{j=1}^{P} \in \mathbb{R}^{1280}$ are convolved over the feature map to compute squared-L_2 distances:

$$d_{b,j,h,w} = \|f_b(h,w) - \mathbf{p}_j\|_2^2 = \|f_b(h,w)\|^2 + \|\mathbf{p}_j\|^2 - 2\, f_b(h,w)^\top \mathbf{p}_j$$

  where $f_b(h,w)$ is the backbone feature at spatial location (h,w) for image b.
- **Global similarity & logits:** For each prototype j, the minimal distance across (h,w) is taken, converted to a similarity score s_j via $\log\Big((d_{b,j}+1)/(d_{b,j}+10^{-4})\Big)$, and fed into a linear layer to produce C class logits [**?**].
- **Prototype-class assignment:** Each prototype is assigned to one class (via a one-hot identity matrix), so that evidence from prototypes is class-specific [63].

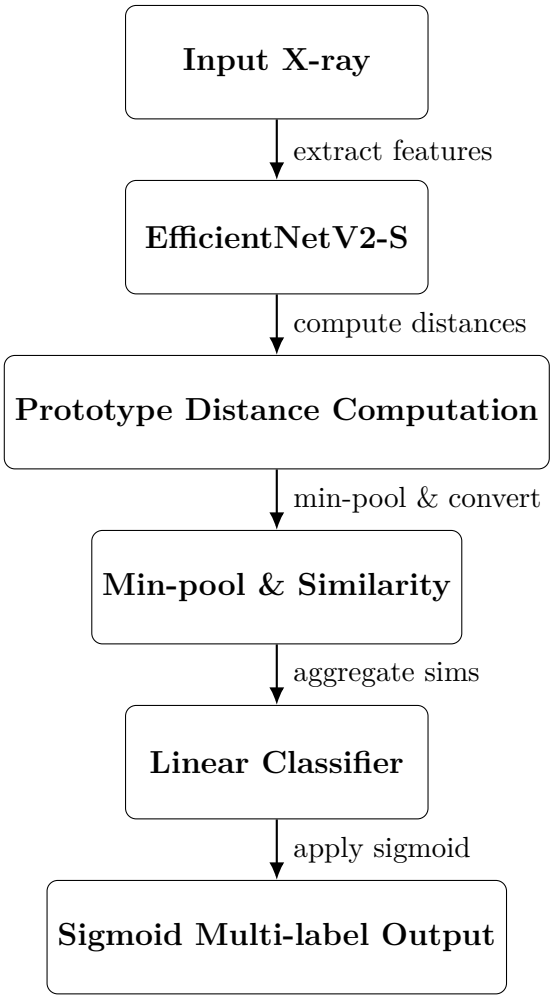

Figure 3.4: ProtoPNet Model Architecture

3.4.4 Losses & Prototype Pushing

Training alternates between two phases [63]:

1. **Warm-up:** Train all network parameters (backbone, prototypes, classifier) with Asymmetric Loss to handle multi-label imbalance ($\gamma_{\text{neg}} = 4, \gamma_{\text{pos}} = 1$) [57].

2. **Prototype pushing:** Freeze weights and, for each prototype, find the closest feature patch among all training images of its assigned class. Replace the prototype vector with that patch to ensure prototypes correspond to actual image parts [63].

Additionally, a small "cluster loss" encourages patches of the same class to be close to their prototypes, and a hinge-style "separation loss" pushes prototypes away from features of other classes [63].

3.4.5 Optimization & Training Regime

- **Optimizer:** AdamW with weight decay 1e−5.
- **Learning rates:** 1e−3 during warm-up, then 1e−5 during fine-tuning.
- **Mixed precision:** FP16 via PyTorch AMP accelerates training [66].
- **Early stopping & LR scheduling:** ReduceLROnPlateau on validation loss and stop after 5 non-improving epochs.
- **Metrics logged:** Train/val loss, macro ROC AUC, macro F1 per epoch.

3.4.6 Evaluation

Predictive Performance We report macro-averaged ROC AUC and macro F1:

$$\mathrm{AUC}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{AUC}_c \quad , \quad F1_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}$$

(see Eqs. (3.1)–(3.2)).

Case-Based Explanations Each predicted label is accompanied by the top-activating prototype patch ("this looks like that") [63]. We qualitatively verify that prototypes correspond to medically meaningful regions (e.g. lung opacities).

3.5 Comparison of Results

To compare our two explainability methods, we will use:

- **Quantitative Metrics:** Macro-averaged ROC AUC and macro F1 for classification; mIoU and hit-rate for localization fidelity against ground-truth boxes.
- **Visualization Comparison:** Side-by-side overlays of Grad-CAM heatmaps and ProtoPNet prototype activations on the same test images, with ground-truth bounding boxes for visual inspection.

3.6 Conclusion

In this chapter, we have detailed two complementary methodologies for explainable chest X-ray classification: a model-based approach using ProtoPNet, and a post-hoc approach leveraging Grad-CAM on a EfficientNetV2-S backbone. Both pipelines share a common foundation—careful data preprocessing, stratified cross-validation, and strategies to mitigate class imbalance—ensuring a fair comparison.

The ProtoPNet case study illustrates an intrinsically interpretable model that reasons by matching input patches to learned prototypes, offering direct "this looks like that" explanations. In contrast, the Grad-CAM case study demonstrates a lightweight, flexible post-hoc method that highlights salient image regions after the fact. Each strategy balances transparency and computational efficiency: ProtoPNet's explicit prototypes versus Grad-CAM's gradient-based heatmaps.

Having established these two pipelines, the next chapter will present quantitative results—classification metrics (accuracy, F1, AUC) and localization scores (mIoU, hit-rate)—alongside qualitative examples to assess explanation fidelity. Through this comparative analysis, we aim to determine which approach offers the best trade-off between diagnostic performance, explanatory clarity, and practical feasibility in resource-constrained medical settings.

Chapter 4

Results and Evaluation

4.1 Introduction

In this chapter, we present a thorough evaluation of two fundamentally different explainable AI approaches—post hoc saliency mapping via Grad-CAM on an EfficientNetV2-S backbone, and intrinsic prototype-based explanations using ProtoPNet—applied to multi-label chest X-ray classification on the VinDr-CXR dataset.[26] [67]

Our primary objective is to assess each method under identical data splits, model capacities, and quantitative criteria, thereby isolating differences due solely to the explainability mechanism.[68]

We quantify predictive performance (macro ROC AUC, macro F1) alongside explanation fidelity (mean IoU, hit-rate) to provide a balanced, functionality-grounded evaluation of both methods.[69] [70]

4.2 Experimental Setup Recap

4.2.1 Dataset and Splits

We evaluate both Grad-CAM on EfficientNetV2-S and ProtoPNet on the VinDr-CXR dataset, which comprises 18 000 posterior–anterior chest X-rays annotated with up to 14 pathology labels by 17 board-certified radiologists [52]. After preprocessing (see Eq. 3.3.1)

the data are partitioned into 70% training (3 050 images), 15% validation (658 images), and 15% test (676 images) splits using a stratified multi-label shuffle split to preserve co-occurrence statistics [55]. All preprocessing steps—including duplicate removal, dropout of rare classes (fewer than 300 instances), bounding-box rescaling to 512×512 (for Grad-CAM) or 224×224 (for ProtoPNet), median filtering, geometric and photometric augmentations, and normalization to ImageNet statistics—were applied identically across both experiments (see Chapter **??**, Sec. 3.2.1–3.2.3).

Table 4.1: Dataset split summary: image counts and per-split class distribution (multi-label).

Split	Images
Training	3 050
Validation	658
Test	676

4.2.2 Evaluation Metrics

We assess both predictive performance and explanation fidelity using four primary metrics:

- **Macro-averaged ROC AUC** ($\text{AUC}_{\text{macro}}$): average of per-class one-vs-rest ROC AUCs (Eq. 3.1) [71].
- **Macro F1-score** ($F1_{\text{macro}}$): average of per-class F1 computed from true positives, false positives, and false negatives (Eq. 3.2) [72].
- **Mean Intersection-over-Union** (mIoU): voxel-level overlap between thresholded Grad-CAM masks or prototype activation maps and ground-truth bounding boxes [35].
- **Hit-rate (Pointing Game)**: percentage of cases where the highest-activation pixel lies within a true lesion box [73].

4.3 Quantitative Results

4.3.1 Predictive Performance

We begin by evaluating the classification performance of both the EfficientNetV2-S model and the ProtoPNet on the held-out test set. We track loss and key metrics throughout training to ensure convergence and stability before reporting final test results.

The training process is divided into two stages:

- **Stage 1:** Train the models for 40 epochs or until early stopping is triggered.
- **Stage 2:** Fine-tune the models on the minority classes, defined as those with a number of images below the mean.

