The Thesis Committee for Hari Priya Kandasamy

certifies that this is the approved version of the following Thesis.

Enhancing Chest X-ray Analysis: A Comparative Study of Deep Learning Models with Explainable AI

APPROVED BY

SUPERVISING COMMITTEE:

Ying Ding, Supervisor

Abhijit Mishra, Co-Supervisor

Enhancing Chest X-ray Analysis: A Comparative Study of Deep Learning Models with Explainable AI

by

Hari Priya Kandasamy

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Information Studies

The University of Texas at Austin

May 2023

Dedication

Dedicated to my parents, Prema and Kandasamy.

Acknowledgements

I would like to express my heartfelt gratitude to the following individuals for their unwavering support and assistance throughout my academic journey.

First and foremost, I would like to thank my supervisor, Dr. Ying Ding, and co-supervisor, Dr. Abhijit Mishra, for their invaluable guidance, patience, and encouragement. Their expertise and mentorship have been instrumental in the successful completion of this thesis.

I am profoundly grateful to my loving parents, who have always believed in me and supported my dreams. Their sacrifices, love, and unwavering belief in me have been my guiding light that enabled me to face the challenges and triumphs of studying abroad.

To my wonderful sister and supportive brother-in-law, who have been my anchors throughout this journey, offering invaluable assistance as I adapted to life away from home. Their unwavering presence, wise guidance, and heartfelt encouragement have served as a source of solace, keeping me grounded and providing stability during turbulent times. My friends and colleagues, both near and far, deserve special mention for their constant support, motivation, and helpful feedback.

I am also grateful to the staff and faculty at The University of Texas at Austin, who provided me with the necessary resources and facilities. Their assistance and commitment to fostering a supportive academic environment have been truly appreciated.

Abstract

Enhancing Chest X-ray Analysis: A Comparative Study of Deep Learning Models with Explainable AI

Hari Priya Kandasamy, M.S. Info.Stds

The University of Texas at Austin,2023

Supervisor : Ying Ding
Co-supervisor : Abhijit Mishra

Analyzing medical images is a vital aspect of contemporary healthcare. While conventional deep learning models have demonstrated potential in this area, their interpretability remains a challenge, especially for intricate medical images like chest X-rays. In our study, we assess the performance of advanced deep learning models, such as InceptionV4 and Vision Transformers, in comparison to traditional models like ResNet50 to determine their effectiveness in classifying chest X-ray images. We use two comprehensive datasets, NIH-CXR-LT and MIMIC-CXR-LT, for evaluating the models' performance.

Our goal is to improve the comprehensibility and usability of these models by implementing explainable AI techniques for visualizations, contributing to the creation of user-friendly AI tools for medical imaging. Our findings indicate that Vision Transformers attain higher AUC scores while reducing training time compared to other models. By employing explainable AI methods like GradCAM and SHAP, we showcase the interpretability of the models, allowing us to pinpoint areas of interest in chest X-ray images that the models depend on for their predictions. These techniques also assist in detecting model biases and recognizing potential errors, which can be crucial for informed clinical decision-making.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1. Introduction

The widespread integration of artificial intelligence (AI) across various fields has resulted in notable progress, particularly in medical imaging. Pioneering studies by Benjio et al. (1993) and Yamashita et al. (2018) have shown that deep learning techniques, such as convolutional neural networks, can improve the precision and effectiveness of medical imaging processes. Inception and Vision Transformer architectures, as outlined by Szegedy et al. (2016) and Lu et al. (2019), have emerged as potential alternatives to traditional CNNs. These advanced models are known for their exceptional performance in a variety of image classification tasks, motivating researchers to investigate their potential for medical imaging applications.

However, the increasing complexity of AI models, including Inception and Vision Transformers, raises questions about their interpretability and transparency. The "black-box" nature of these models makes it difficult to comprehend their decision-making processes, which is essential in medical imaging applications where trust, accountability, and dependability are crucial. As a result, there is a growing interest in explainable AI (XAI) in medical imaging, with the goal of providing healthcare professionals and patients with insights into the decision-making processes of AI models.

In this thesis, we aim to examine whether Inception and Vision Transformer models can surpass traditional CNN models in the context of medical imaging. By comparing the performance of these models on various medical imaging datasets, the study seeks to identify potential advantages and limitations associated with each approach. Our research

also investigates the application of XAI techniques to medical imaging, with an emphasis on enhancing interpretability, trust, and transparency for healthcare professionals and patients. This will contribute to improved patient outcomes and increased confidence in AI-driven diagnostic solutions.

1.1 Background of AI in Medical Imaging

Imaging Artificial Intelligence (AI) has significantly influenced the medical imaging field in recent years. With the growing availability of large volumes of medical imaging data, AI algorithms have been developed and utilized to enhance the accuracy and speed of image analysis. One of the first applications of AI in medical imaging was the development of computer-aided detection (CAD) systems, as illustrated in Fig. 1.1 (Doi et al., 2007). These systems were created to assist radiologists in identifying abnormalities in medical images by employing machine learning algorithms to analyze images and emphasize regions of interest for further examination by a radiologist.

Fig. 1.1 Computer-aided diagnosis (Tao et al. 2023)

Computer-aided diagnosis (CADx) systems have further developed the capabilities of AI in medical imaging (Ginneken et al. 2011). These systems not only detect abnormalities but also provide diagnoses based on image analysis. For instance, AI algorithms can be trained to recognize specific types of cancers or other diseases by identifying patterns in medical images (Hosny et al. 2018).

Over recent decades, AI has progressed from rule-based systems and expert systems to more advanced machine learning and deep learning algorithms, such as convolutional neural networks (CNNs) (LeCun et al. 2015). These algorithms, trained on extensive medical image datasets, have demonstrated superior performance compared to traditional image analysis techniques (Litjens et al. 2017). This improvement has facilitated breakthroughs across various medical imaging applications, including cancer detection and diagnosis, lung disease identification, cardiovascular disorder analysis, and other conditions (Shen et al. 2017).

Although AI has made significant strides in medical imaging, challenges remain in its widespread adoption within healthcare. One major issue is the "black box" nature of AI algorithms, which can be difficult to comprehend and trust (Castelvecchi et al. 2016). This has led to concerns among medical professionals and patients regarding the accuracy, fairness, and ethics of AI-based diagnoses. To address these concerns, there is a growing demand for more understandable and explainable AI, referred to as "Explainable AI" (Arrieta et al. 2020). If successful, AI in medical imaging could result in earlier and more precise diagnoses, reduced invasive procedures, and a more efficient healthcare system, ultimately benefiting patients (McKenzie et al. 2020)

1.2 The Role of Explainable AI

The use of artificial intelligence (AI) in medical imaging has the potential to greatly improve patient outcomes, but as AI models become more complex, it becomes more difficult to understand how they make decisions. This lack of transparency can be a problem in the medical field, where trust and accountability are critical. To address this issue, a field called Explainable AI (XAI) has emerged. The goal of XAI is to make AI models more interpretable, meaning their decision-making processes are transparent and easily understood by humans.

In summary, XAI is important in medical imaging because it helps ensure that AI models are transparent, trustworthy, and reliable, ultimately leading to better patient outcomes.

1.3 Motivation for the Research

The accurate interpretation of chest X-rays is essential for early detection and diagnosis of numerous diseases, including pneumonia, tuberculosis, and lung cancer. However, manual analysis of these images is time-consuming and prone to human error. Consequently, there is a growing demand for automated methods that can assist healthcare professionals in the interpretation process.

Deep learning, specifically Convolutional Neural Networks (CNNs), has revolutionized the field of medical image analysis by providing state-of-the-art performance in various tasks, including chest X-ray interpretation. Despite their success, more advanced models have the potential to further improve diagnostic accuracy.

By conducting a comprehensive comparison of the architectures and investigating their explainability, this thesis aims to advance the field of medical image analysis and contribute to the development of more effective and transparent diagnostic tools. Explainability in AI, also referred to as explainable AI (XAI), refers to the ability of AI systems to provide clear, understandable, and human-interpretable explanations for their decisions and actions (Holzinger et al. 2019; Gilpin et al. 2018). This is particularly important in medicine, where understanding the underlying reasoning behind AI-generated diagnoses is crucial for fostering trust and facilitating informed decision-making by healthcare professionals (Arrieta et al. 2020; Ribeiro et al. 2016).

1.4 Objectives of the Research

This research explores the application of XAI techniques in medical imaging, with a focus on chest X-ray interpretation, aiming to achieve the following objectives. The overview is given in Fig 1.2

1. Review the current state of AI in chest X-ray analysis, including the successes and challenges faced by deep learning models in this domain.
2. To demonstrate the utility of these models in real-world scenarios by applying them to different chest X-ray datasets.
3. To assess the explainability of these models using heatmap and bounding box visualizations, which can provide insights into the decision-making process and help establish trust in the models' predictions.

The model pipeline overview in this context involves processing chest X-ray images through a deep learning model to generate disease predictions or labels, and then applying explainable AI techniques to provide interpretable insights into the model's decision-making.

