Enhancing Visual Interpretability in Computer-Assisted Radiological Diagnosis: Deep Learning Approaches for Chest X-Ray Analysis

Zirui Qiu

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

August 2024

© Zirui Qiu, 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Zirui Qiu**

Entitled: **Enhancing Visual Interpretability in Computer-Assisted Radiological Diagnosis: Deep Learning Approaches for Chest X-Ray Analysis**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

________________________ Chair
Dr. Mirco Ravanelli

________________________ External Examiner
Dr. Wei-Ping Zhu

________________________ Examiner
Dr. Mirco Ravanelli

________________________ Co-supervisor
Dr. Hassan Rivaz

________________________ Co-supervisor
Dr. Yiming Xiao

Approved by ________________________
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

____________ 2024 ________________________
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Enhancing Visual Interpretability in Computer-Assisted Radiological Diagnosis: Deep Learning Approaches for Chest X-Ray Analysis

Zirui Qiu

This thesis delves into the realm of interpretability in medical image processing, focusing on deep learning's role in enhancing the transparency and understandability of automated diagnostics in chest X-ray analysis. As deep learning models become increasingly integral to medical diagnostics, the imperative for these models to be interpretable has never been more pronounced. This work is anchored in two main studies that address the challenge of interpretability from distinct yet complementary perspectives. The first study scrutinizes the effectiveness of Gradient-weighted Class Activation Mapping (Grad-CAM) across various deep learning architectures, specifically evaluating its reliability in the context of pneumothorax diagnosis in chest X-ray images. Through a systematic analysis, this research reveals how different neural network architectures and depths influence the robustness and clarity of Grad-CAM visual explanations, providing valuable insights for selecting and designing interpretable deep learning models in medical imaging. Building on the foundational understanding of interpretability, the second study introduces a novel deep learning framework that enhances the synergy between disease diagnosis and the prediction of visual saliency maps in chest X-rays. This dual-encoder, multi-task UNet architecture, augmented by a multi-stage cooperative learning strategy, offers a sophisticated approach to interpretability. By aligning the model's attention with that of clinicians, the framework not only enhances diagnostic accuracy but also provides intuitive visual explanations that resonate with clinical expertise. Together, these studies contribute to the field of medical image processing by offering innovative approaches to improve the interpretability of deep learning models. The findings underscore the potential of interpretability-enhanced models to foster trust among medical practitioners, facilitate better clinical decision-making, and pave the way for the broader acceptance and integration of AI in healthcare diagnostics. The thesis concludes by synthesizing the insights gained from both projects and outlining prospective pathways for future research to further advance the interpretability and utility of AI in medical imaging.

Acknowledgments

First and foremost, I would like to express my deep gratitude to my supervisors, Dr. Hassan Rivaz and Dr. Yiming Xiao, for their persistent guidance and support throughout my studies. Their innovative ideas, effective leadership, and thorough reviews have provided me with an excellent opportunity to learn and thrive in my research endeavors.

I am also indebted to my colleagues at the IMPACT and HEALTH-X labs, whose help and support have been invaluable. Their collaboration and encouragement have significantly enriched my research experience.

Special thanks to my dear family and friends, whose unwavering support has been fundamental to my success. I am deeply grateful for all the love and encouragement that I have received from them, making this journey possible.

This thesis, and all its findings, was highly dependent on the carefully collected dataset of chest X-ray images and radiologist eye gaze data. I extend my gratitude to all those individuals involved in collecting and annotating this crucial data.

Finally, we acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), and NVIDIA for the donation of the GPU, which were instrumental in conducting this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medical image processing plays an indispensable role in modern healthcare, particularly in extracting precise and actionable insights from imaging data to significantly influence diagnostic and therapeutic decisions. With the advent of deep learning, the capabilities of image processing have expanded tremendously, introducing a new era of accuracy and efficiency. However, the inherent opacity of these deep learning models, often described as ”black-box” approaches, poses a significant challenge in clinical adoptions and regulataroy approvals, especially in the context of medical decision-making, where understanding and trust are paramount.

This thesis focuses on enhancing the interpretability of deep learning models used in medical image processing, specifically targeting lung-related diseases in chest X-ray (CXR) images. It aims to bridge the gap between advanced AI technologies and their practical, understandable applications in healthcare. One of the primary methods explored is the generation of saliency maps that align with the visual attention of clinicians during their diagnostic processes. By ensuring that these saliency maps closely mimic the areas doctors focus on in CXR images, the thesis strives to make the diagnostic outcomes derived from AI models more transparent and convincing.

In addition to saliency maps, this work also incorporates techniques like Gradient-weighted Class Activation Mapping (Grad-CAM), which provides visual explanations for the decisions made by deep learning models. Figure 1.1 illustrates an example of Grad-CAM for CXR imgages. Grad-CAM uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

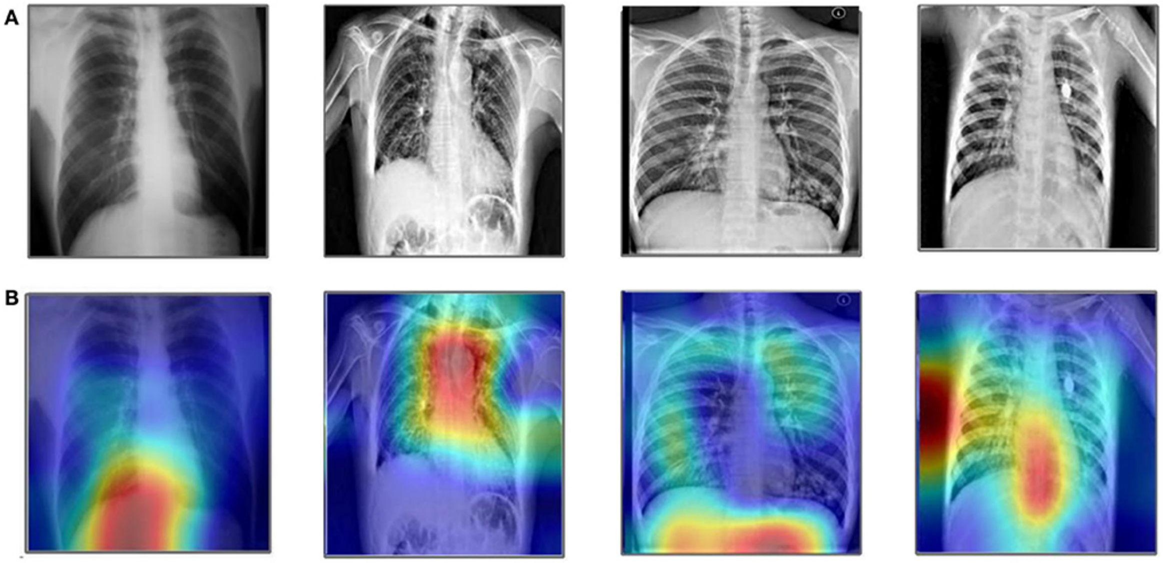


Figure 1.1: An illustration of Grad-Cam for Chest X-Rays (from [17]).

By applying these interpretability techniques to CXR images, particularly those that capture pulmonary conditions, the thesis aims to elucidate the decision-making processes of AI models. This clarification helps build trust between medical professionals and AI models by allowing them to understand and verify the AI-generated insights, thereby facilitating more informed and accurate clinical decision-making.

1.1 A Deeper Look Into Chest X-Ray Disease Classification

Chest X-ray (CXR) classification is a crucial component in the diagnostic processes within healthcare, particularly in the detection and management of various pulmonary conditions. X-rays are among the most common and widely utilized diagnostic imaging tools due to their accessibility, cost-effectiveness, and the ability to quickly provide valuable insights into numerous health conditions, ranging from pneumonia and tuberculosis to lung cancer and COVID-19. The classification of these images, therefore, holds paramount importance, as it aids clinicians in making rapid and accurate medical decisions that are crucial for patient care [37].

The advent of deep learning has significantly revolutionized the field of CXR classification by introducing models capable of achieving and sometimes surpassing human-level accuracy in identifying pathologies. These advancements have not only enhanced diagnostic precision but have also streamlined workflow in medical settings, reducing the burden on radiologists and allowing more timely intervention for patients. However, despite these benefits, the deployment of deep learning models in clinical practice has been hampered by the ‘black-box’ nature of these technologies, where the reasoning behind their predictions is not always clear [25].

This opacity in AI-driven tools poses a significant challenge, as medical professionals seek not only predictive accuracy, but also transparency and interpretability in automated clinical decisions. Addressing these concerns involves navigating complex ethical and operational considerations, ensuring that AI technologies augment healthcare in a manner that is both ethical and beneficial to patient care [9].

1.2 The Need for Interpretability in Chest X-Ray Classification

The necessity for interpretability in chest X-ray (CXR) classification arises from several critical demands in the healthcare domain, particularly around ensuring the safety, reliability, and clinical utility of artificial intelligence (AI) applications. CXRs are instrumental in diagnosing a wide range of pulmonary diseases, from common conditions like pneumonia to more severe pathologies such as lung cancer. While deep learning models have significantly enhanced the accuracy of CXR analysis, their adoption in clinical practice hinges on more than just high-performance metrics; it requires a clear understanding of how these models arrive at their conclusions.

- **Transparency and Trust:** In medical practice, decisions that affect patient care must be based on transparent and reliable methods. Deep learning models, with their complex architectures, often function opaquely. Without a clear understanding of how decisions are made, clinicians may be hesitant to trust and rely on these tools, irrespective of their accuracy. Interpretability helps demystify the decision-making process of AI systems, fostering trust among healthcare providers by aligning model outputs with clinical expectations and reasoning.[38, 20]

- **Clinical Validation and Debugging:** Interpretability facilitates the clinical validation of AI models by allowing medical experts to assess the rationale behind specific diagnostic suggestions. This review process is essential not only for integrating AI tools into routine clinical workflows but also for identifying and correcting potential errors or biases in the models. By understanding the features and patterns that a model considers important, clinicians and data scientists can fine-tune these systems to improve their accuracy and generalizability across diverse patient populations.[55]

- **Incorporating Clinical Expertise:** Despite the advances in visual explanation methods like Grad-CAM, these tools often fall short in precisely highlighting areas of importance identified by clinical experts. Studies have shown that Grad-CAM heatmaps (e.g., Fig. 1.2) do not always align with the critical regions marked by radiologists, potentially leading to misinterpretations and diagnostic oversights [22]. To address these limitations, integrating direct inputs such as eye-gaze data from radiologists into the training and validation of these models is crucial. This integration can enhance the model's focus, ensuring that the AI's attention coincides with clinically relevant features, thereby improving both the interpretability and reliability of the outputs.