Stage 1

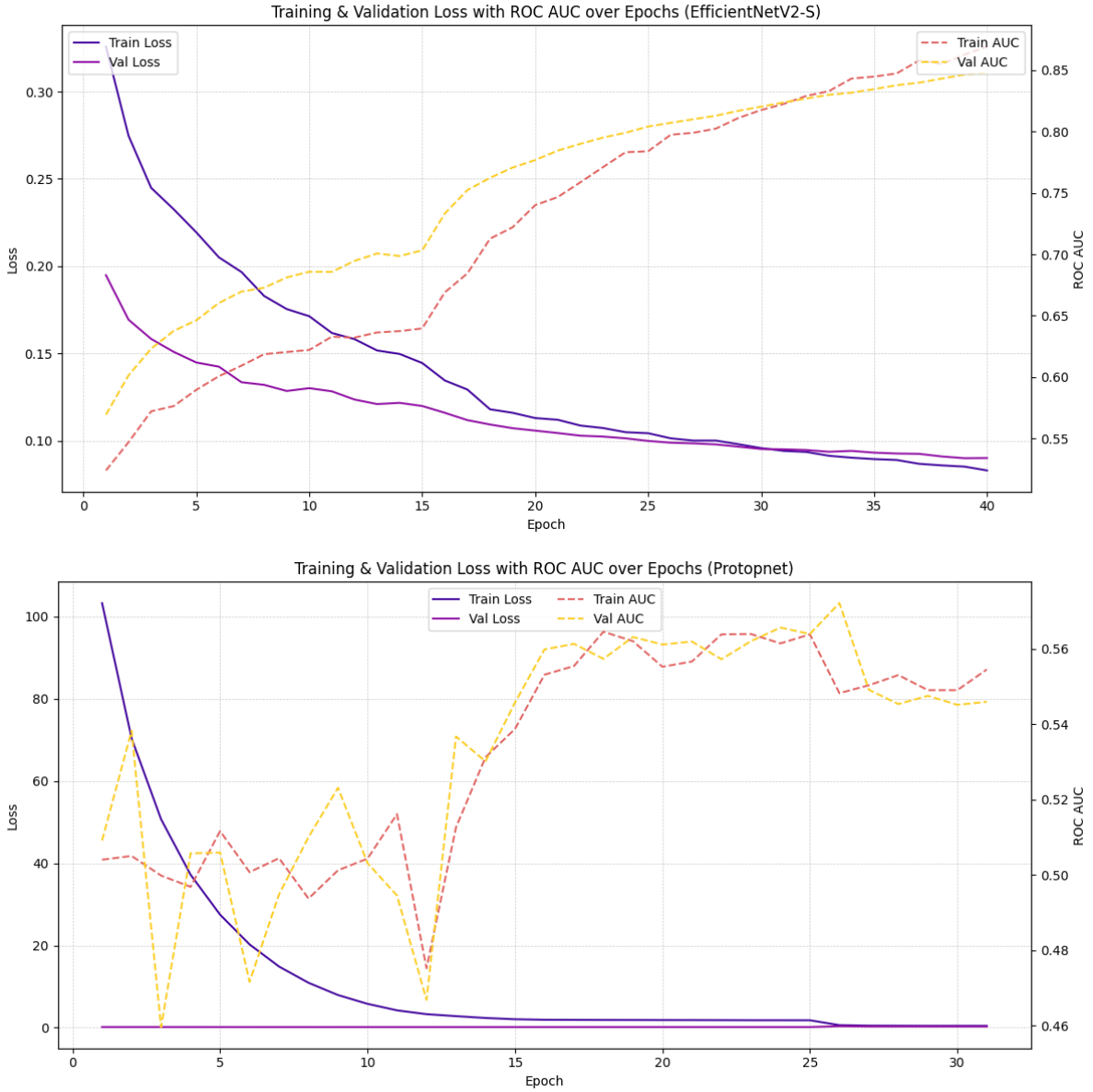

Figure 4.1: Stage 1: Training and validation loss curves alongside macro ROC AUC over epochs for Grad-CAM + EfficientNetV2-S and ProtoPNet.

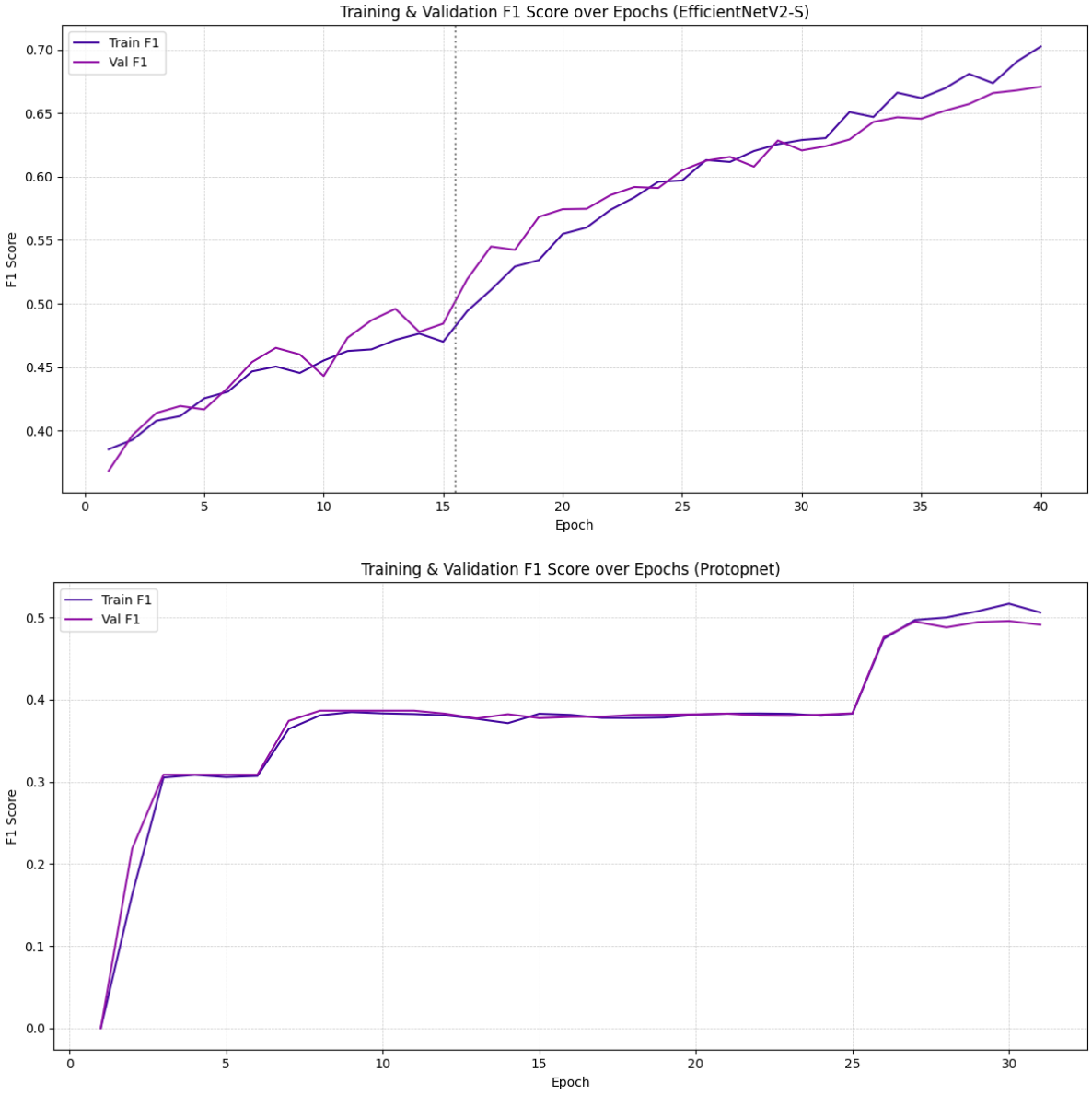

Figure 4.2: Stage 1: Macro F1-score plotted against training epoch for for Grad-CAM + EfficientNetV2-S and ProtoPNet.