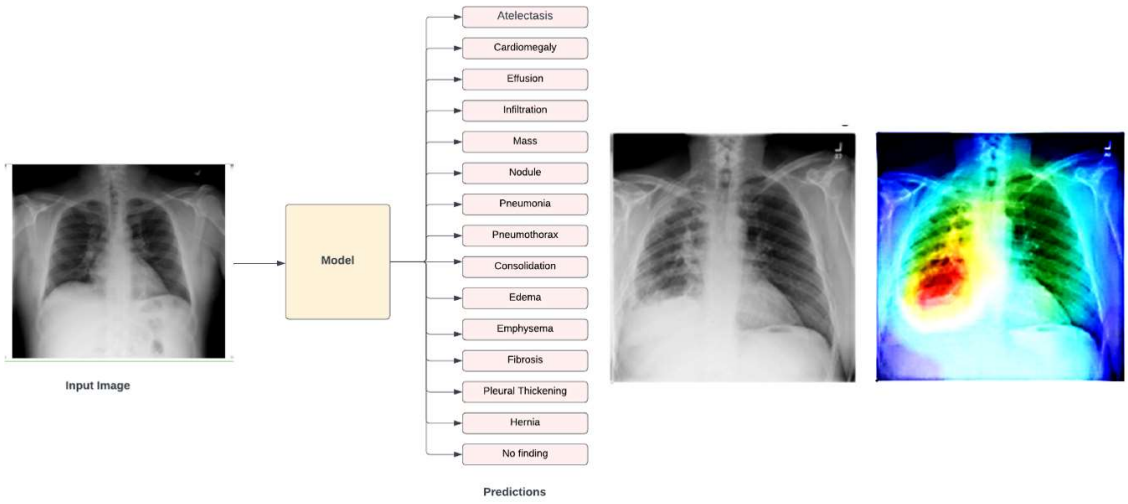


Fig 1.2 Overview of the model pipeline

Chapter 2: Related Work

2. Chest x-ray Interpretation: Current State and Challenges

Chest x-ray (CXR) imaging is one of the most widely used diagnostic tools in medical practice, primarily due to its non-invasive nature, low radiation exposure, and relatively low cost. CXRs can reveal various lung and heart conditions, such as pneumonia, lung cancer, tuberculosis, and congestive heart failure. Accurate interpretation of chest x-rays plays a crucial role in diagnosing and managing these conditions.

Traditionally, radiologists have been responsible for interpreting chest x-rays. However, the growing demand for medical imaging services and the shortage of radiologists have led to an increased interest in the development of automated methods for chest x-ray interpretation. In recent years, computer-aided detection (CAD) systems and artificial intelligence (AI) techniques, specifically deep learning, have shown potential in assisting radiologists in their tasks.

2.1 Deep learning for chest X-ray image analysis

Deep learning has become an increasingly popular approach for image classification tasks in recent years, with convolutional neural networks (CNNs) being the most widely used architecture. However, there have been recent efforts to investigate the effectiveness of other architectures for image classification tasks.

In recent years, deep learning has revolutionized the field of chest X-ray image analysis, enabling automated detection of various diseases with impressive accuracy. Lakhani and Sundaram (2017) highlighted the potential of deep learning in this area, with some models even surpassing radiologist performance. CheXNet by Rajpurkar et al. (2017)

employed a 121-layer DenseNet and outperformed radiologists in detecting pneumonia using the ChestX-ray14 dataset. The results are shown in Fig 2.1

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Fig 2.1 Comparing radiologists and CheXNet on the F1 metric. (Rajpurkar et. al 2017)

Litjens et al. (2017) provided an overview of the use of deep learning techniques in medical imaging, highlighting the potential benefits and challenges of using these techniques in practice. “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases" by Xiaosong Wang comprises 108,948 frontal-view X-ray images of 32,717 unique patients with text-mined eight disease image labels. This study shows that it's possible to identify and locate common chest diseases using a framework that combines weak supervision, multi-label image classification, and disease localization. However, using deep convolutional neural networks to "read" chest X-rays remains a challenging task for fully automated, highly accurate computer-aided diagnosis (CAD) systems.

2.2 Traditional CNN models in Chest X-ray Analysis

Krizhevsky et al. (2012) introduced AlexNet, a traditional CNN model. Shin et al. (2016) reviewed various CNN-based architectures and their applications in medical image analysis, including chest X-ray analysis. Jiashi Zhao et al. (2020) proposed a deep learning method using the AM DenseNet model for classifying chest X-ray images, which achieved an average AUC value of 0.8537. The authors demonstrated the effectiveness of the CBAM attention mechanism and Focal Loss function in enhancing the model's performance.

Rajpurkar et al. (2017) presented CheXNet, a 121-layer convolutional neural network that diagnoses pneumonia and localizes affected areas in chest X-ray images. CheXNet achieved an F1 score of 0.435, higher than the radiologist average of 0.387, and outperformed the best published results on all 14 diseases in ChestX-ray14.

Ikechukwu et al. (2021) proposed a deep learning method to classify chest X-ray images as normal or pneumonia using pretrained models (VGG-19 and ResNet-50) and a custom convolutional neural network. The pretrained models showed superior performance compared to training from scratch, and the proposed model outperformed existing state-of-the-art models as shown in Fig 2.2.

The Thorax-Net is a deep convolutional neural network that uses attention mechanisms to classify 14 thoracic diseases in chest radiography. The authors suggest that incorporating visual attention mechanisms into deep models can help them focus on abnormal regions of images, improving both performance and interpretability of thorax disease diagnosis. Thorax-Net uses ResNet-152 as the classification branch and Grad-CAM embedded into stacked convolutions as the attention branch. The model combines

the outputs of both branches to diagnose each input, and it achieved state-of-the-art performance on the ChestX-ray14 dataset

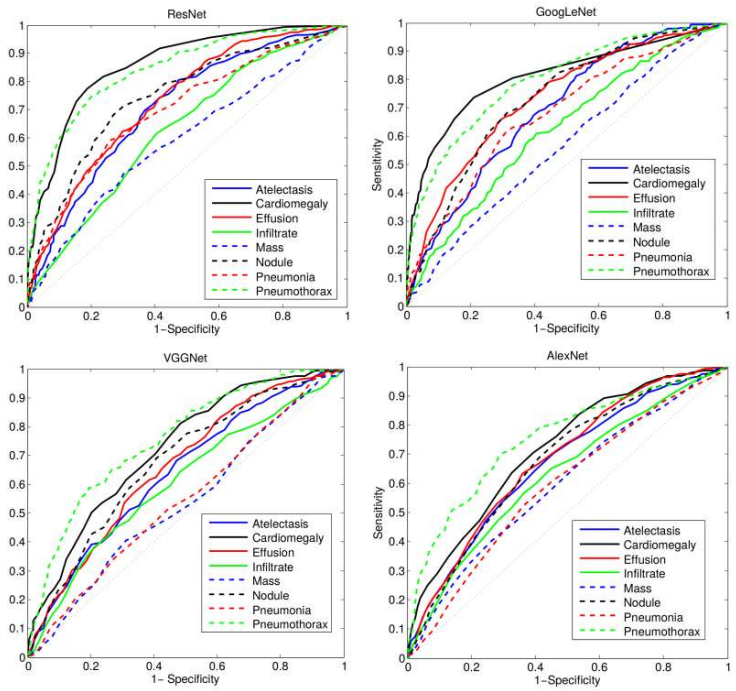


Fig 2.2 Comparison of ResNet,GoogleNet, VGGNet and AlexNet in multiclassification of Chest Xray Interpretation (Wang et. al 2017)

. The authors believe that attention-based deep models can improve medical image analysis and diagnosis by localizing and recognizing multiple objects while identifying the major contributors to the network output. However, recent advances in deep learning models such as Inception and Vision Transformer (ViT) may provide better performance and scalability for chest X-ray analysis, warranting further investigation.

2.3 Inception models

Inception models, first proposed by Szegedy et al. (2015), have demonstrated exceptional performance in image recognition tasks. These models have been successfully applied to medical imaging as well. For instance, Gulshan et al. (2016) used an Inception model to effectively detect diabetic retinopathy. This success indicates that Inception models may have significant potential in the realm of chest X-ray analysis.

The Inception model improves efficiency and accuracy in image classification by using multiple layers of convolutional filters with varying kernel sizes, allowing it to capture features at different scales. In the context of chest X-ray analysis, a modified Inception model has been shown to outperform radiologists in detecting pneumonia (Rajpurkar et al., 2017).

2.4 Vision Transformer models

The Vision Transformer (ViT) model, introduced by Dosovitskiy et al. (2020), is another promising architecture for chest X-ray image analysis. Instead of convolutions, ViT models utilize self-attention mechanisms, achieving competitive performance in image recognition tasks. In the medical imaging domain, Chen et al. (2021) proposed TransUNet, a hybrid ViT model for medical image segmentation, demonstrating the ViT's adaptability to such tasks.