1.3 Thesis Contributions

This thesis advances the interpretability of deep learning models in medical image processing, specifically within the realm of chest X-ray (CXR) classification. Here are the detailed contributions based on the comprehensive analysis presented in the two published/submitted papers:

(1) **Grad-CAM Effectiveness Across Different Architectures:** The first contribution of this thesis is a systematic investigation of Gradient-weighted Class Activation Mapping (Grad-CAM) across a variety of deep learning architectures. Detailed in Chapter 3, this study evaluates the robustness and effectiveness of Grad-CAM in generating interpretable visual explanations for various popular CXR classification models. It explores how variations in network depth and architectural design influence the quality of Grad-CAM visualizations,

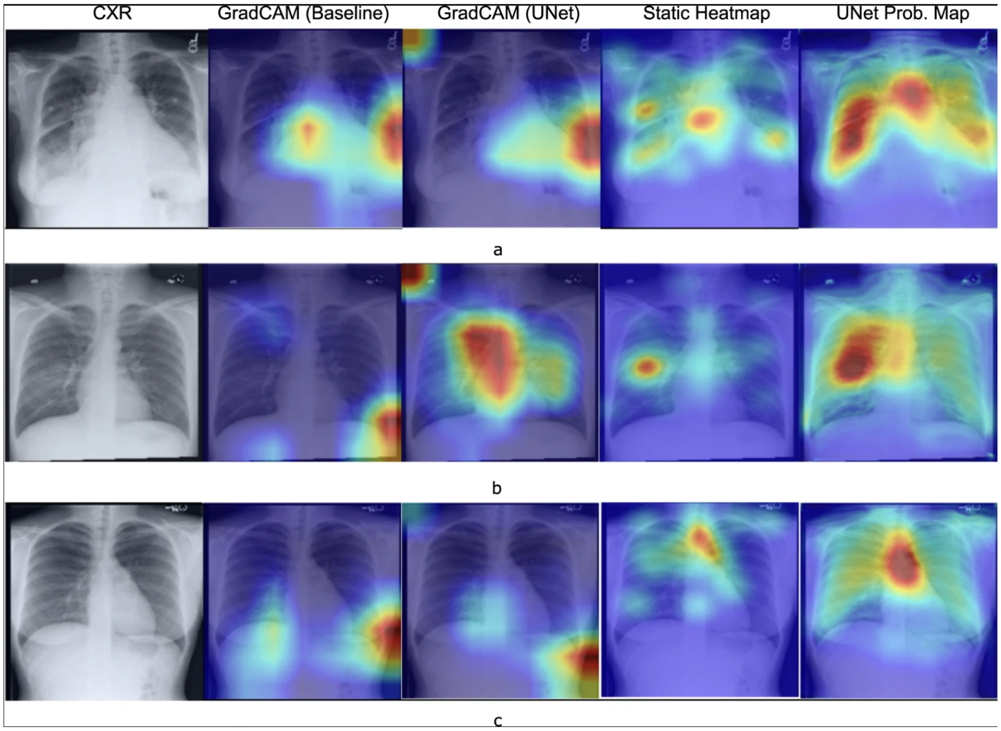


Figure 1.2: Qualitative comparison of the interpretability of U-Net based probability maps in comparison with GradCAM for a few example use cases. (a) Congestive Heart Failure (CHF). The physician's eye gaze tends to fall on the enlarged heart and hila, as well as generally on the lungs, (b) Pneumonia. The physician's eye gaze predictably focuses on the focal lung opacity and (c). Normal. Because the lungs are clear, the physician's eye gaze skips around the image without focus[22]

offering crucial insights for selecting optimal models that balance interpretability with diagnostic accuracy.

(2) **Novel Multi-task Learning Framework for Enhanced Interpretability:** The second and more theoretical contribution of this thesis, which will be extensively discussed in Chapter 4, involves developing a novel multi-task learning framework. This framework is designed for simultaneous disease diagnosis and clinical visual attention prediction using chest X-rays. By integrating a dual-encoder, multi-task UNet architecture that combines a DenseNet201 backbone with a Residual and Squeeze-and-Excitation block-based encoder, this approach significantly improves the generation of visual saliency maps. The introduction of a multi-stage cooperative learning strategy enhances the model's performance and alignment with clinical diagnostic processes, making this approach a major step forward in making deep learning tools more interpretable and clinically viable.

1.4 Thesis Outline

This thesis is organized into a structured format to systematically explore the interpretability of deep learning models in chest X-ray (CXR) analysis. Chapter 2 sets the stage with a background and literature review, detailing the evolution of medical image processing and focusing on the integration of deep learning techniques in CXR analysis. It also reviews existing interpretability methods such as Grad-CAM and saliency maps, discussing their application and relevance in clinical diagnostics. Chapter 3 delves into the initial study of the thesis, which examines the effectiveness of Grad-CAM across different neural network architectures, analyzing how variations in model design influence the clarity and utility of visual explanations. Chapter 4 presents the major contribution of the thesis, a novel multi-task learning framework designed for simultaneous disease diagnosis and visual attention prediction, showcasing a dual-encoder, multi-task UNet architecture enhanced by a multi-stage cooperative learning strategy. The final chapter, Chapter 5, concludes the thesis by summarizing the findings and discussing their implications for medical image processing. It reflects on the potential future directions for research, emphasizing how ongoing advancements could further refine the interpretability and clinical applicability of AI tools in healthcare diagnostics.

Chapter 2

Background and Literature Review

2.1 Deep Learning in Medical Imaging

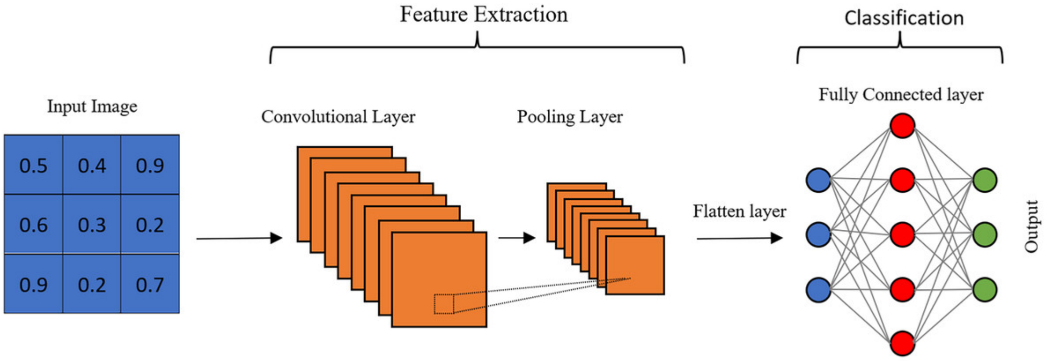
Transformative technologies in medical imaging such as deep learning have substantially improved the analysis and interpretation of diagnostic images. In this section, deep learning techniques in medical imaging are explored with consideration to their theoretical bases, major structures, and definite implementations for chest X-ray (CXR) analysis.

2.1.1 Fundamentals of Deep Learning for Medical Imaging

Deep learning algorithms, particularly those based on neural networks, excel in identifying complex patterns in large datasets. In medical imaging, these algorithms automate the feature extraction and classification processes, traditionally performed manually by radiologists, thereby augmenting diagnostic precision and efficiency.

2.1.2 Key Architectures in Medical Imaging

- **Convolutional Neural Networks (CNNs):** CNNs have transformed the field of medical imaging through their ability to process and analyze complex image data efficiently. The Figure 2.1 shows these networks utilize layers of convolutional filters to extract features from images at various levels of abstraction. In the context of chest radiography (CXR), CNNs can detect subtle visual cues that are indicative of various diseases, ranging from simple infections



0.5	0.4	0.9
0.6	0.3	0.2
0.9	0.2	0.7

Figure 2.1: Convolutional Neural Network Architecture (from [1]).

to complex conditions such as lung cancer. The strength of CNNs lies in their capacity to learn feature hierarchies directly from the data, eliminating the need for manual feature selection, which is often subjective and labor-intensive [27]. This automation significantly enhances the diagnostic process, allowing for high-throughput analysis and greater diagnostic accuracy.

- **U-Net:** Originally designed for segmenting biomedical images, U-Net's architecture is particularly well-suited for medical tasks that require precise localization of features within an image. Its architecture shows in Figure 2.2, which includes a contracting path to capture context and an expansive path that enables precise localization, is ideal for tasks such as delineating the boundaries of tumors or identifying regions affected by pathological changes in CXRs. U-Net has been pivotal in improving the performance of medical imaging tasks by providing detailed and accurate segmentation maps that are crucial for treatment planning and monitoring disease progression [43]. Its effectiveness in medical imaging stems from its ability to work with a limited amount of data and still produce high-quality results, which is often the case in medical scenarios where annotated images are scarce.

- **Transformer:** Transformers [56] have revolutionized the field of deep learning, initially gaining prominence in natural language processing due to their unique ability to manage long-range dependencies within datasets. Unlike traditional convolutional neural networks (CNNs) that process input data through localized filters, Transformers look at all parts of the data simultaneously, which is incredibly useful in medical imaging where a contextual

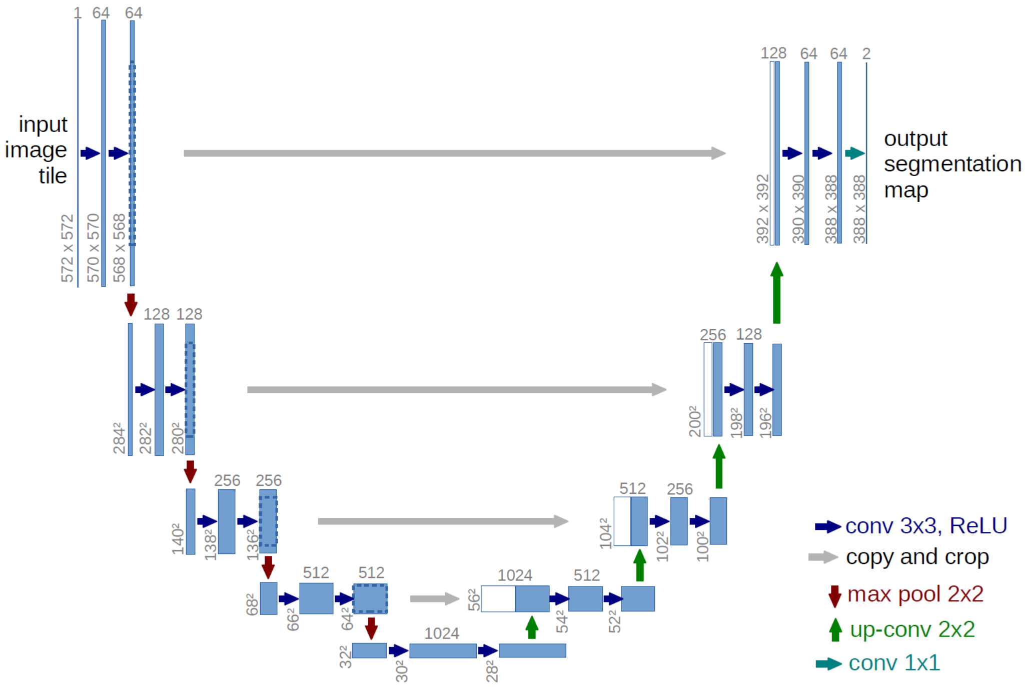


Figure 2.2: UNet model architectur (from[43]).

analysis is often crucial.

Vision Transformer (ViT) is an adaptation of the transformer architecture. In ViT, each image is split into several pieces, known as patches. As illustrated in Figure 2.3, these patches are then converted into a series of flat, linear pieces of data or embeddings and lined up like sentences in a text for the transformer to process. As a result, ViTs have a knack for spotting patterns and connections across the whole image, which often lets them outperform traditional models like CNNs that look at images piece by piece [12].

Adapting Transformers for medical imaging not only boosts diagnostic accuracy, but also expands the potential of AI in new medical image processing applications. Their ability to process entire images at once is a big improvement over models that tend to focus on local features. This could lead to more informed and reliable diagnostic decisions with the consideration of richer contextual information.