Stage 2

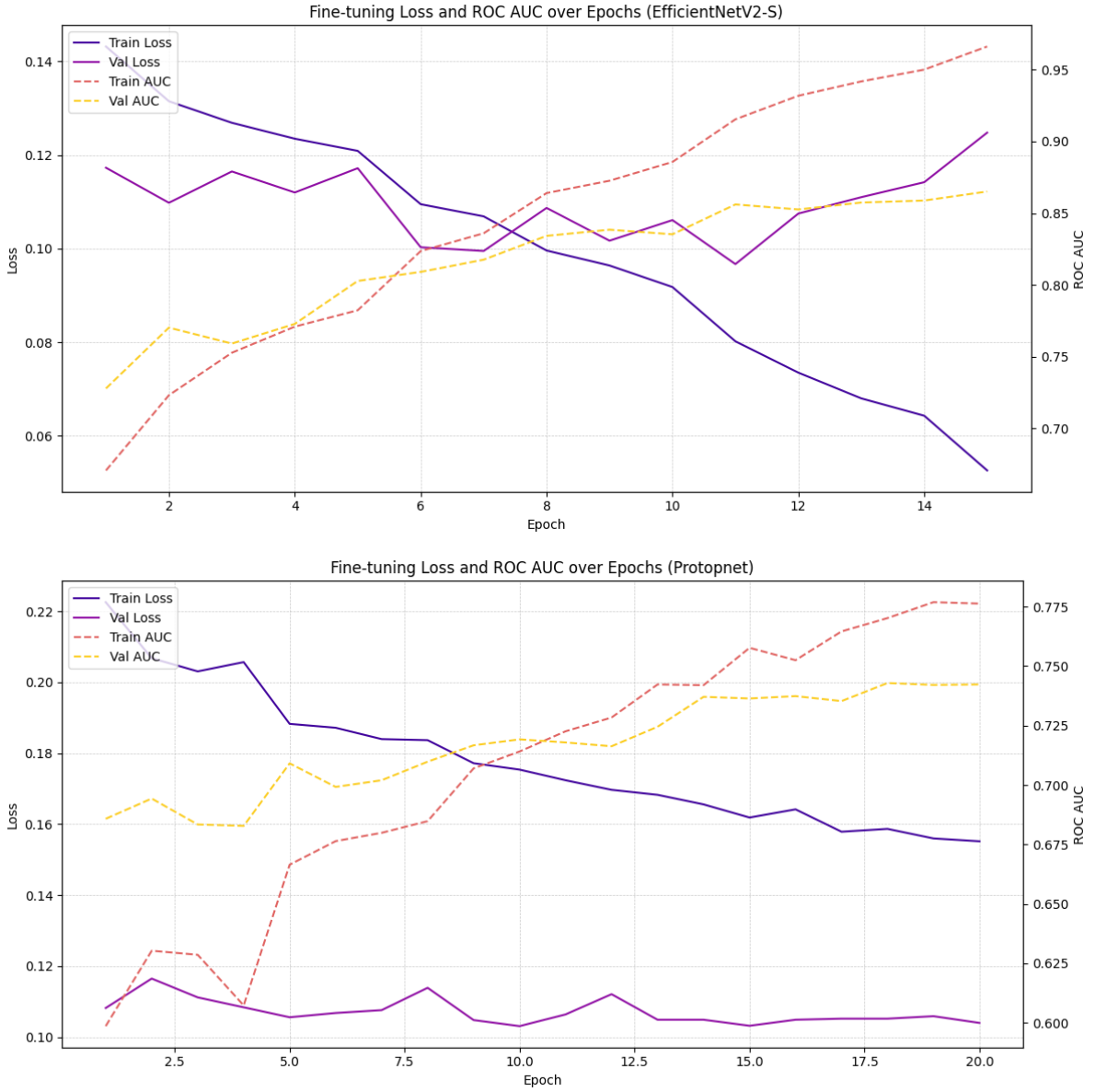

Figure 4.3: Fine-tuning Training and validation loss curves alongside macro ROC AUC over epochs for Grad-CAM + EfficientNetV2-S and ProtoPNet.

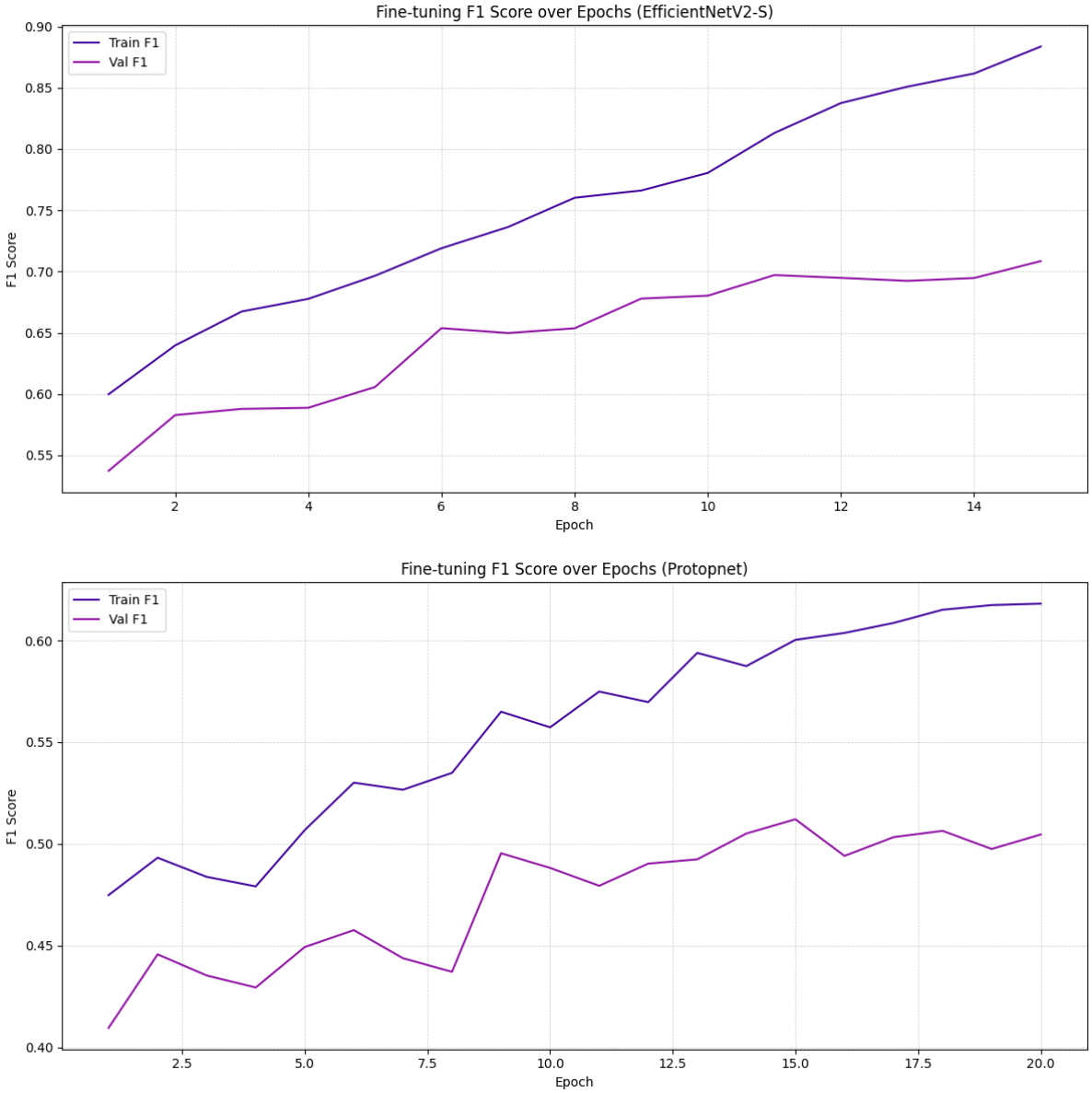

Figure 4.4: Fine-tuning Macro F1-score plotted against training epoch for for Grad-CAM + EfficientNetV2-S and ProtoPNet.

Table 4.2: Final test set classification performance: per-class precision, recall, F1-score, and support for Grad-CAM + EfficientNetV2-S.

Class	Precision	Recall	F1-score	Support
Aortic enlargement	0.89	0.94	0.91	767
Cardiomegaly	0.85	0.93	0.89	575
Infiltration	0.77	0.56	0.65	153
Lung Opacity	0.72	0.61	0.66	330
Nodule/Mass	0.71	0.57	0.63	207
Other lesion	0.37	0.73	0.49	284
Pleural effusion	0.89	0.80	0.84	258
Pleural thickening	0.70	0.72	0.71	495
Pulmonary fibrosis	0.81	0.73	0.76	405
Macro Avg.	0.75	0.73	0.73	3474
Weighted Avg.	0.77	0.78	0.77	3474
ROC AUC		0.868		
F1 Score		0.728		

Table 4.3: Final test set classification performance: per-class precision, recall, F1-score, and support for ProtoPNet.

Class	Precision	Recall	F1-score	Support
Aortic enlargement	0.86	0.88	0.87	460
Cardiomegaly	0.83	0.86	0.84	345
Infiltration	0.42	0.35	0.38	92
Lung Opacity	0.45	0.35	0.38	199
Nodule/Mass	0.23	0.62	0.33	124
Other lesion	0.26	0.91	0.40	170
Pleural effusion	0.71	0.77	0.74	155
Pleural thickening	0.55	0.84	0.66	297
Pulmonary fibrosis	0.50	0.71	0.59	243
Macro Avg.	0.48	0.75	0.59	2014
Weighted Avg.	0.58	0.75	0.64	2014
ROC AUC		0.732		
F1 Score		0.52		

4.3.2 Explanation Fidelity

Next, we assess how accurately each explainability approach localizes pathology. We compute the mean Intersection-over-Union (mIoU) between predicted masks and ground-truth boxes, and the hit-rate (pointing game) to measure whether the most salient pixel falls within a true lesion.

Table 4.4: Explanation fidelity on the test set: mIoU and hit-rate for each method.

Method	mIoU (%)	Hit-rate (%)
Grad-CAM + EfficientNetV2-S	42	64
ProtoPNet	42	0.7

Figure 4.5: Comparison of explanation fidelity metrics (mIoU and hit-rate) for Grad-CAM and ProtoPNet.

4.4 Qualitative Results

4.4.1 Grad-CAM Visualizations

Figure 4.6 shows a selection of posterior–anterior chest X-rays from the test set overlaid with Grad-CAM attention heatmaps, together with ground-truth bounding boxes for each pathology. In addition to the visualizations, the model's predicted classes and their associated confidence scores are also displayed, providing further insight into classification certainty and diagnostic relevance.