ViT, initially developed for language translation tasks, has been adapted for computer vision tasks. It employs self-attention mechanisms to enhance image

classification performance, and has demonstrated effectiveness in object recognition and natural language processing. Recent studies, such as Yang et al. (2022) and Li et al. (2022), have proposed enhancements to ViT specifically for chest X-ray image classification, demonstrating superior performance compared to traditional CNN-based models.

Among the emerging deep learning models, the Vision Transformer (ViT) has gained significant attention due to its superior performance on various image recognition tasks. Yang et al. (2022) proposed an enhanced vision transformer architecture called IEViT (Improved and Efficient Vision Transformer) specifically designed for chest X-ray image classification. The authors addressed the limitations of the original ViT, such as the large number of parameters and computational complexity, by introducing a modified self-attention mechanism and a feature pyramid network (FPN). The proposed IEViT model demonstrated superior performance in terms of accuracy and computational efficiency when compared to traditional CNN-based models, as well as the original ViT, on the COVIDx dataset. Li et al. (2022) introduced the Local Vision Transformer (LVT), which combines the advantages of both CNNs and transformers. LVT incorporates a local perception module that captures local spatial information through small convolutional kernels. This addition mitigates the ViT's difficulty in processing fine-grained information, which is particularly relevant for medical images. The LVT model achieved state-of-the-art performance on various medical image classification tasks, including chest X-ray-based COVID-19 detection.

Wang et al. (2023) presented a novel approach for chest X-ray interpretation called XRAI. The authors combined a ViT-based model with a contrastive learning framework to learn discriminative visual features without explicit supervision. The XRAI model demonstrated competitive performance in detecting COVID-19, pneumonia, and other lung abnormalities from chest X-rays, even when compared to methods that rely on labeled data. Additionally, the model provided human-interpretable explanations for its predictions, which could contribute to increased trust in AI-assisted diagnostics.Singh et al. (2022) explored the use of transformers in a multi-modal context by proposing the Multi-Modal Medical Image Transformer (MMMIT). The MMMIT model leverages both visual and textual information from chest X-ray images and associated radiology reports to improve classification accuracy. The authors demonstrated that the proposed model outperforms both CNN-based and transformer-based unimodal approaches in detecting lung diseases, highlighting the potential of multi-modal learning in medical image analysis.

2.5 Explainability and heatmaps in deep learning

Zeiler and Fergus (2014) introduced neural network activation visualization, a technique that uses deconvolutional networks to map feature activations back to the input pixel space. This method helps visualize the hierarchical structure of features learned by the model and provides insight into the model's decision-making process.

Grad-CAM is a strategy introduced by Zhang et al. (2018) to create visually interpretable heat maps. It operates by analyzing the gradients of the target class that flow into the ultimate convolutional layer to produce a rudimentary localization map that highlights

crucial areas in the input image that are vital for predicting the class. Unlike previous methods, Grad-CAM can be used across a diverse range of CNN architectures and offers superior visual explanations.

Spatially-sensitive pooling (SSP), developed by Wang et al. (2016), is a method that generates interpretable heatmaps and bounding boxes. SSP uses a spatial pyramid structure to aggregate local features into global features, preserving spatial information at multiple scales. This approach enables the identification of discriminative regions in the input image and generates interpretable visualizations.

Rajpurkar et al. (2017) integrated Grad-CAM into the CheXNet model, which helped provide visual explanations for the model's predictions and increased trust in AI-assisted diagnostics.

While current state-of-the-art methods in this domain predominantly include deep learning models like DenseNet and traditional CNNs, the capabilities of Inception and ViT models have yet to be fully explored in medicsl Imaging domain. Given their performance in other imaging tasks, these models may offer significant improvements in chest X-ray analysis.Moreover, our research emphasizes the importance of explainability in deep learning models, particularly in medical applications. Techniques such as Grad-CAM play a crucial role in ensuring model interpretability and trustworthiness.

Chapter 3: Methods

3.1 Datasets

In this study, we utilized two chest x-ray datasets: the NIH Chest X-rays and the MIMIC-CXR-LT. Both datasets were chosen due to their comprehensive nature, large size, and diverse patient populations, making them suitable for training and evaluating deep learning models for chest x-ray interpretation. Table 3.1 contains the statistics of both the datasets.

Table 3.1 Dataset Statistics

Dataset	Number of Images	Image Size	Modality	Labels
NIH Chest X-ray	112,120	1024 x 1024 pixels	Digital Radiography	14 labels
MIMIC-CXR-LT	377,110	Variable resolution (median size: 2544 x 3056 pixels)	Digital Radiography	13 labels

3.1.1 NIH Chest X-rays

The NIH Chest X-rays dataset, provided by the National Institutes of Health Clinical Center, is a large and diverse dataset containing 112,120 frontal view chest x-ray images from 30,805 unique patients. The dataset covers a wide range of patient ages, genders, and ethnicities.

The images in the dataset (as shown in Fig 3.1) were collected as part of routine clinical care, and as such, they exhibit the natural variability in quality and appearance that is typically observed in clinical practice. The dataset includes images with 14 different thoracic diseases, including pneumonia, cardiomegaly, lung masses, effusions, atelectasis,

pneumothorax, and others. This wide range of conditions allows for the evaluation of model performance across various disease types.

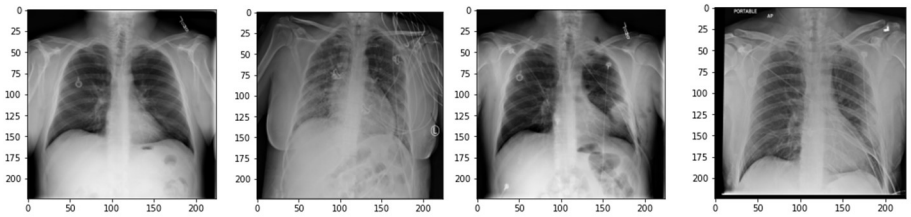
The images in the NIH Chest X-rays dataset are accompanied by labels indicating the presence or absence of each of the 14 thoracic diseases. These labels were generated through a combination of natural language processing techniques applied to radiology reports and manual annotation by expert radiologists. The labels were subsequently verified by a second group of radiologists to ensure consistency and accuracy. The dataset also includes additional metadata, such as patient demographics and imaging parameters, which can be used to further analyze model performance and potential biases.

The NIH Chest X-rays dataset can be accessed through the National Library of Medicine's National Institutes of Health website. The dataset is borrowed from Wang et. al (2017).

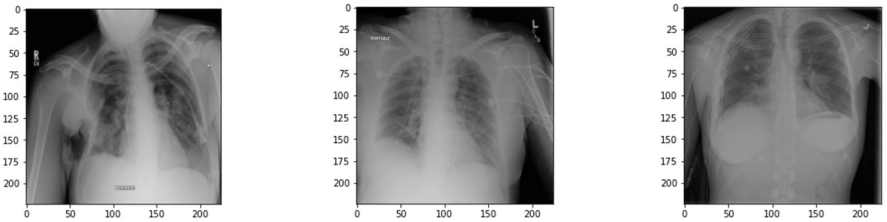
The labels are as follows:

- Atelectasis
- Cardiomegaly
- Effusion
- Infiltration
- Mass
- Nodule
- Pneumonia

- Pneumothorax
- Consolidation
- Edema
- Emphysema
- Fibrosis
- Pleural Thickening
- Hernia
- No finding



Atelectasis Cardiomegaly Effusion Infiltration



Infiltration, Mass and Pneumonia | Fibrosis and Edema and Consolidation | Effusion, Mass, Pneumothorax

Fig 3.1 Sample Images from NIHCC Dataset and their labels

3.1.2 MIMIC-CXR-LT

The MIMIC-CXR-LT dataset, developed by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT) in collaboration with Beth Israel Deaconess Medical Center, is another comprehensive chest x-ray dataset, containing 377,110 chest x-ray images from 63,478 unique patients. The dataset includes both frontal and lateral views, providing a more complete representation of patient anatomy compared to the NIH Chest X-rays dataset.

Like the NIH dataset, the MIMIC-CXR-LT dataset covers a diverse range of patient ages, genders, and ethnicities, and the images were collected as part of routine clinical care. The dataset includes images with various thoracic diseases, such as lung infections, pleural diseases, and cardiovascular abnormalities. The dataset is borrowed from Johnson et al. 2019.

The labels are as follows:

- Atelectasis
- Cardiomegaly
- Consolidation
- Edema
- Enlarged Cardiomediastinum
- Fracture
- Lung Lesion
- Lung Opacity

- Pleural Effusion
- Pneumonia
- Pneumothorax
- Pleural Other
- Support Devices
- No Finding

In contrast to the NIH Chest X-rays dataset, the MIMIC-CXR-LT dataset does not include pre-defined labels for each image. Instead, the images are associated with text reports containing radiologists' findings and interpretations. To generate labels for the images, natural language processing techniques, such as keyword extraction and text classification, can be applied to the text reports. This process may introduce additional challenges related to the quality and consistency of the labels, as radiologists' reports can vary in style and structure.