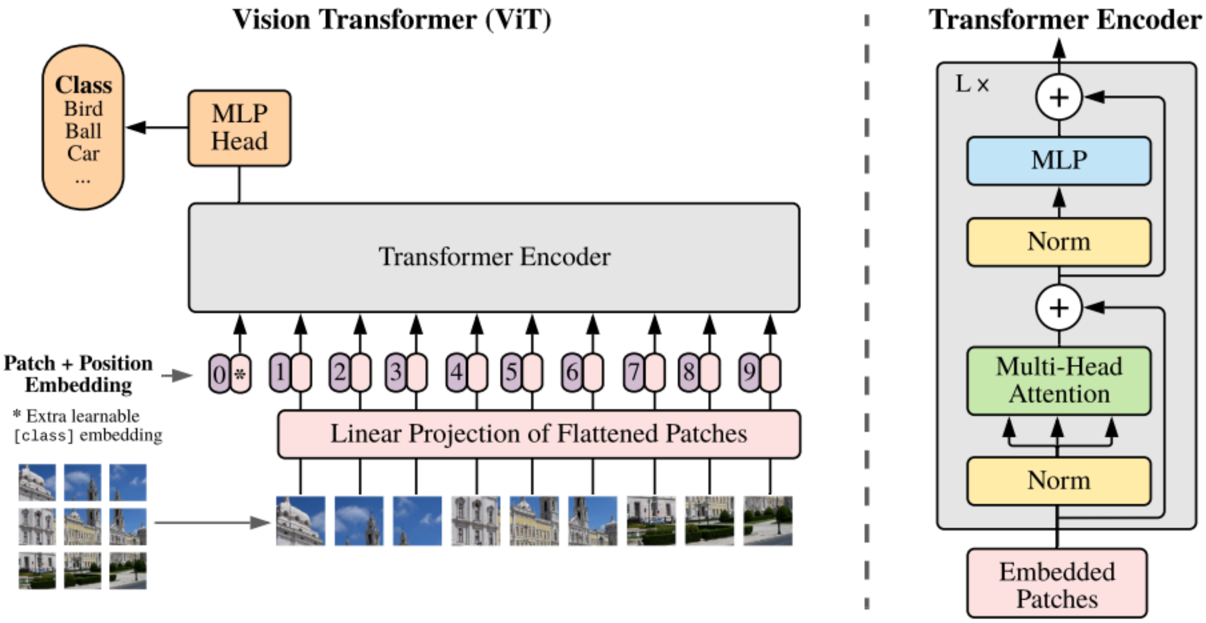


Figure 2.3: ViT model architecture (from[12]).

2.1.3 Application to Chest X-Ray Analysis

The utilization of deep learning in CXR analysis has significantly enhanced the detection and diagnosis of critical pathologies, such as pneumonia and pneumothorax, surpassing traditional image processing methods in both accuracy and speed.

Pneumonia Detection: Pneumonia, often characterized by an increase in lung opacity due to fluid or infection, presents specific challenges in radiographic interpretation. Convolutional Neural Networks (CNNs) are particularly effective in distinguishing these opacities from other benign conditions. Deep learning models, trained on large datasets of annotated CXR images, have demonstrated a remarkable ability to detect pneumonia, significantly reducing diagnostic time and enhancing accuracy. This is crucial for timely intervention, which can be life-saving in severe cases [37].

Pneumothorax Identification: Pneumothorax, the presence of air or gas in the chest cavity, is another area where deep learning models excel. The condition, which can lead to a collapsed lung, requires rapid detection and intervention. Deep learning algorithms, especially those employing

U-Net architectures for segmentation tasks, have proven adept at identifying subtle signs of pneumothorax that may be missed during manual review. These models enhance the radiologist's ability to diagnose pneumothorax quickly and accurately, thereby improving patient outcomes [25].

These deep learning applications not only achieve high diagnostic accuracy, but also reduce the time required for analysis, an essential factor in emergency medical scenarios where quick decision-making is critical. The ability of these models to continuously learn and improve from new data ensures that their effectiveness grows over time, keeping pace with evolving medical standards and practices.

2.1.4 Challenges in Deep Learning Applications

Despite the significant advancements deep learning has brought to medical imaging, several challenges persist that complicate its deployment. One of the primary challenges is data dependency and the need for extensive annotated datasets. Deep learning models require large volumes of high-quality, annotated images to train effectively. In the medical field, obtaining such datasets is particularly arduous due to the need for expert annotation, which is time-consuming, costly, and often constrained by the availability of specialists. Moreover, concerns about patient privacy and stringent regulatory requirements complicate the collection and sharing of medical data [3].

Another critical issue is the inherent complexity of these models, which often results in a lack of transparency about how decisions are made. This opacity is problematic in clinical settings, where trust and reliability are paramount. Without clear insight into the decision-making processes of these models, clinicians may hesitate to rely on them, hindering their integration into routine medical practice [20]. Additionally, the ability of these models to generalize across different demographics and medical equipment without bias is a significant concern. Models trained on limited datasets may not perform well universally, leading to diagnostic discrepancies and potential patient safety risks [62].

Regulatory and ethical considerations also pose substantial challenges. As regulatory bodies scramble to catch up with rapid technological advancements, there remains a significant gap in the guidelines and standards governing the use of AI in healthcare. Ethical concerns such as potential increased patient surveillance, data misuse, and decisions made without sufficient human oversight further complicate the deployment of AI technologies in clinical environments [9]. Moreover, integrating deep learning models into existing healthcare systems presents practical challenges related to scalability, compatibility with existing hardware, and workflow integration. These factors must be carefully considered to ensure that the introduction of AI technologies enhances rather than disrupts clinical practices [13].

2.2 Deep Learning Model Interpretability Techniques

Deep learning model interpretability techniques are crucial when dealing with machine learning models that are not inherently interpretable. These techniques aim to make the decision-making process transparent by showing how predictions are derived from given inputs. The most commonly used methods for achieving interpretability include feature importance methods and instance-based approaches.

2.2.1 Feature Importance Methods

Feature importance methods identify which input features significantly impact the model's output, offering insights into the predictive elements within the data. Two primary types of feature importance methods are:

- **Perturbation-based:** This approach assesses the impact of altering one or more input features on the model's predictions. While straightforward and widely applicable, these methods can be slow and may not fully capture the complexities of nonlinear interactions in advanced models [31, 41].

- **Gradient-based:** These methods calculate the gradient of the output with respect to each input feature, offering a more efficient way to gauge feature significance. Methods like DeepLIFT compare activations to a reference to assign importance scores efficiently [49].

2.2.2 Instance-based Methods

Instance-based methods form a core part of the interpretative framework in this thesis, focusing on making AI decisions in medical imaging comprehensible by mimicking human expert behavior. These methods are particularly effective in providing explanations that are local to specific instances, which in the context of medical imaging, involves using saliency maps that replicate how doctors visually assess diagnostic images.

In this thesis, instance-based methods have been applied by training AI models to generate saliency maps that align with those observed in human experts, such as radiologists. By doing so, the AI models learn to highlight the same areas of interest that a doctor would, thereby making the model's diagnostic reasoning transparent and understandable. This approach not only enhances trust in AI-driven diagnostics by providing clear visual evidence of the model's focus but also helps in verifying the clinical relevance of the features identified by the AI.

2.3 Techniques to Generate Saliency Map for Enhancing Interpretability

In medical imaging, particularly in chest X-ray (CXR) analysis, interpretability of AI-driven diagnostic tools is fundamental. The ability to visually interpret what a model "sees" and "considers important" not only aids clinicians in understanding and trusting AI, but also plays a crucial role in validating the diagnostic decisions made by these models.

2.3.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM as shown in Figure 2.4 is a visualization technique that enhances the interpretability of convolutional neural networks by highlighting the regions within an image that are important for a specific classification decision. The process begins by forwarding an image through the CNN to generate feature maps, followed by performing task-specific computations that produce a score for a target category. To isolate the influence of the desired class, gradients are set to zero for all other classes except the targeted one, which is set to one. These gradients are then backpropagated to the

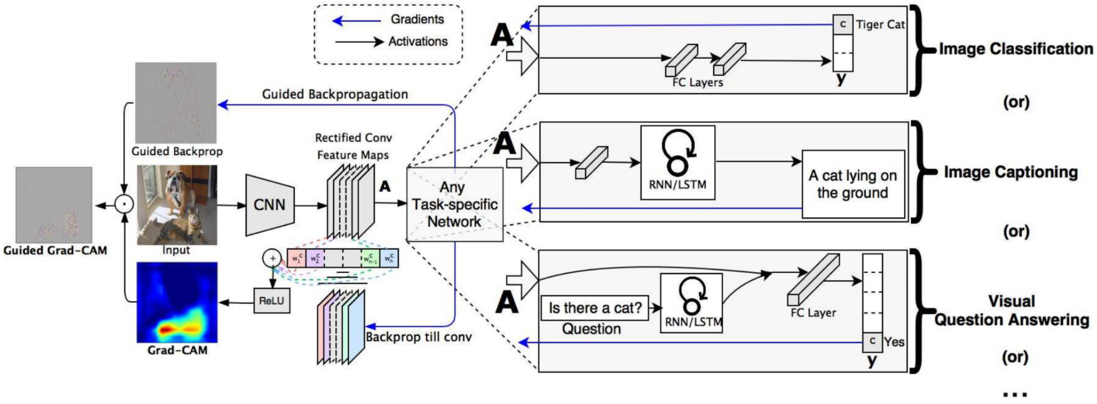


Figure 2.4: Grad-CAM overview (from [47]).

convolutional feature maps, and the resulting data are combined to produce a coarse localization heatmap, illustrating where the model focused to make its decision. This heatmap can be further refined by combining it with guided backpropagation, resulting in Guided Grad-CAM visualizations that offer high-resolution and concept-specific insights into the model's reasoning process [47].

2.3.2 Attention Maps from Transformers

Transformers utilize self-attention mechanisms that assess the importance of all parts of the image relative to each other, rather than being restricted by the local regions typically prioritized by convolutional neural networks. This method allows the model to dynamically adjust its focus across the entire image, making attention maps from transformers particularly valuable for tasks where global context and fine details are crucial. For instance, in tasks like image captioning or object detection in natural scenes, attention maps can vividly illustrate how the model shifts its focus to different objects and background elements, providing a visual explanation of how various components of the image contribute to the final prediction [12] (see Figure 2.5)

2.3.3 Human Gaze Tracking

Complementing machine-generated attention maps, human gaze tracking captures the visual attention of expert radiologists during diagnostic evaluations. By comparing these human-derived attention maps with those produced by transformer models, researchers and clinicians can assess

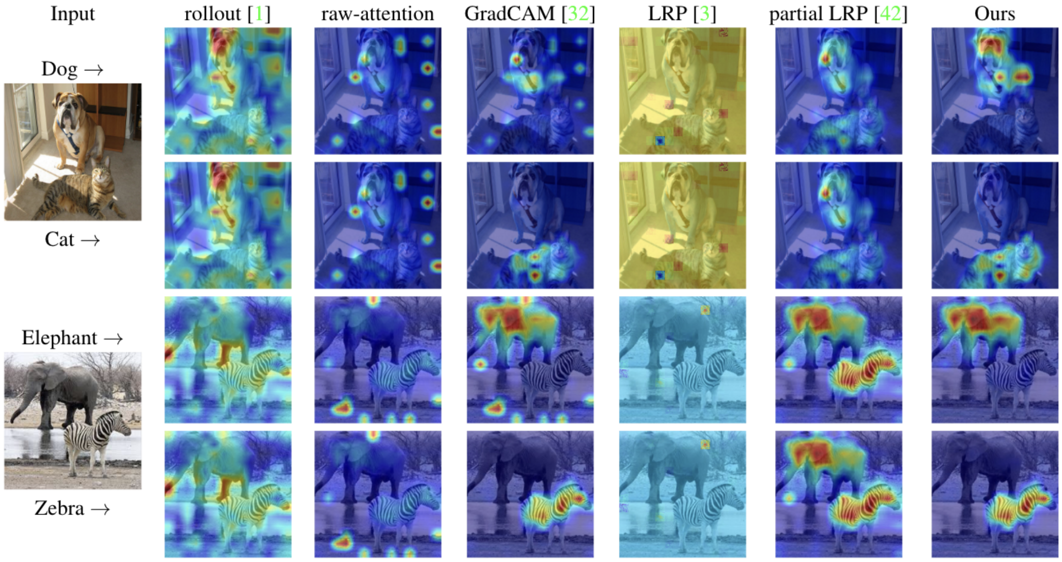


Figure 2.5: Sample attention map results from Hila et al. study of attention map [10], showcasing six different methods: 1. Rollout, which aggregates attention across all layers, highlighting generalized areas of interest. 2. Raw-attention, focusing sharply on specific features like animal faces and bodies. 3. GradCAM, emphasizing critical features through gradients of the last convolutional layer. 4. LRP, illustrating both positive and negative influences on classification. 5. Partial LRP, focusing on selected layers' contributions. 6. Hila et al. improved method, integrating various aspects for a refined visualization of model focus areas.

whether AI systems are focusing on the same key diagnostic features as human experts. This comparison not only validates the AI's diagnostic reasoning, but also ensures alignment with expert human judgment, crucial for clinical acceptance and reliability [6].