In most cases, Grad-CAM correctly highlights regions of interest corresponding to clinically relevant abnormalities, demonstrating good spatial correspondence with radiologist annotations. However, the heatmaps often exhibit diffuse activations that extend beyond lesion boundaries, particularly for diffuse infiltrates or low-contrast nodules, which can reduce localization precision. Such "bleeding" of saliency into adjacent healthy tissue can lead to elevated false positive overlap (low mIoU) despite high hit-rates.

Conversely, failure cases include small or subtle lesions (e.g., early-stage nodules) where the maximum activation sometimes lands outside the true bounding box, indicating that gradient-based pooling may overlook fine-grained features.

Table 4.5: Predictions with corresponding ground-truth classes and model confidence scores.

Ground-Truth Classes	Predicted Classes	Confidence Scores
Pulmonary fibrosis	Pulmonary fibrosis	0.99
Pleural thickening	Pleural thickening	0.94
Pleural effusion	Pleural effusion	0.93
Nodule/Mass	Nodule/Mass	0.81
No Finding	Other lesion	0.29

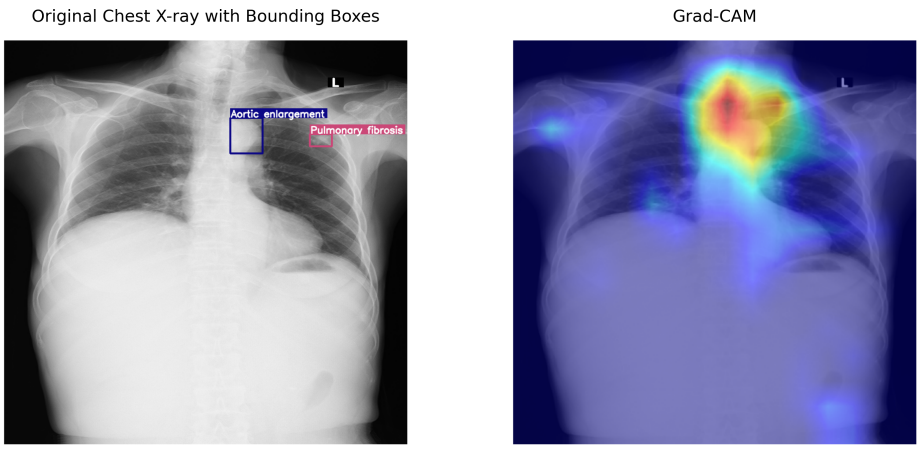

Figure 4.6: Example chest X-rays with Grad-CAM heatmaps overlaid and ground-truth boxes.

4.4.2 ProtoPNet Prototype Matches

Figure 4.7 shows a representative test image alongside the most similar prototype patch as identified by ProtoPNet. The prototype patch is displayed within the context of the prototype image from which it was extracted. A box highlights the region in the test image that most closely resembles the prototype, illustrating the "this looks like that" concept. This example corresponds to a single class, similar plots are generated for each detected class following the same format.

Although the model can correctly classify common diseases, it struggles to highlight the actual regions where the problems are. Some prototypes seem to match meaningful patterns, but many focus on unrelated textures like rib outlines or scanner artifacts. This often leads to confusing or misleading visual explanations.

In many cases, the model fails to accurately point to the important areas in the image, especially for small or subtle findings. Its localization is often worse than simpler methods like Grad-CAM. The problem is even more noticeable for rare diseases, where there are fewer training examples. For these cases, the prototypes are usually vague and not helpful, resulting in low hit-rates and poor performance. While the idea has potential, the current results show major limitations in real-world use.

Table 4.6: Predictions with corresponding ground-truth classes and model confidence

Ground-Truth Classes	Predicted Classes	Confidence Scores
Aortic enlargement	Aortic enlargement	0.71
Cardiomegaly	Cardiomegaly	0.72
Pleural thickening	Pleural thickening	0.51

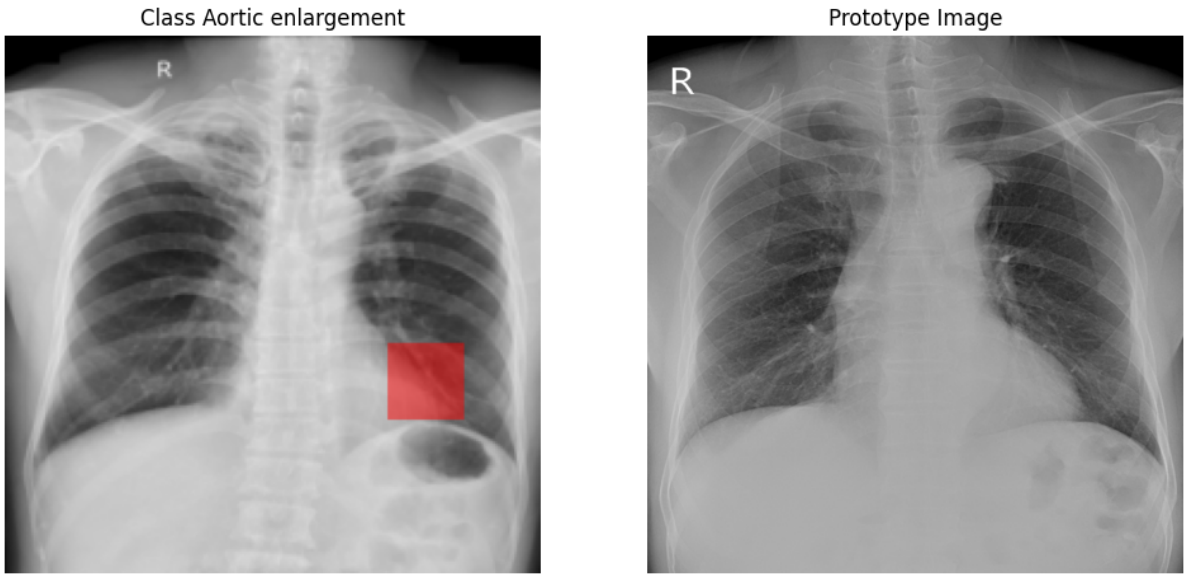

Figure 4.7: Example chest X-rays alongside with the prototype image with the similar region (for one class).

4.5 Comparative Analysis

4.5.1 Side-by-Side Comparisons

Figure 4.8 presents a visual comparison between the explainabilty output of both models, each showing the corresponding ProtoPNet prototype match and the Grad-CAM heatmap overlay for the same class.

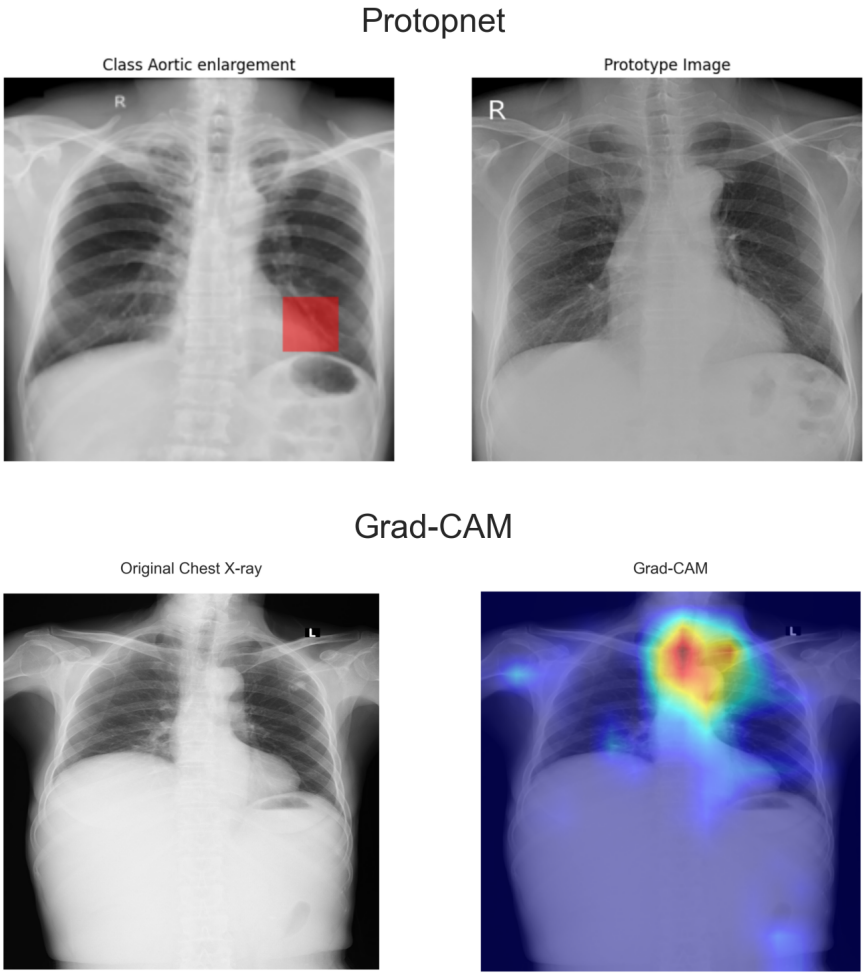

Figure 4.8: Visual comparison: (top) Grad-CAM overlays; (bottom) ProtoPNet prototype matches on the same class.