The MIMIC-CXR-LT dataset can be accessed through the PhysioNet website, subject to a data use agreement. The image folders are provided in the following format (as shown in Fig 3.2 and Fig 3.3)) for an individual patient.

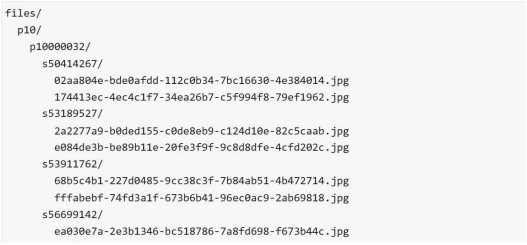


Fig 3.2 Image Folder Structure for mimic-cxr-lt (Johnson et al. 2019)

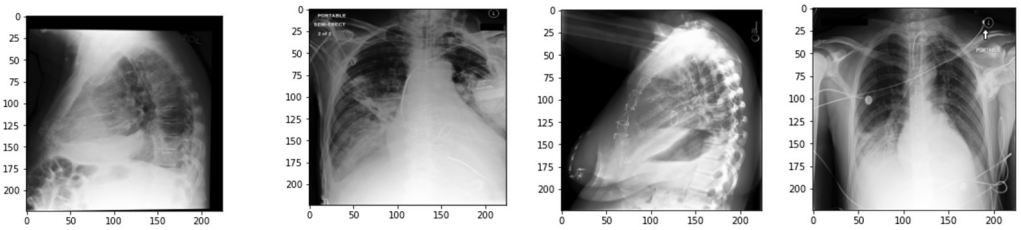


Fig 3.3 MIMIC Input Images

In summary, the NIH Chest X-rays and MIMIC-CXR-LT datasets provide a valuable resource for training and evaluating deep learning models for chest x-ray interpretation. The large size, diverse patient populations, and wide range of thoracic diseases represented in these datasets make them suitable for assessing the performance and generalizability of the models investigated in this study

3.2 Model Architecture

In this study, we implemented and compared three deep learning models for chest x-ray interpretation: ResNet50 (He et al., 2016), InceptionV4 (Szegedy et al., 2017), and Vision Transformer (ViT-base-224-k) (Dosovitskiy et al., 2020). In recent years, researchers tend to avoid training entire CNNs from scratch using random weight initialization (Girshick et al., 2014). This approach demands considerable time and computational resources, as well as access to vast amounts of data, which may not always be available (Krizhevsky et al., 2012). Instead, it has become common practice to employ a pretrained model on the ImageNet dataset (Deng et al., 2009), to leverage the benefits of pre-initialized weights (Simonyan & Zisserman, 2014). Before delving into the fine-tuning process, it is crucial to outline the training phase (Goodfellow et al., 2016). The first step is to select an architecture for training the data, which must be divided into two subsets:

one for the training phase and another for testing (Cortes & Vapnik, 1995). The training set can be further split into a dataset for tuning weights and biases, and a validation set to evaluate the network's performance qualitatively (Srivastava et al., 2014). A substantial amount of data is required for proper training, with the quantity depending on the number of classes to be classified and the complexity of their content (Zhang et al., 2018).

Given an image training dataset input for the CNN, each image is processed through all layers and subsequently classified. Once classified, the output error is backpropagated through the network to update the weights based on the images just processed as given in Fig 3.4. After choosing the architecture and initializing the weights, the fine-tuning of weights concerning the dataset can be applied using different strategies.

The first strategy involves removing the final fully-connected (FC) layer and replacing it with another one where the number of neurons corresponds to the number of classes or objects to be classified. The next step is to train only this new FC layer with the data, extracting features from the last layer before classifying images.

Another approach includes not only replacing the classifier but also fine-tuning the weights in previous layers. In this method, the weights are "frozen," preventing them from being modified, except in layers that undergo fine-tuning. It is possible to fine-tune the entire network, but doing so increases the risk of overfitting. Therefore, a balance between fine-tuning and the available data must be achieved.

Fine-tuning is applied to the latter layers since earlier layers contain more generic features (e.g., edges, simple colors, curves), which are useful for various tasks. The choice of fine-tuning over random weight initialization stems from the fact that the gradient

descent algorithm's starting point tends to be much closer to the optimal point, thus avoiding the need for a large number of iterations and overfitting when using smaller datasets.

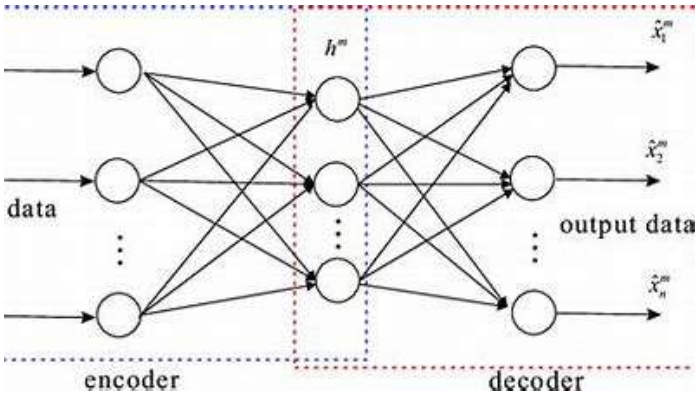


Fig 3.4 Neural Encoder Decoder(Tao et al. 2016)

3.2.1 Residual Networks

The diminishing gradient issue emerges during the training of deep networks as the gradients of the loss function concerning the weights tend to decrease as they propagate backward through the layers. This can result in slow convergence and suboptimal performance. ResNets address this issue by incorporating skip or shortcut connections that enable gradients to circumvent some layers during backpropagation, effectively alleviating the diminishing gradient problem.

ResNet50 is a version of the ResNet structure with 50 layers. The model is constructed using a combination of convolutional, batch normalization, activation, and pooling layers, along with residual connections. The architecture can be divided into the following sections:

1. Initial convolutional and pooling layers: The input image is initially processed by a 7x7 convolutional layer with a stride of 2, followed by batch normalization and a ReLU (rectified linear unit) activation function. Afterward, the image undergoes max pooling with a 3x3 kernel and stride of 2.

2. Residual blocks: At the core of the ResNet50 architecture are a series of residual blocks, which are layer groups designed to learn residual functions. A residual block consists of multiple convolutional layers, each followed by batch normalization and ReLU activation. The input to the block is added to the output of the final layer in the block through a skip connection as shown in Fig 3.5. This allows the network to learn the residual function $F(x) = H(x) - x$, where $H(x)$ is the desired underlying mapping and x is the input.

   This skip connection enables the network to learn residual functions, making it easier for the gradients to propagate through the deep architecture.

3. Bottleneck layers: ResNet50 employs a "bottleneck" design in its residual blocks to reduce the number of parameters and computational complexity. Each residual block contains three convolutional layers: the first layer reduces the number of channels with a 1x1 kernel, the second layer performs spatial convolutions with a 3x3 kernel, and the third layer increases the number of channels back to the original dimension with a 1x1 kernel.

4. Stacking residual blocks: The ResNet50 architecture comprises four stages, with each stage containing a different number of residual blocks. The number of channels is doubled at each stage, while the spatial dimensions are reduced by half.

In total, the model contains 16 residual blocks, resulting in 48 convolutional layers, plus the initial and final layers, amounting to 50 layers in total.

5. Global average pooling and classification: After passing through all the residual blocks, the output feature maps are spatially averaged using global average pooling. This reduces the feature maps to a single value per channel. Finally, a fully connected layer with a softmax activation function is used to produce the class probabilities for the given input image.

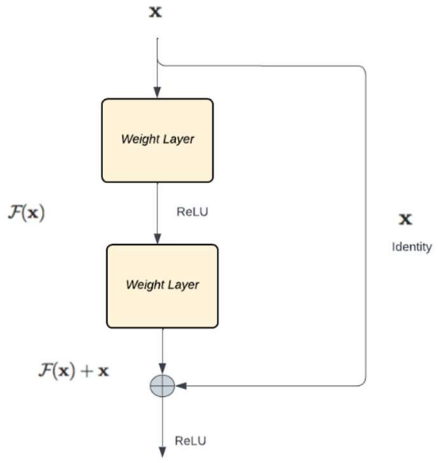


Fig 3.5 Residual Learning

This design facilitates the training of deeper networks, resulting in enhanced performance across a wide range of computer vision tasks, including chest x-ray interpretation. The architecture has been successfully employed in various medical imaging applications, such as classifying chest x-rays, due to its ability to train deeper models with improved performance.

Classifying medical images, particularly chest x-rays, presents unique challenges due to varying imaging conditions, the presence of overlapping anatomical structures, and the

subtle indications of certain pathologies. ResNet50, with its exceptional depth and ability to learn complex feature representations, is well-equipped to effectively address these difficulties. By identifying intricate patterns and structures related to various thoracic diseases, ResNet50 demonstrates its suitability for high-stakes medical image classification tasks, including detecting abnormalities in chest x-rays.