2.4 Multi-Task Learning

Multi-task learning (MTL) is a subset of machine learning that enhances learning efficiency and prediction accuracy by simultaneously solving multiple related problems using a shared representation. This approach is particularly beneficial in medical imaging, where the complexity of diagnostic tasks often benefits from the simultaneous learning of correlated objectives, such as detecting various diseases from chest X-rays or segmenting multiple anatomical structures. This section explores the principles of multi-task learning, its specific applications in medical imaging, and its integration into the systems developed in the second major work of this thesis.

2.4.1 Foundations of Multi-Task Learning:

Multi-task Learning (MTL) optimizes the efficiency and effectiveness of machine learning by addressing multiple related tasks simultaneously using a shared representation. As depicted in the Figure 2.6, the architecture of an MTL system typically features shared layers at the base, which extract and process core features from input data applicable across all tasks. These universal features then feed into task-specific layers tailored to the unique requirements of each task (e.g., Task A, Task B, Task C). This structure allows the system to leverage commonalities across tasks to improve generalization while preserving the ability to specialize through task-specific adaptations. The dual benefit of this approach is especially valuable in medical imaging, where complex diagnostics can be enhanced by learning correlated objectives concurrently, thereby enhancing diagnostic accuracy and reducing computational redundancy [44].

2.4.2 Applications in Chest X-ray Analysis:

In the context of chest X-rays, multi-task learning is particularly helpful. For example, the MT-UNet model discussed by Zhu et al. [64] demonstrates how simultaneous tasks—saliency prediction

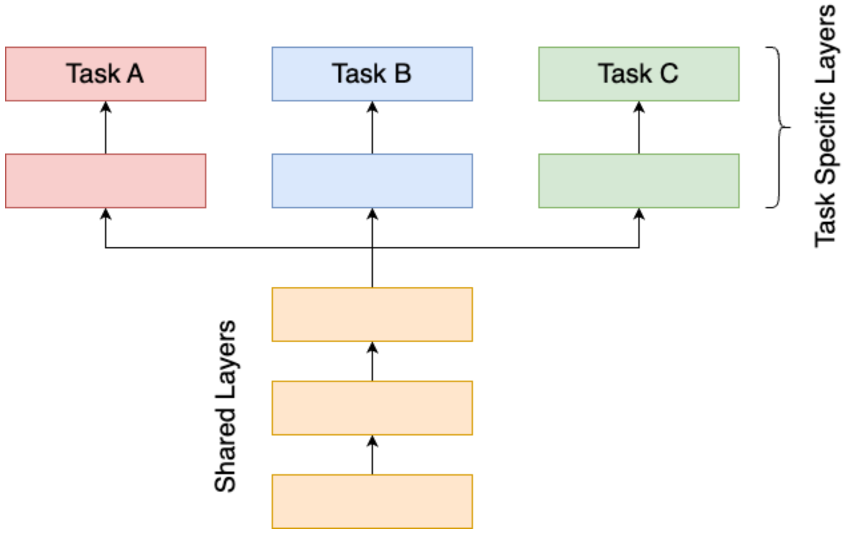


Figure 2.6: General Multi-Task Learning overview (inspired from[44]).

and disease classification—can synergistically improve each other. Saliency prediction helps focus the model's attention on significant areas of the image, potentially leading to more accurate disease classification. Conversely, insights gained from disease classification tasks can inform and refine the processes of identifying key salient features in images.

2.5 Common Metrics for Saliency Evaluation

This section discusses three key metrics commonly used to measure the effectiveness of saliency predictions.

2.5.1 Kullback-Leibler Divergence (KLD):

KLD is a statistical measure used to determine how one probability distribution diverges from a reference probability distribution. In the context of saliency maps, KLD measures the difference between the predicted saliency map and the ground truth (typically derived from expert annotations or eye-tracking data). A lower KLD value indicates that the predicted saliency distribution more closely matches the ground truth, suggesting higher accuracy in the model's attention focusing. The formula for KLD is:

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{1}$$

where P is the predicted saliency distribution within the map and Q is the ground truth distribution.

2.5.2 Pearson's Correlation Coefficient (PCC):

PCC assesses the spatial correlation between the predicted and actual saliency maps. It provides a value between -1 and 1, where 1 means perfect positive correlation, -1 indicates perfect negative correlation, and 0 signifies no correlation. Higher values of PCC indicate that the model's predicted saliency map aligns well with the ground truth in terms of distribution patterns, reflecting the model's ability to focus on relevant areas in the image as human experts do. The formula for PCC is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{2}$$

where X_i and Y_i are the values of the predicted and actual saliency at the ith pixel, respectively, and $\overline{X}$ and $\overline{Y}$ are the means of these distributions.

2.5.3 Histogram Similarity (HS):

HS evaluates the similarity between the histograms of the predicted and ground truth saliency maps. This metric is particularly useful for understanding whether the overall values of saliency across the entire image match, even if their exact spatial distribution does not. HS can be a useful complement to other metrics by providing an aggregate measure of similarity that is less sensitive to precise spatial alignment but focuses on the general distribution of saliency values.

Chapter 3

Is visual explanation with Grad-CAM more reliable for deeper neural networks? a case study with automatic pneumothorax diagnosis

A version of this chapter was presented at the the Machine Learning in Medical Imaging (MLMI 2023) workshop co-hosted with the 26th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023.

- Qiu, Z., Rivaz, H., & Xiao, Y. (2023, October). Is visual explanation with Grad-CAM more reliable for deeper neural networks? a case study with automatic pneumothorax diagnosis. In International Workshop on Machine Learning in Medical Imaging (pp. 224-233). Cham: Springer Nature Switzerland. [35]

3.1 Introduction

With rapid development, deep learning (DL) techniques have become the state-of-the-art in many vision applications, such as computer-assisted diagnosis. Although initially proposed for natural image processing, staple Convolutional Neural Networks (CNNs), such as VGG and ResNet architectures, have become ubiquitous backbones in computer-assisted radiological applications due to their robustness and flexibility to capture task-specific, complex image features. Furthermore, the more recent Vision Transformer (ViT), a new class of DL architecture that leverage the self-attention mechanism to encode long-range contextual information is attracting great attention in medical image processing, with evidence showing superior performance than the more traditional CNNs [11]. Although these DL algorithms can offer excellent accuracy, one major challenge that hinders their wide adoption in clinical practice is the lack of transparency and interpretability in their decision-making process. So far, various explainable AI (XAI) methods have been proposed [2], and among these, direct visualization of the saliency/activation maps has gained high popularity, likely due to their intuitiveness for fast uptake in clinical applications. With high flexibility and ease of implementation for different DL architectures, the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [47], which provides visual explanation as heatmaps with respect to class-wise decision has been applied widely in many computer-assisted diagnostic and surgical vision applications. However, in almost all previous investigations, Grad-CAM outcomes are only demonstrated qualitatively. To the best of our knowledge, the impacts of different deep learning architectures and sizes on the robustness and effectiveness of Grad-CAM have not been investigated, but are important for the research community.

To address the mentioned knowledge gap, we benchmarked the performance of Grad-CAM on DL models of three types of popular architectures, including VGG, ResNet, and ViT, with varying network depths/sizes of each one. We explored the impacts of DL architectures on GradCAM by using pneumothorax diagnosis from chest X-ray images as a case study. Pneumothorax is a condition that is characterized by the accumulation of air in the pleural space, and can lead to lung collapse, posing a significant risk to patient health if not promptly diagnosed and treated. As the radiological features of pneumothorax (i.e., air invasion) can be subtle to spot in X-ray scans, the task provides an excellent case to examine the characteristics of different DL models and Grad-CAM visualization. In summary, our work has two main contributions. **First**, we conducted a comprehensive evaluation of popular DL models including CNNs and Transformers for pneumothorax diagnosis. **Second**, we systematically compared the effectiveness of visual explanation using Grad-CAM across these staple DL models both qualitatively and quantitatively. Here, we analyzed the impact of network architecture choices on diagnostic accuracy and effectiveness of Grad-CAM results.

3.2 Related Works

The great accessibility of public chest X-ray datasets has allowed a large amount works [33] on the diagnosis and segmentation of lung diseases using deep learning algorithms, with a comprehensive review provided by Calli et al. [7]. So far, many previous reports adopted popular DL models that were first designed for natural image processing. For example, Tian et al. [54] leveraged ResNet and VGG models with multi-instance transfer learning for pneumothorax classification. Wollek at al. [60] employed the Vision Transformer to perform automatic diagnosis for multiple lung conditions on chest X-rays. To incorporate visual explanation for DL-based pneumothorax diagnosis, many have adopted Grad-CAM and its variants [60, 61] for both CNNs and ViTs. Yuan et al. [61] proposed a human-guided design to enhance the performance of Saliency Map, Grad-CAM, and Integrated Gradients in visual interpretability of pneumothorax diagnosis using CNNs. Most recently, Sun et al. [53] proposed the Attri-Net, which employed Residual blocks and multi-label explanations that align with clinical knowledge for improved visual explanation in chest X-ray classification.

3.3 Material and Methodology

3.3.1 Deep Learning Model Architectures

Our study explored a selection of staple deep learning architectures, including VGG, ResNet and ViT, which are widely used in both natural and medical images. Specifically, the VGG models [50] are characterized by their multiple 3x3 convolution layers, and for the study, we included VGG-16 and VGG-19. The ResNet models [18] leverage skip connections to enhance residual learning and training stability. Here, we incorporated ResNet18, ResNet34, ResNet50, and ResNet101, which comprise 18, 34, 50, and 101 layers, respectively. Lastly, the Vision Transformers initially proposed by Dosovitskiy et al.[12] treat images as sequences of patches/tokens to model their long-range dependencies without the use of convolution operations. To test the influence of network sizes, we adopted the ViT_small and ViT_base variants, both with 12 layers but differing in input features, and the ViT_large variant with 24 layers. All these models were pretrained on ImageNet-1K and subsequently fine-tuned for the task of Pneumothorax vs. Healthy classification using the curated public dataset.

3.3.2 Grad-CAM Visualization

To infer the decision-making process of DL models, the Grad-CAM technique [47] creates a heatmap by computing the gradients of the target class score with respect to the feature maps of the last convolutional layer. Specifically, for VGG16 and VGG19, we applied Grad-CAM to the last convolution layer. For ResNet models, we targeted the final bottleneck layer and in Vision Transformer variants, the technique was applied to the final block layer before the classification token is processed. Ideally, an effective visual guidance should provide high accuracy (i.e., correct identification of the region of interest by high values in the heatmap) and specificity (i.e., tight bound around the region of interest). Note that for each Grad-CAM heatmap, the value is normalized to [0,1].

3.3.3 Dataset Preprocessing and Experimental Setup

For this study, we used the SIIM-ACR Pneumothorax Segmentation dataset from Kaggle [1]. It contains chest X-Ray images of 9,000 healthy controls and 3,600 patients with pneumothorax. In addition, regions related to pneumothorax were manually segmented for 3576 patients. For our experiments, we created a balanced subset of 7,200 cases (50% with pneumothorax) from the original

[1] SIIM-ACR Pneumothorax Segmentation:https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation/data

dataset. From the curated data collection, we divided the cases into 7,000 for training, 1,000 for validation, and 1,000 for testing while balancing the health vs. pneumothorax ratio in each set. For DL model training and testing, each image was processed using Contrast Limited Adaptive Histogram Equalization (CLAHE) and normlaized with z-transform. In addition, all images were re-scaled to the common dimension of 224×224 pixels. In terms of training, the VGG and ResNet models utilized the cross-entropy loss function with the Adam optimizer, and were trained at a learning rate of 1e-4 for 50 epochs. The ViT models employed a cross-entropy loss function and the Stochastic Gradient Descent (SGD) method for optimization with a learning rate of 1e-4. They were trained for 300 epochs. Finally, to boost our model's performance and mitigate overfitting, we used data augmentations in training, including the addition of random Gaussian noise, rotations (up to 10 degrees), horizontal flips, and brightness/contrast shifts.