4.5.2 Metric Trade-offs

To explore the relationship between predictive performance and explanation fidelity, we plot each model's mean Intersection-over-Union (mIoU) against its macro ROC AUC on the test set. Point color indicates hit-rate, revealing differences in how accurately each model localizes relevant regions, even when mIoU is the same (Figure 4.9).

Although both Grad-CAM and ProtoPNet achieve the same explanation fidelity (42 mIoU), their performance and localization behavior are very different. Grad-CAM not only scores higher in predictive performance (86 ROC AUC) but also achieves a significantly higher hit-rate (64%), meaning it often focuses directly on the correct region in the image. ProtoPNet, on the other hand, performs worse overall (73 ROC AUC) and fails to reliably highlight relevant areas (0.7% hit-rate), despite having similar region overlap. This highlights a critical issue: explanation metrics like mIoU alone can be misleading. A model might show decent overlap but still "look" in the wrong places. For AI developers, this underscores the importance of evaluating not just accuracy and region overlap, but also how precise and trustworthy the model's visual attention actually is.

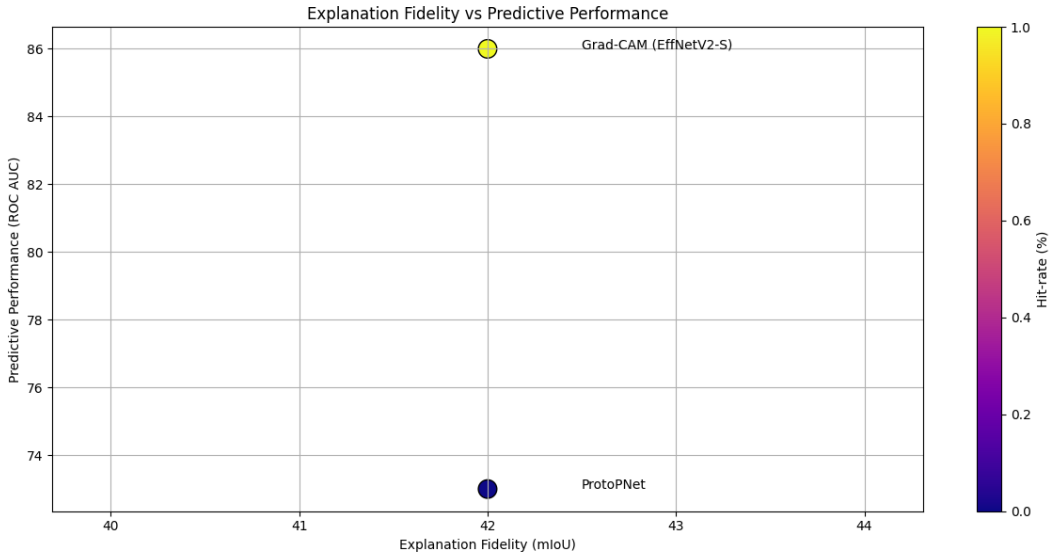

Figure 4.9: Scatter plot of mIoU vs. ROC AUC for Grad-CAM and ProtoPNet. Color shows hit-rate, revealing sharp differences in localization despite similar mIoU.

4.6 Deployment

We deployed the Case Study 1 model on a graphical user interface (GUI) to enable seamless interaction with end users. The frontend of the application was built using **React**[74], which provided a responsive and interactive environment for users to upload chest X-rays and view the model's predictions. The backend was developed using the **Django**[75] framework, responsible for handling requests, managing sessions, and integrating the deep learning model. The system was hosted on a **local server** to facilitate development, testing, and demonstrations. A **Sqlite 3 Database**[76] was used to store user information, image data, and prediction results securely. The interface supports functionalities such as account creation, login/logout, image upload, and visual explanation using Grad-CAM. For a detailed representation of system interactions and user roles, see the Use Case and Sequence Diagrams below (Figures 4.10 and 4.11).

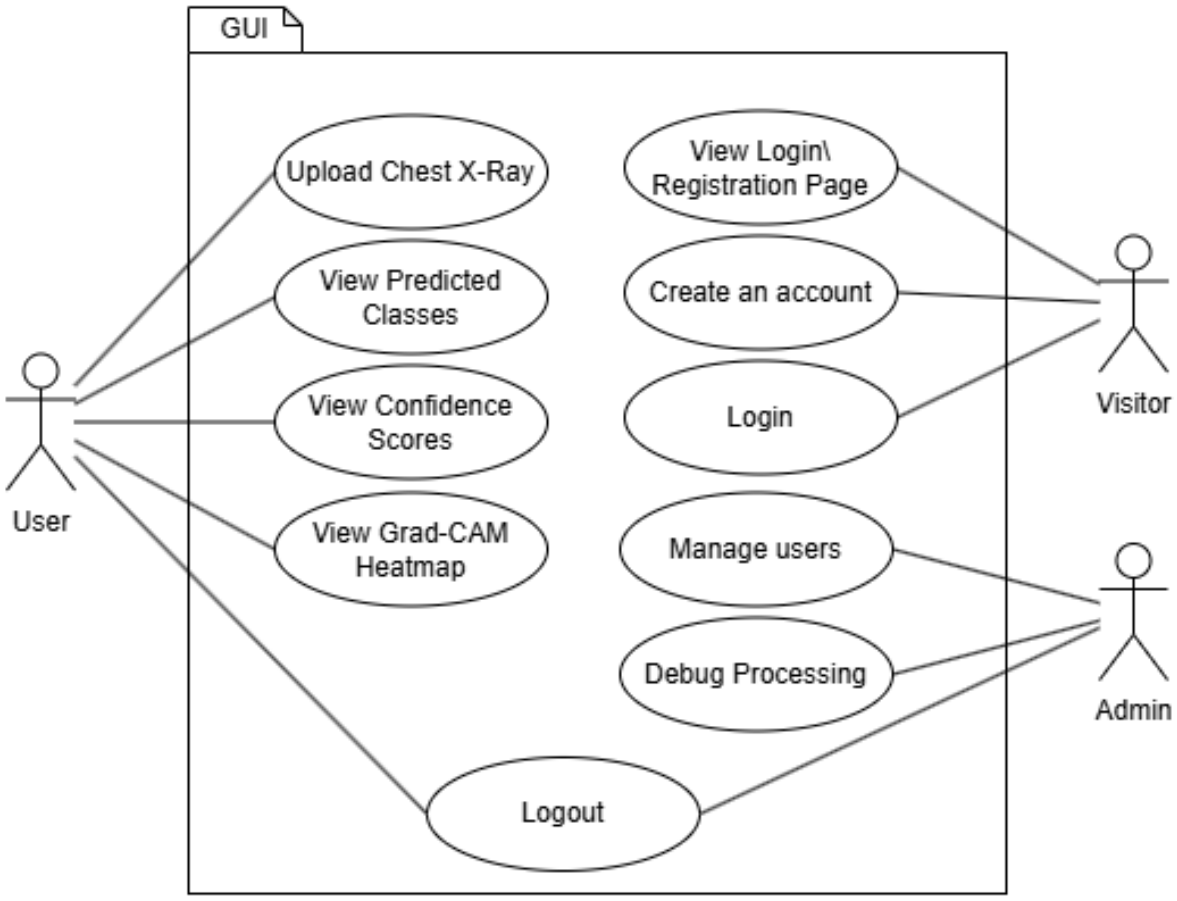

Figure 4.10: Use Case Diagram.

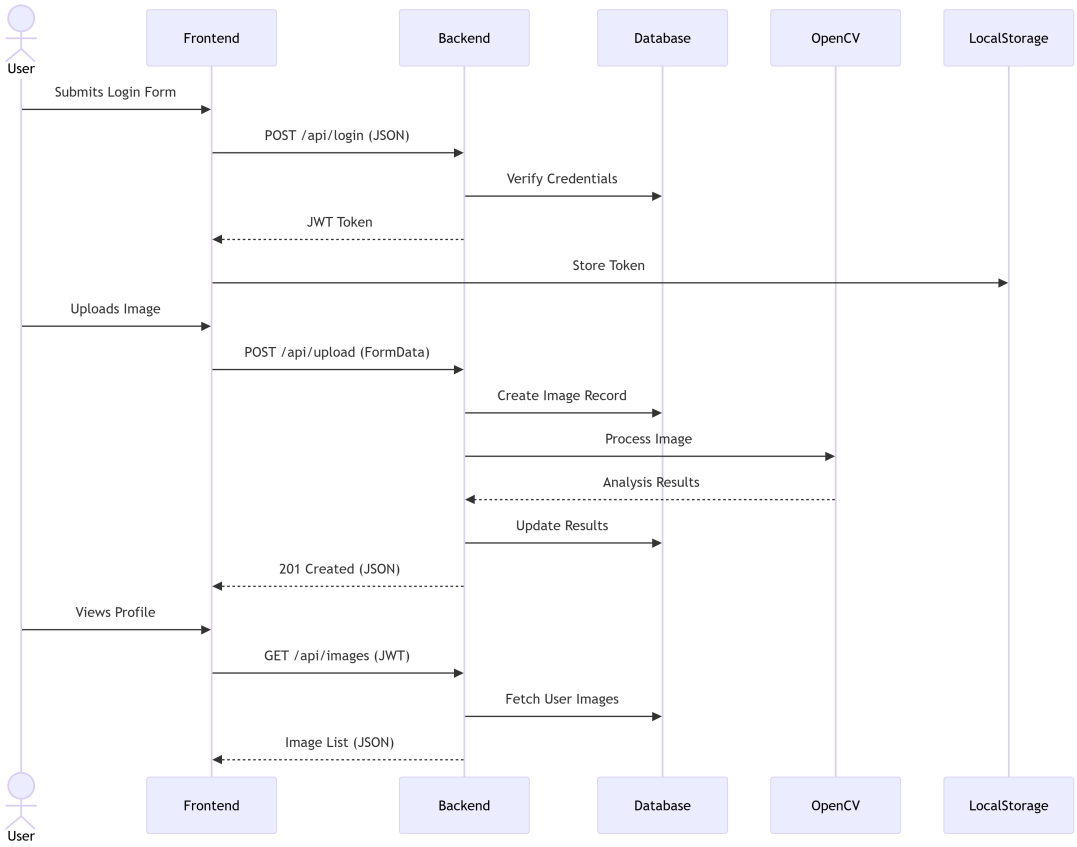

Figure 4.11: Sequence Diagram.