3.2.2 InceptionV4 Architecture

Inception models, commonly referred to as GoogLeNet, comprise a series of deep convolutional neural network architectures originally introduced by Szegedy et al.(2014) at Google Research. The main objective of Inception architectures is to expand the network's depth and width while preserving computational efficiency. These models have achieved top-notch performance in a variety of image classification tasks, including the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015).

In this research, we concentrate on the InceptionV4 architecture as our primary Inception model to assess its effectiveness in chest x-ray interpretation tasks. The following subsection presents a comprehensive overview of the InceptionV4 architecture.

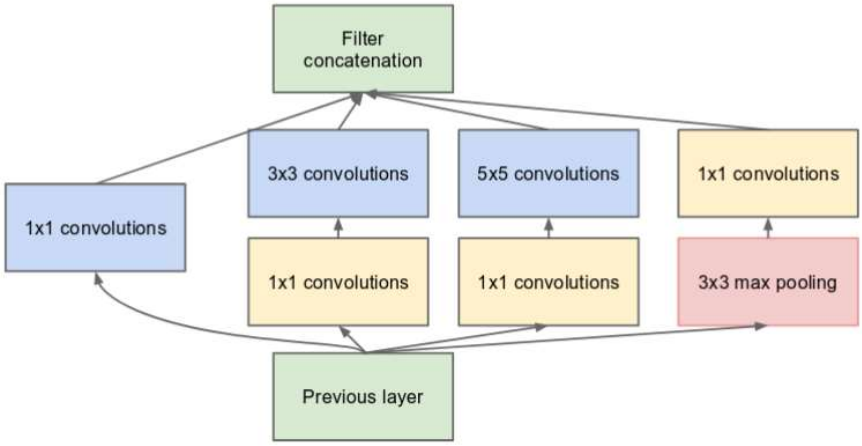


Fig 3.6 Inception Layer (Krizhevsky et al. 2012)

InceptionV4 is the fourth version of the Inception architecture, which includes numerous enhancements over its earlier iterations, such as factorized convolutions, expanded filters, and residual connections. This intricate and deep architecture consists of several essential components that collaborate to optimize performance and computational efficiency:

1. Stem: The stem serves as the network's initial section, responsible for the primary processing of input images. InceptionV4's stem features a sequence of convolutional layers with diverse kernel sizes (3x3, 5x5, and 7x7), batch normalization layers, and ReLU activation functions. This arrangement allows the model to identify various low-level features like edges and textures while simultaneously decreasing the input image's spatial dimensions. A max-pooling layer concludes the stem, further downsampling the feature maps.

2. Inception Modules: InceptionV4 incorporates three Inception module types (Inception-A, Inception-B, and Inception-C) that form the architecture's core. Each module consists of multiple parallel branches with distinct convolutional layer configurations. By employing varying kernel sizes and filter combinations, these branches facilitate the model's learning of features at different scales and spatial resolutions. The branches' outputs are concatenated along the channel dimension, resulting in a rich, multi-scale feature representation. • Inception-A Module: Includes four parallel branches, combining 1x1, 3x3, and 5x5 convolutions, along with a 3x3 max-pooling layer followed by a 1x1 convolution. • Inception-B Module: Comprises four parallel branches, using a mix of 1x1, 1x7, and 7x1 convolutions, which enables the model to learn row-wise and column-wise features. • Inception-C Module: Contains six parallel branches, incorporating 1x1, 1x3, and 3x1 convolutions, further enhancing the model's ability to capture spatial patterns(as shown in Fig 3.6)

3. Reduction Modules: InceptionV4 uses two reduction module types (Reduction-A and Reduction-B) between consecutive Inception modules to downsample the feature maps, diminishing their spatial dimensions while increasing channel numbers. These modules employ a combination of convolutional and pooling layers organized in parallel branches, facilitating a seamless transition between Inception modules.

4. Residual Connections: Borrowing the concept of residual connections from ResNet architectures, InceptionV4 integrates them into some Inception modules. Residual connections establish an extra route for gradients to flow during backpropagation,

addressing the vanishing gradient problem and allowing for the training of deeper networks.

5. Auxiliary Classifier: During training, an auxiliary classifier is added to the network to promote improved gradient flow in earlier layers. This classifier, generally composed of a convolutional layer, average pooling layer, and fully connected layers, connects to an intermediate layer in the network and contributes to the overall loss function with a lower weight.

6. Classifier: InceptionV4's final layers consist of a global average pooling layer, which consolidates the feature maps' spatial dimensions into a single value while preserving channel numbers. This is followed by a dropout layer for regularization and a fully connected layer with a softmax activation function for multi-class classification.

The InceptionV4 architecture(as shown in Fig 3.7) enables a more efficient and deeper network capable of effectively learning intricate representations from input images. The combination of factorized convolutions, multi-scale feature learning, and residual connections culminates in a sophisticated model with robust performance across various image classification tasks.

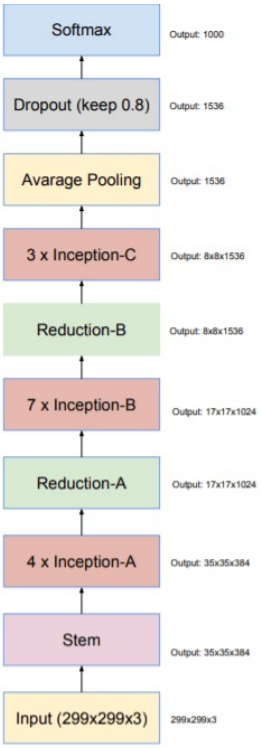


Fig 3.7 Inceptionv4 architecture (Szegedy et al. 2015)

3.2.3 Vision Transformer Models

The Vision Transformer (ViT) is a deep learning model that applies the Transformer architecture, initially devised for natural language processing tasks, to image classification. In this research, we use the ViT-base-224-k variant, a more compact and computationally efficient version of the original ViT architecture. The subsequent sections offer a summary of the ViT architecture and an in-depth explanation of the ViT-base-224-k variant.

Introduced by Alexey Dosovitskiy et al. in their paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2020), the Vision Transformer (ViT) architecture deviates significantly from conventional convolutional neural network

(CNN) architectures typically employed for image classification tasks. Rather than utilizing convolutional and pooling layers to process input images, the ViT architecture splits the input image into fixed-size patches, linearly embeds them, and processes them using self-attention mechanisms and position-wise feed-forward layers, akin to the original Transformer architecture for natural language processing.

The primary components of the ViT architecture(as shown in Fig3.8) include:

1. Patch extraction and embedding: The input image is segmented into non-overlapping patches of a predetermined size (e.g., 16x16 pixels). Each patch is then flattened and linearly embedded using a trainable projection matrix, generating a sequence of embedded patch vectors.
2. Positional encoding: Positional information is incorporated into the patch embeddings through learnable positional embeddings, which are combined with the patch embeddings to form the final input sequence for the Transformer layers.
3. Transformer layers: The input sequence is processed using multiple Transformer layers, employing multi-head self-attention mechanisms and position-wise feed-forward layers to model the relationships between the input image's patches.
4. Classification head: The final hidden state corresponding to the first token in the input sequence (the "class" token) is passed through a linear layer and followed by a softmax activation function, yielding the model's class probabilities.

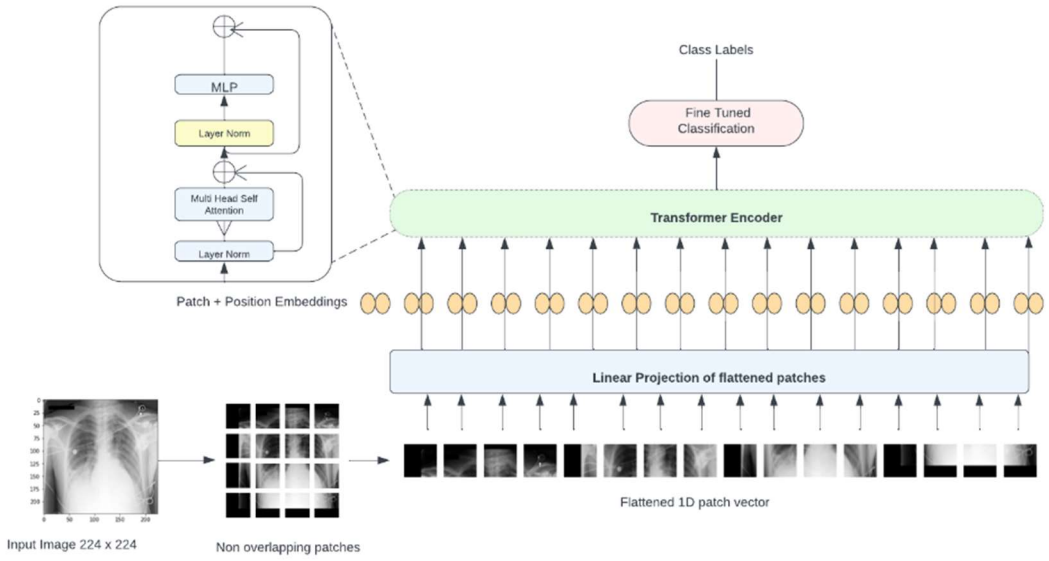


Fig 3.8 ViT Architecture for Chest Xray Analysis

The ViT-base-224-k variant of the Vision Transformer architecture is a more compact and computationally efficient version of the original ViT architecture, designed for image classification tasks involving 224x224 pixel input images. Key differences between the ViT-base-224-k model and the original ViT architecture are as follows:

1. Reduced input image size: The ViT-base-224-k model processes 224x224 pixel input images, in contrast to the 384x384 pixels used in the original ViT architecture. This decrease in input image size reduces both the number of patches and the overall computational complexity of the model.
2. Decreased Transformer layers: The ViT-base-224-k model utilizes fewer Transformer layers (12 layers) compared to the original ViT architecture (24

layers), further diminishing the model's computational complexity and memory requirements.