3.3.4 Evaluation Metrics

To evaluate the performance of different DL models in pneumothorax diagnosis, we assessed the accuracy, precision, recall, and area under the curve (AUC) metrics. In terms of assessing the effectiveness of the Grad-CAM results for each model, we propose to use two different measures. First, we compute the difference between the means of the Grad-CAM heatmap values within and outside the ground truth pneumothorax segmentation, and refer to this metric as $Diff_{GradCAM}$. We hypothesize that an effective Grad-CAM visualization should generate a high positive $Diff_{GradCAM}$ because the ideal heatmap should accumulate high values primarily within the pathological region (i.e., air invasion in the pleural space). The scores from the models were further compared by a one-way ANOVA test and Tukey's post-hoc analysis, and a p-value < 0.05 indicated a statistically significant difference. Second, we compute the Effective Heat Ratio (EHR) [60], which is the ratio between the thresholded area of the Grad-CAM heatmap within the ground truth segmentation and the total threshold area. The thresholds were computed in equidistant steps, and the Area Under the Curve (AUC) is calculated over all EHRs and the associated threshold values to assess the quality of Grad-CAM results. Both metrics reflect the accuracy and specificity of visual explanation for the networks.

3.4 Results

3.4.1 Pneumothorax Diagnosis Performance

The performance of pneumothorax diagnosis for all DL models is listed in Table 3.1. In terms of accuracy, ResNet models offered the best results, particularly with ResNet50 at 88.20%, and the ViT and VGG ranked the second and the last. The similar trend held for precision and AUC. However, as for recall, the obtained scores were similar across different architecture types. When looking into different network sizes for each architecture type, the results showed that deeper neural networks did not necessarily produce superior diagnostic performance. Specifically, the two popular VGG models didn't result in large discrepancy in accuracy, recall, and AUC while VGG16 has better precision. For the ResNet models, ResNet18, ResNet34 and ResNet101 had similar performance, with the accuracy, recall, and AUC peaked slightly at ResNet50. This likely means that ResNet18 has enough representation power to perform the classification, and therefore deeper networks do not improve the results. Finally, for the ViTs, ViT_base resulted in better performance than the small and large versions, with slight performance deterioration for ViT_large .

Table 3.1: Pneumothorax diagnosis performance across all DL models

Model	Accuracy	Precision	Recall	AUC
VGG16	83.70%	0.8224	0.8800	0.9069
VGG19	83.20%	0.7747	0.9080	0.9098
ResNet18	87.60%	0.8821	0.8680	0.9420
ResNet34	87.20%	0.8758	0.8600	0.9417
ResNet50	88.20%	0.8802	0.8820	0.9450
ResNet101	87.60%	0.8798	0.8780	0.9434
ViT Small	85.80%	0.8137	0.8820	0.9302
ViT Base	86.20%	0.8602	0.8740	0.9356
ViT Large	84.40%	0.8356	0.8640	0.9197

3.4.2 Qualitative Grad-CAM Evaluation

We present the Grad-CAM heatmaps for three patient cases across all tested DL models in Fig. 3.1, with the ground truth pneumothorax segmentation overlaid in the X-ray scans. In the presented cases, all the heatmaps correctly indicated the side of the lungs affected by the disease. However, the overlap with the pneumothorax segmentation varied. The VGG16 and VGG19 models pinpointed the pneumothorax areas for the first two cases while VGG19 failed to do so for the third case. Note that both models presented a secondary region. The ResNet18, 34, and 50 models successfully highlighted the problematic area, with the heatmap of ResNet18 slightly off-center, while the ResNet101 model showed activation in two regions. In comparison, ViT models exhibited more dispersed Grad-CAM patterns than the CNNs, and the amount of unrelated areas increased with the model size.

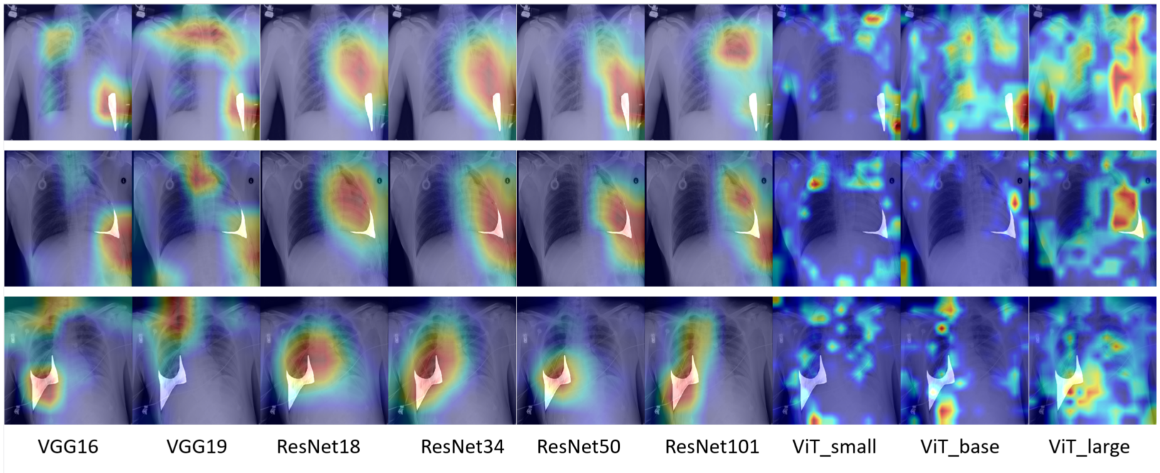


Figure 3.1: Demonstration of Grad-CAM results across different deep learning models for three patients X-ray images (one patient per row), with the manual pneumothorax segmentation overlaid in white color.

3.4.3 Quantitative Grad-CAM Evaluation

Table 3.2: Difference of mean Grad-CAM values within and outside the manual pneumothorax segmentation ($Diff_{GradCAM}$).

	VGG16	VGG19	ResNet18	ResNet34	ResNet50	ResNet101	ViT_small	ViT_base	ViT_large
mean	0.186	0.166	0.162	0.133	0.142	0.183	0.052	0.081	0.051
std	0.212	0.220	0.262	0.273	0.251	0.265	0.143	0.166	0.164
min	-0.213	-0.365	-0.366	-0.452	-0.287	-0.320	-0.251	-0.233	-0.319
max	0.773	0.715	0.787	0.782	0.801	0.780	0.652	0.786	0.685

For the Grad-CAM heatmap of each tested model, we computed the $Diff_{GradCAM}$ and EHR AUC metrics across all test cases to gauge their effectiveness of visually interpreting the decision-making process of DL algorithms. Here, $Diff_{GradCAM}$ and EHR AUC are reported in Table

3.2 and Fig. 3.2, respectively. For $Diff_{GradCAM}$, the ANOVA test showed a group-wise difference ($p<0.001$). In general, the CNNs offered better results than the ViT models, despite ViTs' good pneumothorax diagnosis accuracy. VGG16 and ResNet101 ranked the best ($p<0.05$), but the associated standard deviations of CNN models were also higher. Within each architecture type,the scores for VGG16 and VGG19 were similar ($p>0.05$), the ViT models also don't differ significantly ($p>0.05$), and ResNet101's score was significantly higher than ResNet34 ($p<0.05$). Between VGG and ResNet models, comparisons between ResNet18, ResNet50, VGG16, and VGG19 did not yield any significant differences ($p>0.05$). Among all the tested models, ResNet50 achieved the highest max score and ResNet34 had the lowest min score. As for EHR AUC, similar to the case of $Diff_{GradCAM}$, the CNN models also outperformed the ViT ones, with ResNet101 leading the scores at 0.0319 and VGG16 ranking the second at 0.0243. Compared with VGG16, the EHR AUC score of VGG19 was very similar. Among the ResNet variants, ResNet18, ResNet34, and ResNet50 scored similary in the range of 0.21$\sim$0.23, generally lower than those of VGG models. With the deepest architecture among the ResNet models, ResNet101 had a large increase of the EHR AUC metric even though its pneumothorax diagnosis accuracy was similar to the rest. For the ViT models, the EHR AUC improved gradually with the increasing size of the architecture, ranging from 0.0145 to 0.0171.

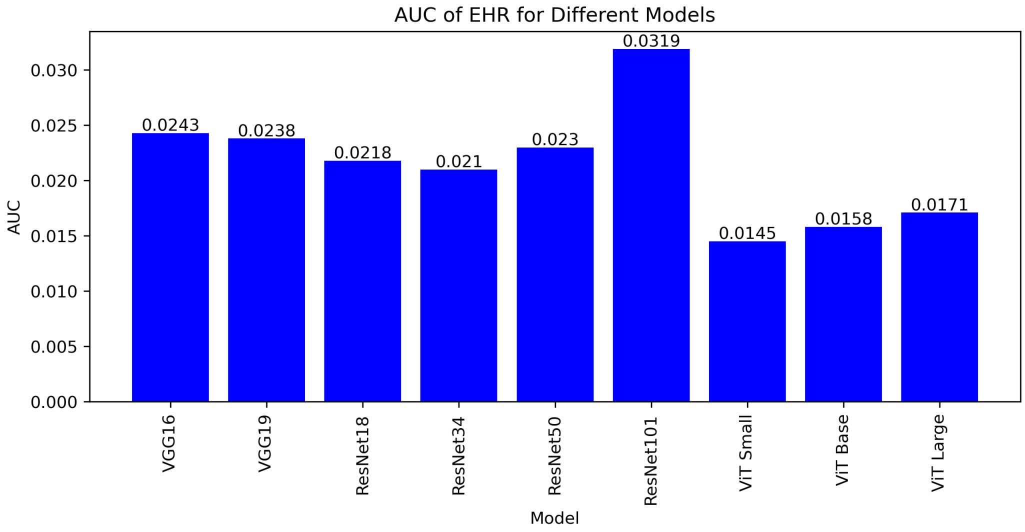


Figure 3.2: EHR AUC results of of different models

3.5 Discussion

Deep learning architectures like VGG, ResNet, and ViT all achieved commendable diagnostic performance in the range of 0.84$\sim$0.88%. As far as the experiments are concerned, deeper networks and the use of ViT do not contribute to better accuracy. Furthermore, with limited performance discrepancies between all tested models in pneumothorax diagnosis, the quality of Grad-CAM visualizations didn't necessarily correlate with the model's accuracy. The observed distinct behaviors of the Grad-CAM heatmaps may largely be due to the DL model's respective architecture types (pure CNN vs. Residual blocks vs. Transformer) and varying network sizes/depths. In our case study, while the Grad-CAM heatmaps of VGG16 and VGG19 accurately pinpointed areas of interest, VGG19 sometimes missed the pathological region, indicating that depth alone doesn't guarantee perfect feature localization. In addition, their activation maps often captured the upper chest areas, which may not be directly relevant for clinical interpretability. On the other hand, ResNet models, particularly ResNet18, 34, and 50, consistently highlighted relevant regions with a single cluster, albeit with slight deviations from the pneumothorax region in some cases. This could be attributed to the network's ability to focus on the most critical features through its residual connections. However, the much deeper ResNet101 model was prone to have two distinct areas in the Grad-CAM heatmaps. This was also noted by Seo et al. [48] in their vertebral maturation stage classification with ResNets, likely due to the fact that deeper architectures can capture more intricate representations [50]. In contrast, ViT models produced more dispersed Grad-CAM patterns compared to the included CNN ones. The inherent global contextual perception ability of Transformers might be responsible for this observation. As we only employed a single Chest X-ray dataset with a limited size, it's plausible that the ViTs could benefit from a larger dataset, potentially leading to better visualization outcomes [51]. Moreover, as the size of the ViT model increases, the proportion of irrelevant areas in the Grad-CAM visualizations also appears to increase, implying the interplay between model architecture and dataset size and the adverse cascading effects in deep Transformer models [63]. In previous investigations for DL-based computer-assisted diagnosis [28][48], multiple network models of different designs and natures were often benchmarked together. However, to the best of our knowledge, a systematic investigation for the impact of network architecture types and network depths on Grad-CAM visualization, especially with quantitative assessments has not been conducted to date. As transparency is becoming increasingly important for the safety and adoptiblity of DL algorithms, the relevant insights are of great importance to the XAI community.