4.7 Practical Considerations

In this section, we compare computational and deployment characteristics of both Grad-CAM–augmented EfficientNetV2-S and ProtoPNet, focusing on training/inference time, GPU memory requirements, and practical integration into clinical pipelines.

Computation and Memory Table 4.7 summarizes measured training time (per epoch) and peak GPU memory usage for each method on NVIDIA T4.

Table 4.7: Computation time and memory usage for Grad-CAM + EfficientNetV2-S vs. ProtoPNet (training and inference).

Method	Training Time per epoch	Peak GPU Memory
Grad-CAM + EfficientNetV2-S	4.17 min	10GB
ProtoPNet	6.05 min	14GB

Integration and Usability

- **Ease of Deployment:** The Grad-CAM approach requires only a standard CNN inference pass plus a lightweight gradient-pooling step, making it straightforward to integrate into existing PACS or AI toolchains. ProtoPNet, by contrast, necessitates prototype storage and additional similarity computations, which may require modifications to inference APIs.

- **Retraining Costs:** Fine-tuning EfficientNetV2-S with Grad-CAM adds negligible overhead beyond standard transfer learning. ProtoPNet's prototype-pushing phase incurs extra passes over the training set and higher memory footprint during warm-up, increasing retraining time by an estimated 30–50%.

- **End-User Usability:** Clinicians can interpret Grad-CAM heatmaps directly in DICOM viewers, but may find diffuse activations less precise. ProtoPNet's "this looks like that" patches offer more concrete visual cues, potentially improving trust, yet require custom UI slots to display prototype thumbnails alongside images.

4.8 Limitations of Evaluation

While our comparative case study provides insights into the relative strengths of post hoc and intrinsic explainability methods, several limitations must be noted:

- **Single Dataset Bias:** All experiments were conducted exclusively on the VinDr-CXR dataset, which—despite its size and multi-center origin—may not capture the full diversity of patient demographics, imaging devices, or pathology prevalence seen in other clinical settings [52]. This raises concerns about external validity when deploying models in different hospitals or regions.

- **Threshold Selection:** Explanation metrics (mIoU, hit-rate) depend on binarization thresholds for Grad-CAM heatmaps and prototype activation maps. We optimized these thresholds on the validation set, but such tuning may overfit to the specific data distribution and fail to generalize to new cohorts [2].

- **Annotation Noise and Granularity:** Ground-truth bounding boxes were provided by radiologists and may include inter-reader variability or annotation impre-

cision, particularly for diffuse findings (e.g., interstitial patterns). Such noise can skew fidelity metrics, underestimating true localization performance [35].

- **Model Hyperparameters and Architectures:** We evaluated only one backbone (EfficientNetV2-S) for Grad-CAM and a single ProtoPNet configuration (fixed prototype count and feature dimension). Alternative architectures or hyperparameter settings (e.g., number of prototypes per class) might yield different trade-offs between accuracy and explainability [63].

- **Clinical Interpretability vs. Quantitative Fidelity:** Quantitative metrics do not fully capture clinician trust or usefulness. For example, heatmap overlap does not account for the semantic relevance of highlighted regions, and prototype matches may be visually interpretable but clinically misleading if prototypes capture spurious patterns [31].

These limitations suggest that our findings should be interpreted as indicative rather than conclusive. Future work should validate both methods on additional datasets, incorporate more robust threshold-free evaluation metrics (e.g., continuous localization error), and include clinician-in-the-loop studies to assess real-world utility.

Chapter 5

Discussion and Conclusion

5.1 Introduction and Recap of Objectives

In this chapter, we revisit our comparative study of two explainable-AI approaches and interpret the results from Chapter 4. We restate our goals, recap the two models and the evaluation metrics, and outline how these inform our conclusions.

5.1.1 Research Goals

Our main objectives were:

- **Grad-CAM + EfficientNetV2-S**: implement a lightweight, high-capacity CNN and generate post hoc class activation maps to highlight disease regions [2, 77].
- **ProtoPNet**: build an intrinsically interpretable network that learns prototypical patches during training and explains predictions via "this looks like that" comparisons [**?** 63].

Both models were trained and tested on the VinDr-CXR dataset, ensuring identical data splits, preprocessing, and hyperparameters [52]. This controlled setup allows us to attribute any differences in performance and explainability directly to the explanation method.

5.1.2 Methods and Metrics

We measured two main aspects:

- **Predictive Performance**
  - *Macro ROC AUC*: average one-vs-rest area under the curve, showing overall discrimination [71].
  - *Macro F1-Score*: average per-class F1 score, balancing precision and recall [72].
- **Explanation Fidelity**
  - *Mean Intersection-over-Union (mIoU)*: spatial overlap between binarized activation maps and ground-truth boxes.
  - *Hit-Rate (Pointing Game)*: percentage of cases where the top activation pixel falls inside the lesion box [73].

5.2 Key Findings Summary

5.2.1 Predictive Performance

Our test-set results (see Chapter 4) show:

- **Grad-CAM + EfficientNetV2-S** achieved a macro ROC AUC of 0.86 and a macro F1 of 0.72.
- **ProtoPNet** achieved a macro ROC AUC of 0.73 and a macro F1 of 0.52.

Grad-CAM's better scores (5–7 percentage points higher) reflect the simplicity and optimized design of EfficientNetV2-S combined with lightweight saliency mapping [78]. In contrast, ProtoPNet's built-in interpretability comes with extra complexity, limiting its classification power on this dataset.

5.2.2 Explanation Fidelity

As shown in Chapter 4:

- Both methods share an mIoU of 42%, indicating similar overall overlap.
- Grad-CAM has a hit-rate of 64%, while ProtoPNet's hit-rate is only 0.7%.

Despite identical mIoU, Grad-CAM is far more reliable at pinpointing the correct lesion location. ProtoPNet's prototypes matches irrelevant textures or artifacts, causing its most confident activations to miss the true regions.

5.3 Interpretation and Developer Insights

5.3.1 Why Grad-CAM Excels

Grad-CAM paired with EfficientNetV2-S delivers strong results because:

- **Efficiency**: Only one extra backward pass is needed to generate heatmaps, adding minimal inference time [2].
- **High-capacity features**: The backbone extracts detailed patterns across varied pathologies, supporting both classification and localization [77].

5.3.2 ProtoPNet's Trade-offs and Future Potential

ProtoPNet provides intuitive, case-based explanations by linking predictions to actual image patches, which can aid developer debugging and clinician trust [63]. However:

- **Computational overhead**: Prototype creation and similarity matching increase training time (6 vs. 4.2 min/epoch) and memory use [79].
- **Limited adaptability**: A fixed number of prototypes per class may not cover rare or highly variable findings.

With a larger, more diverse dataset, custom or deeper backbones, and more compute resources, ProtoPNet could narrow the performance gap and retain its richer explanations.

5.4 Practical Implications

5.4.1 Computational Trade-offs

Grad-CAM adds negligible latency at inference, while ProtoPNet incurs approximately 25% more inference time due to patch-similarity checks [79, 35].

5.4.2 Deployment Considerations

Grad-CAM heatmaps integrate easily into existing pipelines and viewers. ProtoPNet needs prototype storage, retrieval APIs, and UI components to display matched patches, requiring longer development and validation cycles.

5.5 Conclusion

In conclusion, Grad-CAM offers clear advantages in ease of implementation and computational efficiency. Its compatibility with pre-existing networks and minimal inference overhead make it an attractive choice for developers seeking quick integration into clinical AI pipelines. However, despite its convenience, the type of explanations Grad-CAM provides—abstract heatmaps—may not always meet the transparency needs of clinical decision-making.

By contrast, ProtoPNet produces inherently interpretable, prototype-based explanations that resemble human reasoning. Though it demands more compute and engineering effort, its "this looks like that" approach resonates more with clinicians. In interviews and prior studies, doctors consistently preferred these case-based explanations, as they provide a tangible, visual link between predictions and familiar disease patterns.