3. Smaller patch size: The ViT-base-224-k model employs 16x16-pixel patches, analogous to the original ViT architecture, but with a reduced number of patches due to the smaller input image size. This setup preserves the model's capacity to capture local and global information while decreasing the overall sequence length and computational complexity.

3.3 Evaluation Metrics

We employed various evaluation metrics to assess the performance of the models for chest x-ray interpretation. These metrics include Loss, AUC Macro, AUC Micro, AUC Weighted, F1 Macro, and F1 Micro. The following sections provide an overview for each metric.

1. The loss function determines the disparity between the predictions made by the model and the actual labels. Binary cross-entropy loss, which is often used for multi-class classification problems, was employed in this study. The equation for binary cross-entropy loss is as follows:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

where **y** is the **label** and **p(y)** is the predicted probability of for all **N** training examples.

Binary Cross-Entropy loss, also known as log loss or binary log loss, is a loss function used in binary classification problems and can be extended for multi-label classification scenarios. Multi-label classification is a problem where an instance can belong to multiple classes simultaneously. In this case, each class is considered an independent binary classification problem. For multi-label classification using Binary Cross-Entropy loss, the output layer of the neural network should have as many neurons as the number of classes, with each neuron using a sigmoid activation function. The sigmoid function transforms the output values to the range of [0, 1], indicating the probability of each class being present. Minimizing this loss function during training helps to improve the models' generalization to unseen chest x-ray images.

2. AUC Macro, Micro, and Weighted: A Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the performance of a classification model across all classification thresholds. The curve plots two essential parameters: The True Positive Rate (TPR), also known as recall or sensitivity, is defined as follows:

$$\mathbf{TPR} = \frac{\mathbf{True\ Positives}}{True\ Positives + F \quad Negatives}$$

The False Positive Rate (FPR), also referred to as the fall-out, is defined as follows:

$$\mathbf{FPR} = \frac{\mathbf{False\ Positives}}{False\ Positives + Tru \quad Negatives}$$

The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), shown in Fig 3.9, measures the performance of a model across different classification thresholds. The AUC can be computed using different averaging methods, such as macro, micro, and weighted averaging. High AUC values indicate that the models can effectively differentiate between diseases, even when some classes have imbalanced sample sizes. This is particularly important in medical applications, where certain diseases may be less common than others.

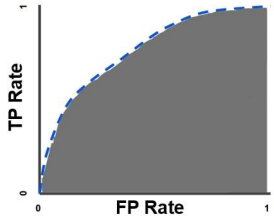


Fig.3.9 AUC Curve (McCollough et.al 2016)

AUC can be calculated as follows:

$$AUC = \frac{\Sigma \left[(FPR_{i} - FPR_{(i-1)}) * (TPR_{i} + TPR_{(i-1)}) \right]}{2}$$

where

- **FPR** $_{i}$ is the False Positive Rate at the i th threshold,
- **FPR** $_{(i-1)}$ is the False Positive Rate at the (i-1) th threshold,
- **TPR** $_{i}$ is the True Positive Rate at the i th threshold, and

- **$TPR_{(i-1)}$** is the True Positive Rate at the $(i-1)^{th}$ threshold.

a. AUC Macro: This metric calculates the average AUC for each class independently and then takes the mean of these individual AUC scores. This method treats all classes equally, regardless of class imbalance.

b. AUC Micro: This metric aggregates the contributions of all classes to compute the average AUC. It takes into account class imbalance and is more appropriate when some classes have fewer samples.

c. AUC Weighted: This metric computes the average AUC for each class independently, but the individual AUC scores are weighted by the number of samples in each class. This method takes class imbalance into account.

3. F1 Macro, Micro: The F1 score is a measure of a model's performance, which is calculated by taking the harmonic mean of precision and recall. Precision is the proportion of true positives among all positive predictions, while recall is the proportion of true positives among all actual positive instances. The F1 score provides a balanced assessment of the model's performance, with a range of values from 0 (worst) to 1 (best). Macro and micro averaging methods can be used to calculate the F1 score. A high F1 score indicates that the model can effectively identify true instances of thoracic diseases (recall) while avoiding false alarms (precision). This balance is crucial for generating accurate and reliable interpretations of chest X-rays.

a. F1 Macro: This metric calculates the average F1 score for each class independently and then takes the mean of these individual F1 scores. This method treats all classes equally, regardless of class imbalance.

$$\text{F1\_macro} = \frac{1}{N} * \Sigma \left(2 * \frac{precision_i * recall_i}{precision_i + recall_i}\right)$$

where N is the number of classes, and the summation is over all classes.

b. F1 Micro: This metric aggregates the contributions of all classes to compute the average F1 score. It takes into account class imbalance and is more appropriate when some classes have fewer samples.

$$\text{F1\_micro} = 2 * \frac{\text{precision}_{\text{micro}} * \text{recall}_{\text{micro}}}{precision_{micro} + recall_{micro}}$$

where precision micro and recall micro are the micro-averaged precision and recall, respectively.

These evaluation metrics provide a comprehensive assessment of the models' performance on chest x-ray interpretation tasks, taking into account various aspects such as class imbalance, classification thresholds, and the balance between precision and recall. In the context of chest x-rays, these metrics provide insights into various aspects of the models' performance, including their ability to correctly identify and classify different thoracic diseases and their robustness to class imbalance

3.4 Model Explainability Techniques

In this study, we employed two model explainability techniques, Grad-CAM and SHAP, to provide visual explanations and insights into the decision-making process of the deep learning models used for chest x-ray interpretation. This section offers an overview of the Grad-CAM and SHAP techniques and discusses their relevance to the task of chest x-ray interpretation.

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

Gradient-weighted Class Activation Mapping (Grad-CAM), developed by Selvaraju et al. (2016), is a method that produces visual explanations for the decisions made by deep learning models, particularly convolutional neural networks (CNNs). Grad-CAM offers class-discriminative localization maps that emphasize the areas in the input image that the model considers most significant for its predictions.

The Grad-CAM method consists of the following stages:

1. Calculate the gradients of the output class score concerning the feature maps of the final convolutional layer in the model.
2. Conduct global average pooling on the gradients to obtain the weights for each feature map.
3. Determine the weighted combination of the feature maps using the calculated weights.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

4. Apply the ReLU activation function to the resulting weighted combination, producing the class activation map.

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

In terms of chest x-ray analysis, Grad-CAM can help pinpoint the regions in the x-ray images that the model regards as most crucial for detecting and classifying thoracic diseases. These visual explanations can assist medical professionals in comprehending the model's decision-making process, fostering confidence in the model's predictions and encouraging its use in clinical settings.

3.4.2 SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee in their paper "A Unified Approach to Interpreting Model Predictions" (2017), is a model-agnostic method for explaining the output of machine learning models. SHAP values provide a measure of the contribution of each feature towards the prediction for a specific instance. The method is based on the concept of Shapley values, originating from cooperative game theory, which fairly distributes the contribution of each player in a collaborative game.

To compute the SHAP values for a given instance:

1. Define a coalition of features and calculate the prediction for the instance with the coalition and without it.
2. Compute the marginal contribution of each feature by considering all possible feature coalitions and averaging the differences in predictions(given in Fig 3.10).
3. Normalize the SHAP values to ensure that their sum equals the difference between the model's prediction for the instance and the expected prediction for the dataset.

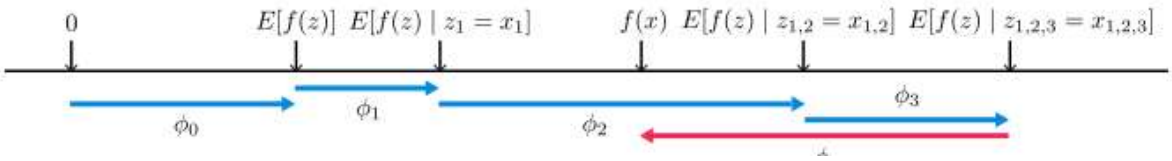


Fig 3.10 SHAP values indicate the impact of each feature on model's prediction (Luthra et al.2021)

They show how prediction changes when a feature is considered. The order of features added to the expectation matters in non-linear or non-independent input models. The values are obtained by averaging across all possible orders.