The presented study has a few limitations. First, a typical issue in medical deep learning is the lack of large, well-annotated datasets. To facilitate the training process, we employed DL models that were pre-trained using natural images and then fine-tuned using domain-specific data. As noted by Lee et al. [28], in comparison to training from scratch, model fine-tuning also has a better advantage to provide more sparse and specific visualization for the target regions in Grad-CAM heatmaps. Furthermore, additional data augmentation was also implemented to mitigate overfitting issues. Second, as visual explanation of the DL algorithms is intended to allow easier incorporation of human participation, user studies to validate the quantitative metrics would be beneficial [42]. However, this requires more elaborate experimental design and inclusion of clinical experts, and will be included in our future studies. Lastly, we utilized pneumothorax diagnosis in chest X-ray as a case study to investigate the impact of DL model architectures on Grad-CAM visualization. It is possible that the observed trend may be application-specific. To confirm this, we will explore different datasets with varying disease types and imaging contrasts in the future.

3.6 Conclusion

In this study, we performed a comprehensive assessment of popular DL architectures, including VGG, ResNet and ViT, and their variants of different sizes for pneumothorax diagnosis in chest X-ray, and investigated the impacts of network depths and architecture types on visual explanation provided by Grad-CAM. While the accuracy for pneumothorax vs. healthy classification is similar across different models, the CNN models offer better specificity and accuracy than the ViTs when comparing the resulting heatmaps from Grad-CAM. Furthermore, the network size can affect both the model accuracy and Grad-CAM outcomes, with the two factors not necessarily in synch with each other. We hope the insights from our study can help better inform future explainable AI research, and we will further confirm the observations with more extensive studies involving more diverse datasets and DL models in the near future.

Chapter 4

Joint chest X-ray diagnosis and clinical visual attention prediction with multi-stage cooperative learning: enhancing interpretability

A version of this chapter was submitted to the 15th Machine Learning in Medical Imaging (MLMI) workshop co-hosted with the 27th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2024.

- Qiu, Z., Rivaz, H., & Xiao, Y. (2024). Joint chest X-ray diagnosis and clinical visual attention prediction with multi-stage cooperative learning: enhancing interpretability. arXiv preprint arXiv:2403.16970.[36]

4.1 Introduction

The continuous advancement of deep learning (DL) in medical imaging has led to a new era in diagnostic radiology, offering diagnostic tools with unprecedented accuracy and efficiency. Particularly in the realm of chest X-ray (CXR) analysis, where the imaging modality has poor soft tissue contrast, DL models have shown significant potential and efficiency in identifying and classifying various pulmonary conditions [16][25]. However, despite their prowess, the black-box nature of many DL-based computer-assisted diagnosis algorithms poses barriers for their clinical adoption, as medical professionals seek not only predictive accuracy, but also transparency and interpretability in automated clinical decisions. So far, there have been great efforts in explainable AI (XAI) to improve the credence of DL-based diagnostic algorithms. Notably, the flexible Class Activation Map (CAM) [46] and architecture-dependent attention maps from various attention mechanisms (e.g., self-attention [39] and attention gates [34]) have been widely adopted to provide human-attention-like visual interpretations for classification and regression tasks. Nevertheless, their resemblance to the real visual attention of radiologists, which could provide richer insights regarding the diagnostic process, is still limited. Recently, several DL models [23],[58] have been proposed to leverage recorded gaze data as an input or as an auxiliary task to enhance diagnostic accuracy. As gaze data collection requires specialized devices and postprocessing, the latter strategy in a multi-task learning framework appears more attractive in practice. Although multi-task learning could potentially boost the performance of individual tasks in synergy [64], the loss of each task can overfit at different rates if care is not taken, and may limit their performance.

To address the aforementioned challenges, we introduce a novel multi-task DL model based on a cooperative learning strategy to enhance classification performance and clinical visual saliency map prediction simultaneously for chest X-ray. Our contributions are threefold: **First**, to mitigate the issue of asynchronous training schedules of individual tasks, we proposed a multi-stage cooperative learning strategy, which helps with robust training of the designed multi-task network. **Second**, we designed a novel dual-encoder UNet with multi-scale feature-fusion to enhance multi-task collaboration for optimal outcomes. **Lastly**, our experiments demonstrated the superior performance of the proposed method than existing methods, and we share the trained model publicly.

4.2 Related Works

Public datasets have catalyzed data-driven advancements in chest X-ray image analysis, utilizing networks like ResNet, DenseNet, and EfficientNet for CXR-based disease classification [40].

Besides these advancements, with the potential benefits of enhancing the performance and/or interpretability of relevant DL models, a few groups have directly integrated tracked gaze data during diagnosis into their DL approaches. In GazeRadar, Bhattacharya et al. [4] fused radiomics and visual attention features to perform pulmonary disease classification. Later, the same group [5] proposed RadioTransformer that utilizes visual attention in a cascaded global-focal Transformer framework for CXR classification. Zhu et al. [65] used the visual attention maps to guide the class activation mapping to boost diagnostic accuracy. More recently, Wang et al. [58] proposed GazeGNN, which uses gaze duration to weigh local image features for CXR diagnosis through a graph neural network. Finally, Zhu et al. [64] proposed a multi-task UNet for joint CXR diagnosis and clinical attention map prediction by adopting an elaborate uncertainty-based loss balancing scheme with trainable parameters. In our study, we decided to tackle the challenge with a multi-stage cooperative learning strategy to boost the outcomes further.

4.3 Methods and Materials

An overview of our proposed technique is illustrated in Fig. 4.1, which is composed of a dual-encoder residual squeeze-and-excitation UNet (Res_SE-UNet) to predict visual saliency maps, a DenseNet-201 to encode image features, and a classifier module with multi-scale feature-fusion to provide CXR diagnosis. In our cooperative learning framework, these components were trained in three main stages, in the order of the DenseNet-201 feature encoder, residual squeeze-and-excitation UNet, and multi-scale feature-fusion classifier, to allow a gradual introduction of cooperation of the two learning tasks.

4.3.1 Stage 1: DenseNet-201 Feature Encoder

As DenseNets have shown great performance in image classification tasks, we used the DenseNet-201 as a key feature encoder for the designated tasks. To augment its robustness in feature representation, we first pretrained the network using the CXR data with a contrastive triplet loss [19], which minimizes the distances between similar image pairs while maximizing those between dissimilar ones. Then, the pretrained DenseNet201 was finetuned for the task of classifying a CXR scan into normal, pneumonia, or heart failure, with a cross-entropy loss. Here, we used all layers of the DenseNet-201 before the last dense block as our image feature encoder.

4.3.2 Stage 2: Visual Saliency Map Prediction

The objective of Stage 2 was to generate a visual saliency map for an input CXR scan that mirrors the attention patterns of medical professionals when diagnosing conditions. Following Stage 1, the feature from the trained DenseNet201 encoder was fed into a modified UNet model at the beginning of its decoding path, together with the compressed feature of the same input image from the UNet's encoding branch [43]. Note that here, we modified the UNet's encoder blocks with Residual and Squeeze-and-Excitation (SE) blocks (see Fig. 4.1) to enhance its robustness and training stability [29]. By leveraging features from two distinct image encoders to extract richer and more nuanced information from the CXR images, we intended to enhance the accuracy of visual saliency map prediction at the decoder end. During training, the DenseNet201 feature encoder was kept frozen, while the Res_SE-UNet was trained using paired CXR images and visual attention maps, and a Kullback–Leibler (KL) divergence loss.

4.3.3 Stage 3: Multi-Scale Feature-Fusion Classifier

In the final stage, we concatenated the feature from the DenseNet-201 encoder and that from the last upsampling layer of the Res_SE-UNet, and fed them into a simple CNN classifier (Fig. 4.1) to conduct CXR diagnosis (normal, pneumonia, or heart failure). This enabled full collaboration of two designated tasks. Here, both the DenseNet-201 encoder and Res_SE-UNet were frozen, and only the classifier was trained based on a cross-entropy loss.

4.3.4 Dataset and Evaluation Metrics

In our study, we used the "chest X-ray dataset with eye-tracking and report dictation" dataset [22] to develop and validate our proposed algorithm. The dataset contains 1083 chest X-ray scans from the MIMIC-CXR dataset [21], with their corresponding diagnostic results (normal, pneumonia, or heart failure), anatomical segmentation, radiologists' audio with dictation, and finally, additional eye-tracking data collected during radiological diagnosis. The eye-tracking data was obtained with a GP3 gaze tracker by Gazepoint (Vancouver, Canada), with an accuracy of around 1° of visual angle at a sampling rate of 60 Hz. For our study, we utilized the visual saliency maps that represent the gaze location and attention over time as a heat map. As the X-ray images have different resolutions and dimensions, we padded and resampled them to the resolution of 640 × 512 pixels, and normalized the pixel intensity to [0,1] per image.

For the proposed algorithm and comparison methods, we focus on the accuracy evaluation of CXR diagnosis (classification of normal, pneumonia, or heart failure) and visual saliency map prediction. Specifically, for diagnostic quality, we measured the area under the curve (AUC) metric for multi-class classification [15] and classification accuracy (ACC) for the overall classification performance, as well as AUCs for three individual classes. In terms of visual saliency prediction quality, we adopted the KL divergence, Pearson's correlation coefficient (PCC), and histogram similarity (HS), which measure the distribution similarity between the predicted and gaze heat maps. While a lower KL divergence is preferred, higher values for the rest of the metrics signify better results. For the metrics that assess the quality of visual saliency map production, two-sided paired sample t-tests were performed to confirm the superiority of our proposed technique against the comparison methods, and a p-value lower than 0.05 was used to declare statistical significance.

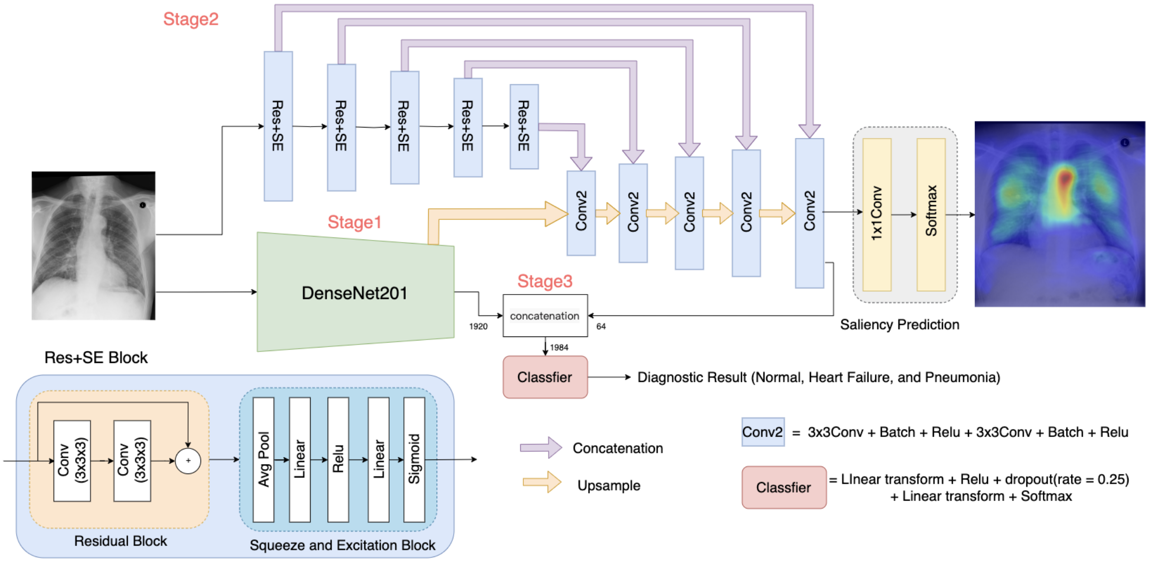


Figure 4.1: An overview of the proposed dual-encoder multi-task UNet architecture with multi-stage cooperative learning.