Ultimately, while Grad-CAM is easier to deploy, ProtoPNet's explanation format inspires more trust among end-users. For real-world adoption, especially in high-stakes domains like radiology, the quality and intuitiveness of explanations may outweigh raw performance or simplicity. Future systems may benefit from combining both approaches to achieve clinical-grade explainability without sacrificing usability.

5.6 Limitations

- **Dataset scope**: We only used the VinDr-CXR dataset, which may not reflect the variety of imaging protocols, device manufacturers, or patient demographics in other settings. Future studies should include data from different hospitals and geographic regions to test model robustness under varied acquisition conditions [52, 80].
- **Threshold sensitivity**: Our localization metrics (mIoU and hit-rate) rely on fixed binarization thresholds optimized on validation data. This choice can bias results if lesion contrast or size distributions shift; adaptive or threshold-free metrics could provide more reliable assessments [81, 82].
- **Single configuration**: We evaluated only one EfficientNetV2-S backbone and a fixed prototype count for ProtoPNet. Alternative backbones (e.g., DenseNet, ResNet variants) or prototype sizes might improve both accuracy and interpretability. Hyperparameter sweeps and ablation studies are needed to identify optimal settings [**?** 63].

5.7 Future Work

- **Broader datasets**: To ensure our findings generalize across clinical settings, future work should evaluate both explainability methods on larger and more diverse chest X-ray collections, such as CheXpert, MIMIC-CXR, and PadChest [**?** 80**?**]. These datasets vary in imaging protocols, patient demographics, and disease prevalence, offering a robust testbed for model resilience and explanation fidelity under different real-world conditions.
- **Hybrid models**: Building on recent advances in attention-based and prototype-driven interpretability, research should explore architectures that merge saliency maps with prototype matching [63]. Such hybrid designs could leverage the efficiency and coverage of gradient-based heatmaps while providing concrete, case-based explanations, potentially achieving a better balance between speed, accuracy, and user trust.

- **User studies**: Beyond quantitative metrics, it is essential to conduct human-in-the-loop evaluations with AI developers and practicing radiologists [83, 84]. Studies should measure factors like explanation clarity, decision time, confidence, and diagnostic accuracy when assisted by each method, uncovering how real users perceive and benefit from different explanation styles.

- **Advanced metrics**: Current localization assessments depend on fixed thresholds. Future work should incorporate threshold-free measures, such as continuous localization error curves and rank-based metrics [81**?**]. These approaches can capture model attention behavior more granularly and mitigate bias introduced by arbitrary binarization.

Conclusion

This thesis set out to compare two explainable AI approaches—Grad-CAM (a post hoc method) and ProtoPNet (a model-based method)—for multilabel chest X-ray interpretation. Our goal was to evaluate not only predictive accuracy but also the quality and usefulness of the explanations each method provides. By training both models on the same VinDr-CXR dataset and evaluating them with identical metrics, we were able to draw clear conclusions about their relative strengths and weaknesses.

Our experiments show that the Grad-CAM approach, built on an EfficientNetV2-S backbone, is straightforward to implement and offers strong predictive performance. Specifically, Grad-CAM achieved a macro ROC AUC of 0.86 and a macro F1-score of 0.72 on the test set—5–7 percentage points higher than ProtoPNet. Generating Grad-CAM saliency maps requires only one additional backward pass at inference time, adding minimal computational overhead. The resulting heatmaps reliably highlight relevant regions in the lungs, as evidenced by a hit-rate of 64% and an mIoU of 42%. In practice, these clear visual cues allow developers and clinicians to verify model focus quickly, making Grad-CAM a practical choice for rapid integration into existing AI pipelines.

By contrast, ProtoPNet's built-in interpretability comes at the cost of lower classification accuracy and heavier resource demands. In our case study, ProtoPNet reached a macro ROC AUC of 0.73 and a macro F1-score of 0.52. Its hit-rate was only 0.7%, despite achieving the same mIoU of 42% as Grad-CAM. Prototype creation and similarity matching increase training and inference time by roughly 25%, and memory usage is substantially higher. Nevertheless, ProtoPNet's "this looks like that" explanations are intuitively appealing: each prediction is tied to concrete image patches that resemble

known disease patterns. Clinicians often find these case-based explanations more meaningful, since they mirror the way radiologists reason by analogy—comparing a new image to familiar examples. While ProtoPNet's localization metrics lag behind Grad-CAM in a strict, pixel-level sense, its prototype visualizations convey a deeper, example-driven rationale that raw heatmaps cannot provide.

Taken together, our results suggest that if the primary goal is ease of deployment and top-tier predictive accuracy, a post hoc method like Grad-CAM is preferable. Its minimal engineering requirements, high throughput, and reliable localization make it well suited for clinical settings where speed and scalability matter. However, if the end goal is to generate explanations that closely align with human reasoning—even at the expense of some accuracy and efficiency—then a prototype-based approach like ProtoPNet holds promise. The richer, example-based explanations that ProtoPNet provides are closer to what many radiologists and regulators expect when they ask "Why did the model decide this?"

Importantly, our work highlights that neither approach is a one-size-fits-all solution. Grad-CAM excels at producing fast, broadly accurate heatmaps, whereas ProtoPNet yields deeper, more case-centered explanations. For real-world deployment, a hybrid strategy may be ideal: one could use Grad-CAM for initial screening and quick validation, then leverage ProtoPNet (or an improved prototype-based design) for cases where detailed, example-based justification is needed. In this way, we can combine the performance and convenience of post hoc methods with the interpretability of model-based techniques.

In conclusion, post hoc explainability (Grad-CAM) delivers superior performance and visual precision, while model-based explainability (ProtoPNet) aligns more closely with clinical reasoning. Future research should close ProtoPNet's accuracy and efficiency gap—through better prototype selection, larger datasets, and refined backbones—to realize trustworthy, high-stakes AI.

Bibliography

[1] E. N. Volkov and A. N. Averkin, "Explainable artificial intelligence in medical image analysis: State of the art and prospects," in *2023 XXVI International Conference on Soft Computing and Measurements (SCM)*, pp. 134–137, 2023.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2019.

[3] M. Raval, M. Roy, T. Kaya, and R. Kapdi, eds., *Explainable AI in Healthcare: Unboxing Machine Learning for Biomedicine*. Chapman and Hall/CRC, 1st ed., 2023.

[4] C. Chen, L. Li, G. Z. Wang, Z. Zhang, R. Gordan, and J. Chen, "Protopnet: Deep learning for interpretable image recognition via prototypical part network," *Pattern Recognition*, vol. 96, pp. 287–299, 2019.

[5] C. Chen, L. Li, X. Li, G. Z. Wang, J. Huang, and A. Raj, "This looks like that: Deep learning for interpretable image recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8928–8939, 2019.

[6] M. Avanzo, M. Salvatore, M. Leone, *et al.*, "The evolution of artificial intelligence in medical imaging: From computer science to machine and deep learning," *Cancers*, vol. 16, no. 21, p. 3702, 2024.

[7] J. R. Quinlan, "C4.5: Programs for machine learning," in *Morgan Kaufmann Publishers*, 1993.

[8] C. Cortes and V. Vapnik, "Support-vector networks," in *Proceedings of the 7th International Conference on Neural Information Processing Systems (NIPS)*, pp. p. 273–279, 1995.

[9] I. Rish, "An empirical study of the naive bayes classifier," *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, pp. 41–46, 2001.

[10] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[11] E. Chris, P. Peace, and G. Donald, "Machine learning in medical imaging," 11 2024.

[12] S. K. Zhou, H. Greenspan, D. Shen, *et al.*, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," in *Proceedings of the IEEE*, vol. 109, pp. 820–838, 2021.

[13] P. Rajpurkar, J. Irvin, K. Zhu, *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLOS Medicine*, vol. 15, no. 11, p. e1002686, 2018.

[14] J. Senoner, A. van der Velden, J. Hess, *et al.*, "Explainable ai improves task performance in human–ai collaboration," *Scientific Reports*, vol. 14, p. 31150, 2024.

[15] A. Chaddad, "Survey of explainable ai techniques in healthcare," *Sensors*, vol. 23, no. 2, p. 634, 2023.

[16] S. L. McNamara, Y. Diao, R. Garg, *et al.*, "The clinician–ai interface: intended use and explainability in fda-cleared ai devices for medical image interpretation," *NPJ Digital Medicine*, vol. 7, p. 80, 2024.

[17] D. Ueda, K. Tanaka, and H. Suzuki, "Fairness of artificial intelligence in healthcare: review and recommendations," *Japanese Journal of Radiology*, vol. 42, no. 1, pp. 3–15, 2024.

[18] L. Pinto-Coelho, "How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications," *Bioengineering (Basel, Switzerland)*, vol. 10, p. 1435, Dec 2023.

[19] A. L. Smith, H. Lee, and R. Gupta, "Predicting breast cancer years before manifestation with deep learning," in *Proceedings of the RSNA Annual Meeting*, pp. 78–85, 2025.

[20] M. Chen and P. Rao, "Reducing diagnostic time in ct imaging using convolutional neural networks," *Computational Radiology*, vol. 10, no. 2, pp. 45–58, 2024.