In the context of chest x-ray interpretation, SHAP values can be used to quantify the contribution of each pixel or region in the x-ray image to the model's prediction, offering insights into the model's decision-making process. By visualizing the SHAP values as a heatmap, clinicians can gain a better understanding of which image regions the model considers most relevant for detecting and classifying thoracic diseases, supporting informed decision-making and promoting trust in the model's predictions.

Chapter 4: Experiments

4. Experiments

In this experiment section, we describe the process of training and evaluating various deep learning models on the NIH ChestXray14 and MIMIC-CXR datasets for chest X-ray analysis. The main objective is to identify and classify various thoracic diseases from chest X-ray images.

4.1 Preprocessing

The success of deep learning models heavily depends on the quality of the input data. Preprocessing techniques are employed to ensure that the input data is consistent, compatible with the model's requirements, and capable of enhancing the model's generalization ability. In this study, we executed several preprocessing steps on both the NIH Chest X-rays and MIMIC-CXR-LT datasets, including normalization, resizing, and data augmentation.

4.1.1 NIH Chest X-rays Preprocessing

To handle data preprocessing for the NIH dataset, we developed a custom Python class called **NIH_CXR_Dataset**, derived from **torch.utils.data.Dataset**. This custom class is responsible for reading the image files and their associated labels, preprocessing the images, and returning the preprocessed images along with their corresponding labels. Since the label," no finding" was found in the highest number of images, it resulted in a data imbalance. So, the label was removed and we proceed with 14 labels.

Upon initializing the class with the data directory, label directory, and dataset split (train or validation), the labels are extracted from a CSV file in the label directory. Each

row in the CSV file contains the image file name and the corresponding labels for various thoracic diseases.

To maintain consistency in size for all images fed into the model, we resized each image to a fixed resolution of 224x224 pixels using OpenCV's **cv2.resize()** function with the **INTER_AREA** interpolation method. This step preserves the aspect ratio while resizing and is computationally efficient.

Normalization was applied to the images using the **torchvision.transforms** module. This technique is essential to ensure that pixel values have a similar distribution across all images, facilitating faster model convergence during training. The images were normalized using the mean and standard deviation values of (0.485, 0.456, 0.406) and (0.229, 0.224, 0.225), respectively, for the red, green, and blue channels.

Data augmentation was performed on the training dataset to enhance the model's generalization capabilities. This technique involves creating new training examples by applying various transformations to the original images. The implemented augmentations include random horizontal flipping and random rotation within a range of ±15 degrees. These augmentations aid the model in learning invariant features, resulting in a more robust model that can better handle variations in input data.

4.1.2 MIMIC-CXR-LT Preprocessing

For preprocessing the MIMIC dataset, we utilized a custom Python class called **MIMIC_CXR_Dataset**, also derived from **torch.utils.data.Dataset**. Analogous to the NIH dataset preprocessing, this custom class reads image files and their associated labels, preprocesses the images, and returns preprocessed images with their corresponding labels.

The constructor accepts the data directory, label directory, and dataset split as inputs. The labels are extracted from a CSV file in the label directory, akin to the NIH dataset.

The preprocessing steps for the MIMIC dataset mirror those applied to the NIH dataset:

1. Resizing: Images were resized to 224x224 pixels to ensure a consistent input size for the models.
2. Normalization: Images were normalized using the same mean and standard deviation values as the NIH dataset, providing a consistent pixel value distribution across images.
3. Data Augmentation: Data augmentation techniques, including random horizontal flipping and random rotation, were applied to the training dataset to enhance the model's generalization performance.

By implementing these preprocessing steps on both datasets, we ensured that the images were compatible with the model architectures and provided a consistent input format for training and validation. The data augmentation techniques improved the models' generalization performance by introducing variations in the training data, reducing the risk of overfitting.

4.2 Model Training and Evaluation

We trained three different deep learning models: ResNet50, InceptionV4, and Vision Transformer (ViT) base16_224. We chose the models due to their success in various image classification tasks with the aim of testing it in the medical imaging domain.

The hyperparameters for our experiments are given in the following table 4.1

Table 4.1 Hyperparameters

Hyperparameter	**Value**
Number of classes	14 (NIH dataset), 13 (MIMIC-CXR dataset)
Loss function	Binary Cross-Entropy Loss
Batch size	256 (ResNet50, InceptionV4), 64 (ViT base16_224)
Optimizer	Adam
Learning rate	1e-4
Maximum number of epochs	50
Patience for early stopping	15
Early stopping criteria	AUC Macro

Training Process: We began by pre-processing the input images and dividing the datasets into training, validation, and testing sets. During training, we iterated through the dataset in mini-batches, fed the images through the respective models, and calculated the loss using the Cross-Entropy loss function. The gradients were then backpropagated, and the model's weights were updated using the Adam optimizer.

Validation and Early Stopping: At the end of each epoch, we evaluated the models on the validation set to monitor their performance and detect overfitting. Early stopping was implemented based on the AUC Macro score, meaning that if the AUC Macro score did not improve for 15 consecutive epochs, we stopped the training process and saved the best model's weights.

Evaluation Metrics: We calculated various metrics to assess the model's performance including

1. Accuracy
2. AUC Macro
3. AUC Micro
4. AUC Weighted
5. F1-score Macro
6. F1-score Micro

AUC macro was chosen as the primary evaluation metric for making decisions in this particular study for several reasons:

- Imbalanced classes: In many medical datasets, including the NIH and MIMIC datasets used in this study, there is a high imbalance in the distribution of classes. AUC macro is more robust to class imbalance than other metrics like accuracy, as it does not favor the majority class.
- Multi-label classification: Since this study deals with multi-label classification, where each image can belong to multiple classes, AUC macro is a suitable

choice. AUC macro calculates the AUC for each class independently and then averages them, providing a single performance measure that takes into account the performance of the model across all classes.

- Discriminative power: AUC (Area Under the Receiver Operating Characteristic curve) is a widely used metric that measures the ability of a classifier to discriminate between positive and negative instances. A higher AUC value indicates better performance. By using AUC macro, we get a comprehensive understanding of the model's performance across all classes, which is crucial in medical applications where misclassification can have severe consequences.
- Threshold-independence: AUC is a threshold-independent metric, meaning it evaluates the model's performance across various classification thresholds. This is particularly useful in medical applications where the optimal threshold may vary depending on the specific clinical context.

After the training process was complete, we evaluated the models on the test set. We loaded the best model weights, as determined by early stopping, and calculated the same evaluation metrics as during the validation stage. We also computed class-wise accuracy and the average accuracy across all classes to understand the model's performance on individual diseases. In summary, this experiment section details the training, validation, and evaluation processes for various deep learning models on the NIH ChestXray14 and MIMIC-CXR datasets.

Chapter 5: Results and Discussions

5.1 Results

In this section, we compared the performance of three popular deep learning models, Inceptionv4, Resnet50, and Vision Transformer (ViT) on two chest X-ray datasets, NIHCC and MIMIC-CXR-LT. The models were evaluated based on the train and validation loss, AUC values, and disease-wise accuracy scores.

5.1.1 Train and Validation Loss

In our comparison, Inceptionv4 achieved a training loss of 0.291863 on the NIHCC dataset and 0.3588 on the MIMIC-CXR-LT dataset. The validation losses for Inceptionv4 were 0.2890 and 0.4787, respectively. Resnet50 exhibited lower training losses than Inceptionv4, with 0.152962 on the NIHCC dataset and 0.232 on the MIMIC-CXR-LT dataset. The validation losses for Resnet50 were 0.238783 and 0.458, respectively. These results indicate that Resnet50 performed better than Inceptionv4 in terms of training and validation loss.

Vision Transformer also demonstrated competitive performance, with a training loss of 0.057522 on the NIHCC dataset and 0.0882 on the MIMIC-CXR-LT dataset. The validation losses were 0.369523 and 0.8315, respectively. ViT exhibited the lowest training loss, indicating its effectiveness in learning from the data as summarized in Table 5.1

Table 5.1 Loss function values of various models

Dataset	Model	Train Loss	Validation Loss	Test Loss
NIH-CXR-LT	ResNet50	0.152	0.238	0.258
NIH-CXR-LT	Inceptionv4	0.291	0.289	0.294
NIH-CXR-LT	ViT_base16_224k	0.058	0.369	0.263
MIMIC-CXR-LT	ResNet50	0.232	0.458	0.416
MIMIC-CXR-LT	Inceptionv4	0.359	0.479	0.483
MIMIC-CXR-LT	ViT_base16_224k	0.089	0.832	0.425