4.3.5 Experimental Setup and Ablation Studies

For our experiment, 100 CXR scans, along with the corresponding labels and visual saliency maps, were randomly selected as a test set while the remaining 983 samples were allocated for training purposes [58]. Our proposed method involved training at different stages, with the steps detailed in Section 3.1 $\sim$ 3.3. For each stage, the training process is conducted over 50 epochs. We utilized the Adam optimizer with its default parameters [24] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$) with a learning rate of 0.0001. Early stopping was employed to avoid overfitting, and model training was performed on a desktop computer with Intel Core i9 CPU, Nvidia GeForce RTX 3090 GPU, and 24GB RAM. To validate the performance of our proposed technique, we compared it against the previously established MT-UNet [64], GazeGNN [58], Inception ResNet v2 [14], and UNet [43] for the two target tasks, using the same training setup as our proposed method.

To help gain further insights into the individual elements of our proposed architecture, we conducted a series of four ablation studies. **First**, to investigate the benefit of adopting Residual and SE blocks in UNet for visual saliency prediction, we compared the heatmap generation quality between only using the Res_SE-UNet and a standard UNet (UNet_S), without the boosting from the DenseNet-201. **Second**, to confirm the contribution of the pretrained DenseNet-201 encoder, we further compared the saliency map prediction accuracy with and without the encoder module (full network vs. Res_SE-UNet). **Third**, the image classification accuracy of full network vs. DenseNet-201 with contrastive learning was investigated to inspect the benefit of multi-scale feature fusion. **Finally**, to reveal the advantage of contrastive learning for pretraining the DenseNet-201, we compared the results with and without contrastive learning pretraining for the network in CXR classification.

4.4 Results

4.4.1 Performance of the proposed method

The quantitative accuracy assessments of CXR classification and saliency map prediction for our proposed dual-encoder feature-fusion UNet and the comparison techniques. including the evaluations for ablation study purposes are presented in Tables 4.1 and 4.2. The arrows within the tables indicate the desired trends of the associated metrics. Overall, our proposed method has achieved an AUC of 0.925 and an accuracy of 80% for CXR diagnosis, outperforming all the comparison methods, including GazeGNN [58], which ranked as the second best method and uses both the X-ray scan and gaze data as inputs. When looking at the AUC results per class, our method achieved the best score for all class-wise AUCs, except for the category of Normal CXR. In terms of visual saliency prediction, our proposed technique significantly outperformed the MT-UNet and standard UNet ($p<0.01$) in all evaluation metrics while reaching better scores on average over our proposed Res_SE-UNet ($p>0.05$ for CC, $p<0.01$ for KL and HS). In addition, to qualitatively assess the visual saliency map generation, we present the results from our proposed method, against those produced by the MT-UNet [64], the gradCAM results from the DenseNet-201 component of our proposed DL architecture, and the ground truths. We can see that our proposed method has a higher resemblance to the ground truths while the gradCAM outputs have a large discrepancy from the human gaze pattern.

Table 4.1: Accuracy assessment of chest X-ray classification for our method, MT-UNet [64], DenseNet201 with contrastive pretraining (DNet201-CL), DesenNet201 (DNet201), Inception ResNet v2 (IRNetv2)[14], and GazeGNN [58].

Metric	**Ours**	MT-UNet	DNet201-CL	DNet201	IRNetv2	GazeGNN
AUC (Normal) ↑	0.953	0.935	0.961	**0.964**	0.878	0.899
AUC (Heart failure) ↑	**0.927**	0.881	0.865	0.897	0.873	0.881
AUC (Pneumonia) ↑	**0.894**	0.687	0.859	0.849	0.602	0.823
AUC ↑	**0.925**	0.847	0.889	0.880	0.794	0.868
Accuracy ↑	**0.800**	0.640	0.750	0.690	0.600	0.730

Table 4.2: Accuracy assessment of visual saliency map prediction (mean±std) for our method, MT-UNet [64], UNet with a modified Residual-Squeeze-and-Excitation encoder (Res_SE-UNet), and a standard UNet (UNet_S).

Metric	**Ours**	MT-UNet	Res_SE-UNet	UNet_S
KL ↓	**0.706 ± 0.183**	0.747 ± 0.185	0.747 ± 0.193	0.781 ± 0.202
CC ↑	**0.576 ± 0.113**	0.545 ± 0.109	0.560 ± 0.108	0.531 ± 0.109
HS ↑	**0.552 ± 0.055**	0.535 ± 0.055	0.539 ± 0.058	0.527 ± 0.056

4.4.2 Ablation studies

We performed four ablation studies for our proposed framework. **First**, when comparing the saliency map generation between the Res_SE-UNet and the standard UNet, we observed a significant improvement in all metrics ($p<0.05$), indicating the positive impact of SE and residual blocks to better capture task-relevant image features. **Second**, when incorporating the pretrained DenseNet-201 encoder into the Res_SE-UNet to form the dual-encoder setup (i.e., the full model), the quality of visual saliency prediction was further enhanced, with a noticeable drop of KL divergence from 0.747 to 0.706. **Third**, in terms of CXR classification, the full model of our proposed technique was compared against the DenseNet-201 pretrained with contrastive learning, and showed great improvement (Accuracy: 80% vs. 75%). Here, both the second and third studies showcased boosted performance by leveraging the collaboration of the classification and prediction tasks. **Lastly**, to confirm the benefit of contrastive pretraining for the DenseNet-201 feature encoder, the CXR classification performance of the network with and without CL pretraining were compared in Table 4.1. With the AUC and accuracy increases of 0.009 and 6%, respectively, the version with CL pretraining is shown to be superior.

4.5 Discussion

One major feature of our proposed DL framework is to employ a three-stage cooperative learning strategy to address the challenge in multi-task learning, where the loss functions of individual tasks fail to converge at the same rate, resulting in sub-optimal outcomes. Although Zhu et al. [64] successfully demonstrated the adoption of the uncertainty-based training strategy, it requires more elaborate setups and additional learnable parameters to achieve the best outcomes. In contrast, our strategy first used a DenseNet-201 feature encoder pretrained on CXR classification to facilitate visual saliency map prediction, and then further enhance the classification quality by allowing full

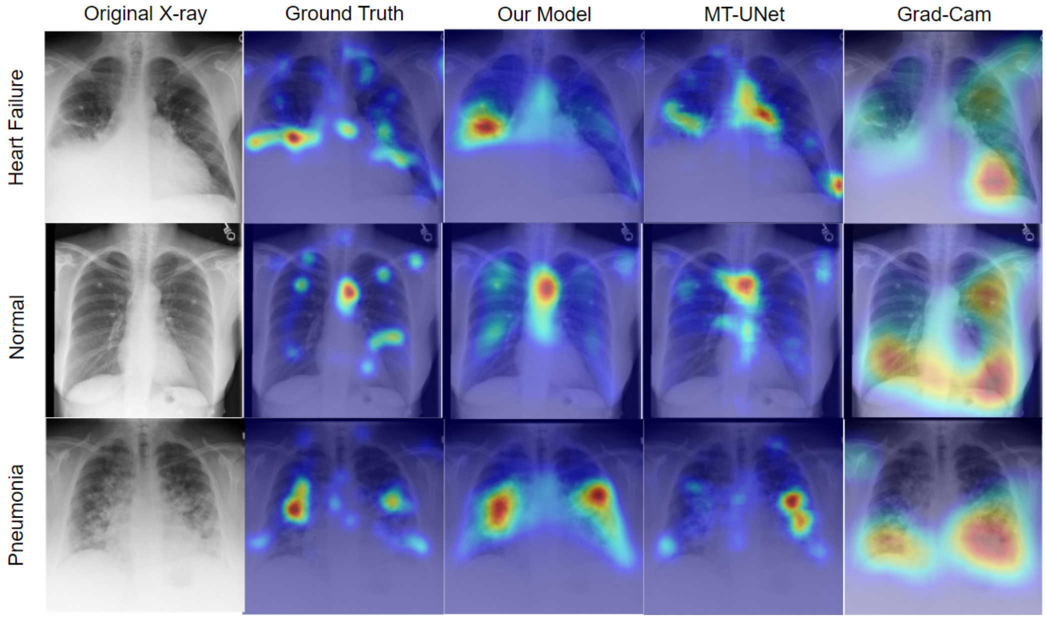


Figure 4.2: Qualitative evaluation of visual saliency map prediction across our proposed model and the MT-UNet [64], against the gradCAM outputs from the DenseNet-201 from our proposed model, the original X-ray scan and the ground truths for exemplary heart failure, normal, and pneumonia cases.

cooperation of the saliency map and DenseNet-201 features. Based on the ablation studies (see Section 4.2), we showed that such task collaboration with dual-encoder and multi-scale feature-fusion has boosted the performance. In addition, our proposed method outperformed the relatively recent MT-UNet [64] and GazeGNN [58], as well as other baseline techniques in the designated tasks. As in the related reports [64], [58], both MT-UNet and GazeGNN have been validated against several state-of-the-art DL models, including ResNets, EfficientNet, Swin Transformers, and VGGSSM [8], and showed better results, we decided not to expand the comparison to these models in our study.

In our ablation studies, we confirmed the positive role of contrastive learning with a triplet loss in pretraining the DenseNet-201 feature encoder. Although more recent self-learning techniques have provided better results, they could also be resource demanding. We will further explore these methods in the near future. For saliency map prediction, we proposed the Res_SE-UNet, with Residual and SE blocks. While the SE block can increase feature representation by dynamic channel-wise feature recalibration, the residual block facilitates gradient flow during training. This modification was proven to be instrumental in our method. As the gaze attention during diagnosis reflects the positions and importance of local features relevant to a diagnosis, such information may efficiently guide DL-based algorithms for radiological tasks. In previous investigations [59][45],[26][30], many have confirmed the constructive impacts of incorporating gaze attention maps on improving radiological diagnosis, either as input features with the medical scans or through auxiliary tasks (e.g., regularization of CAM results). This is also well echoed in our study. As shown in Fig. 4.2, although the highly flexible CAM has been widely used to provide visual explanation for various DL models, in terms of CXR diagnosis, the CAM pattern does not necessarily coincide with real human attentions. Interestingly, in Fig. 4.2, CAM highlighted the heart for the heart failure case while the human attention focuses on a broader area. Thus, it could be beneficial to employ both visual explanations jointly, and our proposed method can support this option.

In our current study, we utilized the visual attention map, which is an accumulation of temporally sampled gaze locations to provide a potential explanation for the diagnostic results. However, additional representation of the gaze pattern, such as the scanpaths can also be used to enrich the understanding of the diagnostic procedure, identify the skill levels of clinicians to ensure data quality [57], and further enhance the accuracy of the diagnostic algorithm to open doors for additional human-machine-interaction. Although it is still a challenging task and has not been adopted for clinical scans, ongoing scanpath prediction research in natural images [52] is progressing steadily. We will seek to explore venues to incorporate such information in radiological diagnosis in the future.