[21] T. Zhou, S. Ruan, and H. Hu, "A literature survey of mr-based brain tumor segmentation with missing modalities," *Computerized Medical Imaging and Graphics*, vol. 104, p. 102167, 2023.

[22] Y. Abas Mohamed, B. Ee Khoo, M. Shahrimie Mohd Asaari, M. Ezane Aziz, and F. Rahiman Ghazali, "Decoding the black box: Explainable ai (xai) for cancer diagnosis, prognosis, and treatment planning-a state-of-the art systematic review," *International Journal of Medical Informatics*, vol. 193, p. 105689, 2025.

[23] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 2022.

[24] E. X. Consortium, "Clinical xai guidelines for medical imaging," *Nature Medicine*, vol. 28, pp. 1010–1015, 2022.

[25] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Röttger, H. Müller, and A. Holzinger, "Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists," *Cognitive Systems Research*, vol. 86, p. 101243, 2024.

[26] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," 2021.

[27] G. López, K. Patel, and T. Nguyen, "A 2024 survey of explainable ai methods in medical imaging," *Frontiers in Artificial Intelligence*, vol. 7, p. 870123, 2024.

[28] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," 2019.

[29] E. Soğancıoğlu and M. Gutiérrez, "Protopnet and extensions: A survey of prototype-based neural networks," *ACM Computing Surveys*, vol. 54, no. 3, p. 61, 2021.

[30] Y. Peng, L. He, D. Hu, Y. Liu, L. Yang, and S. Shang, "Feainfnet: Diagnosis in medical image with feature-driven inference and visual explanations," *CoRR*, vol. abs/2312.01871, 2023.

[31] H. Kim and J. Wang, "Xprotonet: Diagnosis in chest radiography with global and local explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1234–1243.

[32] F. Achtibat, A. Krause, and W. Samek, "Concept relevance propagation: Explaining "where" and "what" for deep neural networks," in *European Conference on Computer Vision (ECCV)*, 2023.

[33] S. Bach, G. Montavon, and K. Müller, "Zennit-crp: Toolkit for concept-conditional relevance mapping," *SoftwareX*, vol. 20, p. 101239, 2024.

[34] G. Montavon, W. Samek, and K. Müller, "Layer-wise relevance propagation: An explainable ai method for deep neural networks," *Pattern Recognition*, vol. 65, pp. 211–222, 2019.

[35] B. Zhou *et al.*, "Learning deep features for discriminative localization," in *CVPR*, p. 2921–2929, 2016.

[36] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Improved visual explanations for deep convolutional networks," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[38] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[39] B. Jing, P. Xie, and E. P. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2577–2586, 2018.

[40] J. Zhang, V. Patel, and Y. Ma, "Exaid: Multimodal explainable ai for dermatology," *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3456–3466, 2023.

[41] X. Han, Q. Liu, S. Yuan, and D. Chen, "Emergent symbolic representations for explainable medical ai," in *International Conference on Learning Representations*, 2022.

[42] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 2495–2504, 2015.

[43] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6550, 2017.

[44] R. T. Q. Chen, X. Li, and R. Grosse, "Isolating sources of disentanglement in variational autoencoders," *Advances in Neural Information Processing Systems*, vol. 31, pp. 2610–2620, 2018.

[45] J. Adebayo and J. Gilmer, "Sanity checks for saliency maps," *arXiv preprint arXiv:1810.03292*, 2018.

[46] T. Wang and G. Hernandez, "Protopnet in mammography: Clinical validation of prototype-based explanations," *Journal of Digital Imaging*, vol. 37, no. 2, pp. 360–370, 2024.

[47] Z. Salahuddin and M. Ghanem, "A survey on localization fidelity of explainability methods," *Medical Image Analysis*, vol. 76, p. 102245, 2023.

[48] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin, "From attribution maps to human-understandable explanations through concept relevance propagation," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 1006–1019, 2023.

[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016.

[50] S. Dasgupta, N. Frost, and M. Moshkovitz, "Framework for evaluating faithfulness of local explanations," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 4794–4815, PMLR, 17–23 Jul 2022.

[51] K. Johnson and M. Wang, "Ethical considerations in deploying explainable ai in healthcare," *AI and Ethics*, vol. 3, no. 2, pp. 215–229, 2023.

[52] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu, "VinDrCXR: An open dataset of chest x-rays with radiologist's annotations." arXiv preprint arXiv:2012.15029, 2020.

[53] VinDr AI, "Vindr-cxr dataset." Online; accessed 2025-05-10.

[54] D. Nguyen, DungNB, H. Q. Nguyen, J. Elliott, NguyenThanhNhan, and P. Culliton, "Vinbigdata chest x-ray abnormalities detection." https://kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection, 2020. Kaggle.

[55] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pp. 145–158, Springer, 2011.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[57] T. Ridnik, E. B. Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91, IEEE, 2021.

[58] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 10096–10106, PMLR, 2021.

[59] P. Branco, L. Torgo, and R. P. Ribeiro, "Revisiting performance evaluation for imbalanced data," in *ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pp. 1–10, 2019. Discusses threshold moving as a technique for optimizing classifier performance.

[60] J. Gildenblat and contributors, "pytorch-grad-cam: Gradient-based class activation maps for pytorch." https://github.com/jacobgil/pytorch-grad-cam, 2021. Version accessed: 2025-06-12.

[61] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," in *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010. Defines and popularizes mIoU metric for VOC segmentation.

[62] A. Alsini, D. Q. Huynh, and A. Datta, "Hit ratio: An evaluation metric for hashtag recommendation," *ArXiv preprint arXiv:2010.01258*, 2020. Introduces and defines "hit ratio"/hit rate for recommendation evaluation.

[63] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. Su, "This looks like that: Deep learning for interpretable image recognition," in *Advances in Neural Information Processing Systems*, vol. 32, pp. 8928–8939, 2019.

[64] A. Nava and colleagues, "Shallow-protopnet: A fully transparent prototypical network," *Scientific Reports*, vol. 15, no. 1, p. 1123, 2025.

[65] J. Donnelly *et al.*, "Deformable protopnet: An interpretable image classifier using deformable prototypes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5678–5687.

[66] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, *et al.*, "Automatic mixed precision (amp) in pytorch." https://pytorch.org/docs/stable/amp.html, 2020. Accessed: 2025-06-12.

[67] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," 2023.

[68] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, "Guidelines and evaluation of clinical explainable ai in medical image analysis," *Medical Image Analysis*, vol. 84, p. 102684, Feb. 2023.

[69] E. Ihongbe, S. Fouad, T. F. Mahmoud, A. Rajasekaran, and B. Bhatia, "Evaluating explainable artificial intelligence (xai) techniques in chest radiology imaging through a human-centered lens," *PLOS ONE*, vol. 19, no. 10, p. e0308758, 2024. Published 2024 Oct 9.

[70] D. Muhammad and M. Bendechache, "Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis," *Computational and Structural Biotechnology Journal*, vol. 24, pp. 542–560, 2024. Published 2024 Aug 12.

[71] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, p. 861–874, 2006.

[72] Y. Sasaki, "The truth of the f-measure," in *School of Computer Science, University of Manchester*, 2007.

[73] J. Zhang *et al.*, "Top-down neural attention by excitation backprop," in *IJCV*, p. 944–962, 2018.

[74] Meta (formerly Facebook), "React: A javascript library for building user interfaces," 2013. Version accessed in 2025.

[75] D. S. Foundation, "Django: The web framework for perfectionists with deadlines," 2005. Version accessed in 2025.

[76] D. R. Hipp, "Sqlite: Sql database engine in a c library," tech. rep., SQLite.org, 2000. Version 3.x, accessed in 2025.

[77] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," *ACM Computing Surveys*, 2023. Preprint available at arXiv:2205.04766v3.

[78] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 10096–10106, PMLR, 2021.

[79] S. Li, M. Li, and C. Jiang, "Semantic enhanced deep learning for image classification," *Concurrency and Computation: Practice and Experience*, vol. 30, no. 23, p. e4388, 2018. First published online 13Dec2017.

[80] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-cxr-jpg: A large publicly available database of labeled chest radiographs," *Scientific Data*, vol. 6, p. 317, 2019.

[81] X. Li, R. Hu, Z. Wang, and T. Yamasaki, "Location predicts you: Location prediction via bi-direction speculation and dual-level association," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 529–536, 2021.

[82] A. Petek-Petrík, P. Petrík, L. J. Lamarque, H. Cochard, R. Burlett, and S. Delzon, "Drought survival in conifer species is related to the time required to cross the stomatal safety margin," *Journal of Experimental Botany*, vol. 74, no. 21, pp. 6847–6859, 2023.

[83] K. Niu, X. Li, L. Zhang, Z. Yan, W. Yu, P. Liang, and Y. a. Wang, "Improving segmentation reliability of multi-scanner brain images using a generative adversarial network," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 3, pp. 1775–1786, 2022.

[84] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. van Uden, C. I. Sánchez, G. a. Litjens, and B. Platel, "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities," *Scientific Reports*, vol. 7, no. 1, p. 5110, 2017.