4.6 Conclusion

We have proposed a novel multi-task DL model using a dual-encoder UNet with multi-scale feature-fusion for CXR diagnosis and visual saliency prediction. The proposed DL model benefits from our multi-stage cooperative learning strategy to best optimize the training of individual tasks, with a gradual introduction of collaboration. With excellent results, our proposed technique tackles the important challenge of transparency in DL-based radiological diagnosis, with a great potential to be extended for other applications.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis has significantly advanced the field of interpretable medical image processing through two main contributions. The first contribution detailed in the MLMI 2023 paper involved a comprehensive evaluation of the Gradient-weighted Class Activation Mapping (Grad-CAM) across different deep learning architectures for pneumothorax diagnosis in chest X-ray images. This study highlighted the variability in Grad-CAM's effectiveness depending on the architecture depth and type, providing valuable insights into the adaptability and utility of visual explanation tools in medical diagnostics.

The second, and major, contribution presented in the MICCAI MLMI2024 workshop submission introduced a novel multi-task learning framework that effectively combines disease diagnosis with clinical visual attention prediction. This work utilized a dual-encoder multi-task UNet architecture, enhancing the interpretability of automated diagnostic decisions by aligning AI-generated saliency maps with clinician attention patterns. The introduction of a multi-stage cooperative learning strategy improved the synergy between these tasks, significantly boosting both diagnostic performance and the quality of visual explanations.

Both studies not only address the critical need for transparent and interpretable AI in medical imaging but also demonstrate practical implementations that could facilitate wider adoption in clinical settings. The improvements in both the reliability of visual explanations and the accuracy of automated diagnostics hold great promise for enhancing patient outcomes and supporting medical professionals in their decision-making processes.

5.2 Future Work

5.2.1 Expansion to Other Imaging Modalities:

First, future research should explore the applicability of the developed methodologies, such as Grad-CAM and the novel multi-task learning framework, to other imaging modalities like MRI, CT scans, and ultrasound. Extending these interpretability approaches to a broader range of imaging techniques could offer valuable insights into their generalizability and effectiveness across different clinical settings. This expansion would not only validate the robustness of the proposed methods but also enhance the scope and impact of AI-assisted diagnostics across various forms of medical imaging, potentially improving diagnostic accuracy and interpretability in a more extensive clinical context.

5.2.2 Integrating Cooperative and Federated Learning:

The concepts of cooperative learning, as discussed in this thesis, and federated learning both represent powerful strategies for enhancing AI systems. While cooperative learning focuses on leveraging the synergy between multiple related tasks to boost model performance and interpretability, federated learning offers a pathway to train models on decentralized datasets without compromising data privacy. Combining these approaches could lead to the development of robust, interpretable models that are trained on diverse, multi-institutional data while ensuring that each participating entity retains control over its own data. This integration promises significant advancements in creating AI tools that are both powerful and privacy-preserving, making them suitable for widespread adoption in diverse healthcare environments.[32]

5.2.3 Enhancing Model Interpretability:

A significant aspect of the research presented in this thesis is its focus on enhancing interpretability through a result-oriented, instance-based approach. This method models the decision-making process of AI systems on how experts, such as medical professionals, analyze and reason about medical images. By simulating human cognitive processes and decision patterns, the AI system's outputs become more intuitive and understandable to practitioners. This instance-based approach not only helps in aligning the AI's focus with that of human experts but also ensures that the outputs are relevant and easily interpretable in a clinical context.

Continuing to enhance the interpretability of AI systems remains a pivotal area for future research. Given the promising results of the instance-based approach in this thesis, further development and refinement of this methodology could yield even more robust and intuitive systems. Future work could explore:

- **Deepening Instance-based Methodologies:** Enhancing the current models to more closely mimic complex human reasoning processes, potentially incorporating more nuanced aspects of expert decision-making to improve the depth and accuracy of the interpretability.

- **Exploring Alternative Models of Explainability:** While instance-based methods are effective, diversifying interpretability approaches could include developing complementary techniques such as counterfactual explanations, rule-based systems, or causal inference models. These models could provide different perspectives on AI decisions, offering a more comprehensive understanding of model behavior.

- **Hybrid Approaches:** Combining instance-based methods with other interpretability frameworks could provide a more holistic view of AI decisions. This hybrid approach would leverage the strengths of various interpretability methods, ensuring that AI systems are not only effective but also fully accountable and transparent in their operations.

Bibliography

[1] Ahmad Alsaleh and Cahit Perkgoz. A space and time efficient convolutional neural network for age group estimation from facial images. *PeerJ Computer Science*, 9:e1395, 2023.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[3] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[4] Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna. Gazeradar: A gaze and radiomics-guided disease localization framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 686–696. Springer, 2022.

[5] Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna. Radiotransformer: a cascaded global-focal transformer for visual attention–guided disease classification. In *European Conference on Computer Vision*, pages 679–698. Springer, 2022.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[7] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021.

[8] Ge Cao, Qing Tang, and Kang-hyun Jo. Aggregated deep saliency prediction by self-attention network. In *Intelligent Computing Methodologies: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part III 16*, pages 87–97. Springer, 2020.

[9] Danton S Char, Nigam H Shah, and David Magnus. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981, 2018.

[10] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[11] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Kelly N DuBois. Deep medicine: how artificial intelligence can make healthcare human again. *Perspectives on Science and Christian Faith*, 71(3):199–201, 2019.

[14] Khalid El Asnaoui, Youness Chawki, and Ali Idri. Automated methods for detection and classification pneumonia based on x-ray images using deep learning. In *Artificial intelligence and blockchain for future cybersecurity applications*, pages 257–284. Springer, 2021.

[15] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[16] Shimpy Goyal and Rajiv Singh. Detection and classification of lung diseases for pneumonia and covid-19 using machine and deep learning techniques. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3239–3259, 2023.

[17] Ameer Hamza, Muhammad Attique Khan, Shui-Hua Wang, Majed Alhaisoni, Meshal Alharbi, Hany S Hussein, Hammam Alshazly, Ye Jin Kim, and Jaehyuk Cha. Covid-19 classification using chest x-ray images based on fusion-assisted deep bayesian optimization and grad-cam visualization. *Frontiers in Public Health*, 10:1046296, 2022.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015.

[20] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[21] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317, 2019.

[22] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific data*, 8(1):92, 2021.

[23] Sota Kato, Masahiro Oda, Kensaku Mori, Akinobu Shimizu, Yoshito Otake, Masahiro Hashimoto, Toshiaki Akashi, and Kazuhiro Hotta. Classification and visual explanation for covid-19 pneumonia from ct images using triple learning. *Scientific Reports*, 12(1):20840, 2022.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284 (2):574–582, 2017.

[26] Ricardo Bigolin Lanfredi, Ambuj Arora, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays. *arXiv preprint arXiv:2112.11716*, 2021.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[28] Yeon-Hee Lee, Jong Hyun Won, Seunghyeon Kim, Q-Schick Auh, and Yung-Kyun Noh. Advantages of deep learning with convolutional neural network in detecting disc displacement of the temporomandibular joint in magnetic resonance imaging. *Scientific Reports*, 12(1):11352, 2022.

[29] Hao Liu, Jiancheng Luo, Bo Huang, Haiping Yang, Xiaodong Hu, Nan Xu, and Liegang Xia. Building extraction based on se-unet. *Journal of Geo-Information Science*, 2019.

[30] André Luís, Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Anderson Maciel, Joaquim Jorge, and Catarina Moreira. Integrating eye-gaze data into cxr dl approaches: A preliminary study. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 196–199. IEEE, 2023.

[31] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[32] Priyanka Mary Mammen. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*, 2021.

[33] Maad M Mijwil. Implementation of machine learning techniques for the classification of lung x-ray images used to detect covid-19 in humans. *Iraqi Journal of Science*, pages 2099–2109, 2021.

[34] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[35] Zirui Qiu, Hassan Rivaz, and Yiming Xiao. Is visual explanation with grad-cam more reliable for deeper neural networks? a case study with automatic pneumothorax diagnosis. In *International Workshop on Machine Learning in Medical Imaging*, pages 224–233. Springer, 2023.

[36] Zirui Qiu, Hassan Rivaz, and Yiming Xiao. Joint chest x-ray diagnosis and clinical visual attention prediction with multi-stage cooperative learning: enhancing interpretability. *arXiv preprint arXiv:2403.16970*, 2024.

[37] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[38] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

[39] Amirhossein Rasoulian, Soorena Salari, and Yiming Xiao. Weakly supervised intracranial hemorrhage segmentation using head-wise gradient-infused self-attention maps from a swin transformer in categorical learning. *arXiv preprint arXiv:2304.04902*, 2023.

[40] Vinayakumar Ravi, Harini Narasimhan, Chinmay Chakraborty, and Tuan D Pham. Deep learning-based meta-classifier approach for covid-19 classification using ct scan and chest x-ray images. *Multimedia systems*, 28(4):1401–1415, 2022.

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.

[42] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: user studies for model explanations. *arXiv preprint arXiv:2210.11584*, 2022.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[44] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[45] Khaled Saab, Sarah M Hooper, Nimit S Sohoni, Jupinder Parmar, Brian Pogatchnik, Sen Wu, Jared A Dunnmon, Hongyang R Zhang, Daniel Rubin, and Christopher Ré. Observational supervision for medical image classification using gaze data. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 603–614. Springer, 2021.

[46] Pranav Kumar Seerala and Sridhar Krishnan. Grad-cam-based classification of chest x-ray images of pneumonia patients. In *Advances in Signal Processing and Intelligent Recognition Systems: 6th International Symposium, SIRS 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 6*, pages 161–174. Springer, 2021.

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[48] Hyejun Seo, JaeJoon Hwang, Taesung Jeong, and Jonghyun Shin. Comparison of deep learning models for cervical vertebral maturation stage classification on lateral cephalometric radiographs. *Journal of Clinical Medicine*, 10(16):3591, 2021.

[49] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[51] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[52] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. Scandmm: A deep markov model of scanpath prediction for 360° images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[53] Susu Sun, Stefano Woerner, Andreas Maier, Lisa M Koch, and Christian F Baumgartner. Inherently interpretable multi-label classification using class-specific counterfactuals. *arXiv preprint arXiv:2303.00500*, 2023.

[54] Yuchi Tian, Jiawei Wang, Wenjie Yang, Jun Wang, and Dahong Qian. Deep multi-instance transfer learning for pneumothorax classification in chest x-ray images. *Medical Physics*, 49 (1):231–243, 2022.

[55] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[57] Stephen Waite, Arkadij Grigorian, Robert G Alexander, Stephen L Macknik, Marisa Carrasco, David J Heeger, and Susana Martinez-Conde. Analysis of perceptual expertise in radiology–current knowledge and a new perspective. *Frontiers in human neuroscience*, 13:213, 2019.

[58] Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Ahmet Cetin, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. Gazegnn: A gaze-guided graph neural network for disease classification. *arXiv preprint arXiv:2305.18221*, 2023.

[59] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 41(7):1688–1698, 2022.

[60] Alessandro Wollek, Robert Graf, Saša Čečatka, Nicola Fink, Theresa Willem, Bastian O Sabel, and Tobias Lasser. Attention-based saliency maps improve interpretability of pneumothorax classification. *Radiology: Artificial Intelligence*, 5(2):e220187, 2022.

[61] Han Yuan, Peng-Tao Jiang, and Gangming Zhao. Human-guided design to explain deep learning-based pneumothorax classifier. In *Medical Imaging with Deep Learning, short paper track*, 2023.

[62] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[63] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *ArXiv*, abs/2103.11886, 2021.

[64] Hongzhi Zhu, Robert Rohling, and Septimiu Salcudean. Multi-task unet: Jointly boosting saliency prediction and disease classification on chest x-ray images. *arXiv preprint arXiv:2202.07118*, 2022.

[65] Hongzhi Zhu, Septimiu Salcudean, and Robert Rohling. Gaze-guided class activation mapping: leveraging human attention for network attention in chest x-rays classification. *arXiv preprint arXiv:2202.07107*, 2022.