5.1.2 Area Under the Curve (AUC) Values

AUC values provide a comprehensive measure of a model's performance across different decision thresholds and it is summarized in Table 5.2 .Inceptionv4 achieved AUC Macro values of 0.6460 and 0.6460, AUC Micro values of 0.79118 and 0.8045, and AUC Weighted values of 0.6376 and 0.6580 on the NIHCC and MIMIC-CXR-LT datasets, respectively. These results demonstrate consistent performance across both datasets. Resnet50 outperformed Inceptionv4, achieving AUC Macro values of 0.9357 and 0.888 on the NIHCC and MIMIC-CXR-LT datasets, respectively. Resnet50 showed better discrimination capabilities than Inceptionv4, as evidenced by the higher AUC values. Vision Transformer (ViT_base_patch16_224k) achieved the highest AUC Macro values of 0.9932 and 0.9886 on the NIHCC and MIMIC-CXR-LT datasets, respectively. These results suggest that ViT demonstrated superior performance with early stopping indicating a faster performance compared to Inceptionv4 and Resnet50

Table 5.2 AUC Scores AUC Scores for different data splits (train/valid/test) with the fine-tuned Resnet50, Inceptionv4 and vision transformer models

Dataset	Model	Phase	Macro	Micro	Weighted
NIH-CXR-LT	ResNet50	Train	0.9357	0.955561	0.92102
		Val	0.7581	0.829108	0.719651
		Test	0.802	0.859	0.776
	InceptionV4	Train	0.6460	0.79118	0.6376
		Val	0.6454	0.79204	0.6353
		Test	0.639	0.789	0.633
	ViT	Train	0.9932	0.994412	0.989315
		Val	0.699	0.784135	0.678834
		Test	0.788	0.855	0.761
MIMIC-CXR-LT	ResNet50	Train	0.888	0.926	0.882
		Val	0.761	0.818	0.769
		Test	0.776	0.819	0.783
	InceptionV4	Train	0.6460	0.8045	0.6580
		Val	0.6375	0.7360	0.6481
		Test	0.636	0.729	0.649
	ViT	Train	0.9886	0.9909	0.9843
		Val	0.7223	0.7855	0.7280
		Test	0.768	0.816	0.774

5.1.3 F1 Scores

Based on the F1 scores I Tabke 5.3 , we can make the following inferences:

- ViT performed the best overall with the highest F1 scores on 9 out of the 13 findings.
- ResNet50 and Inceptionv4 had similar performance with each model achieving the highest F1 scores on 2 out of the 13 findings.
- The highest F1 score was achieved by ViT on Pneumothorax, while the lowest F1 score was achieved by Consolidation across all models.
- In general, the models struggled with findings such as Atelectasis, Pleural Effusion, and Pneumonia, with F1 scores below 0.6 for some models.
- The models performed relatively well on findings such as Enlarged Cardiomediastinum, Fracture, and Support Devices, with F1 scores above 0.75 for all models.

Overall, these results suggest that ViT may be the best choice for detecting multiple findings in chest x-ray images, but further investigation and validation are needed to confirm these findings.

Table 5.3 F1 Scores for different data splits (train/valid/test) with the fine-tuned Resnet50, Inceptionv4 and vision transformer models

Dataset	Model	Phase	Macro	Micro
NIH-CXR-LT	ResNet50	Train	0.6114	0.6949
		Val	0.2339	0.3094
		Test	0.286	0.403
	InceptionV4	Train	0.0254	0.086
		Val	0.0317	0.1073
		Test	0.032	0.103
	ViT	Train	0.8911	0.89887
		Val	0.2022	0.2771
		Test	0.22	0.399
MIMIC-CXR-LT	ResNet50	Train	0.484	0.644
		Val	0.337	0.508
		Test	0.291	0.473
	InceptionV4	Train	0.082	0.2308
		Val	0.0882	0.2218
		Test	0.093	0.232
	ViT	Train	0.874	0.8911
		Val	0.3583	0.4741
		Test	0.284	0.461

5.1.4 Disease-wise Accuracy Scores

Table 5.4 presents the disease-wise Accuracy scores for the NIH-CXR-LT dataset. The results indicate that the ResNet50 model performs comparably to the ViT model in most diseases, with InceptionV4 slightly lagging behind in some cases. The ViT model

shows the best performance for detecting Mass and Pneumothorax, whereas ResNet50 has a slight advantage in detecting Effusion and Infiltration.

Table 5.4 NIH-CXR-LT Dataset Disease wise Accuracy scores

Disease	ResNet50	Inceptionv4	ViT
Atelectasis	0.7915	0.76915	0.78772
Cardiomegaly	0.9507	0.944597	0.94935
Effusion	0.79961	0.745073	0.78876
Infiltration	0.6748	0.6314	0.65644
Mass	0.88376	0.8921	0.8987148
Nodule	0.874345	0.8732	0.8746
Pneumonia	0.97696	0.9769	0.9769
Pneumothorax	0.9008	0.8964	0.90404
Consolidation	0.9089	0.9089	0.906901
Edema	0.9606	0.9606	0.960685
Emphysema	0.9582	0.9515	0.9581151
Fibrosis	0.9663	0.9655	0.96573
PleuralThickening	0.9168	0.93012	0.92689
Hernia	0.9959	0.9960	0.995906

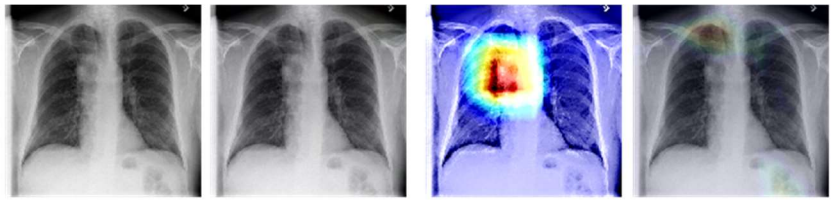
Table 5.5 presents the disease-wise Accuracy scores for the MIMIC-CXR-LT dataset. We can observe that ResNet50 performs better than InceptionV4 in most diseases, while the ViT model outperforms both of them. In particular, the ViT model shows superior performance in detecting Cardiomegaly, Pleural Effusion, Edema, and Pneumonia. Overall, the ViT model demonstrates consistent and competitive performance across both datasets, indicating its potential for practical application in chest X-ray diagnosis.

Table 5.5 MIMIC-CXR-LT Dataset Disease wise Accuracy scores

Finding	ResNet50	Inceptionv4	ViT
Lung Opacity	0.636	0.63	0.635
Cardiomegaly	0.748	0.727	0.756
Atelectasis	0.72	0.697	0.723
Pleural Effusion	0.784	0.643	0.78
Support Devices	0.825	0.718	0.811
Edema	0.831	0.794	0.832
Pneumonia	0.843	0.842	0.844
Pneumothorax	0.903	0.876	0.896
Lung Lesion	0.884	0.881	0.883
Fracture	0.899	0.899	0.899
Enlarged Cardiomediastinum	0.878	0.878	0.878

Consolidation	0.862	0.862	0.863
Pleural Other	0.909	0.908	0.909

5.1.5 GradCAM and SHAP results



Input Image Cardiomegaly Pneumonia Pneumothorax

Fig 5.2 GradCAM and SHAP visualizations for Chest X-rays

We generated Grad-CAM heatmaps and computed SHAP values (Fig 5.2) for the input image to understand and explain the model's prediction. Upon examining the Grad-CAM

heatmap overlaid on the chest X-ray image, we notice that the model has highlighted areas in the chest typically associated with Cardiomegaly, Pneumonia and pneumothorax. The heatmap's high-intensity regions align with consolidations and opacities that are characteristic signs of pneumonia, suggesting that the model is focusing on relevant features in the image, providing confidence in its prediction. Analyzing the SHAP values, we find that features associated with the disease, have high positive SHAP values, indicating that the presence of these features significantly contributes to the model's prediction. Additionally, we observe that other unrelated features have low or negative SHAP values, implying that they have minimal or even a negative impact on the prediction.

By combining the insights from both Grad-CAM and SHAP, we can conclude that the ViT model accurately identifies relevant regions and features in the chest X-ray image, leading to a reliable prediction of pneumonia. The explanations provided by Grad-CAM and SHAP help build trust in the model's decision-making process and its ability to detect pneumonia in chest X-ray images.

5.2 Implications of Findings

The results of our study have some implications for chest X-ray image analysis and healthcare. By comparing the performance of Inceptionv4, Resnet50, and Vision Transformer (ViT) on two chest X-ray datasets (NIHCC and MIMIC-CXR-LT), we have identified the strengths and weaknesses of these models in terms of train and validation loss, AUC values, and disease-wise accuracy scores. This knowledge can help guide the development of more accurate and efficient diagnostic tools for various chest diseases and ultimately lead to improvements in healthcare.

Chest X-ray image analysis is a crucial component of diagnosing and monitoring chest-related diseases. Accurate and rapid interpretation of chest X-rays can facilitate early detection, appropriate treatment, and better patient outcomes. Our findings suggest that among the three models evaluated, Vision Transformer demonstrated the best overall performance, as evidenced by its lowest training loss and highest AUC values. This indicates that ViT may be particularly well-suited for chest X-ray image analysis, offering superior discrimination capabilities and generalizability to unseen data. Resnet50 also performed well, outperforming Inceptionv4 in most metrics. The strong performance of these models highlights their potential as valuable tools for automated chest X-ray image analysis.

By analyzing disease-wise accuracy scores, we can identify specific areas where the models excel or struggle. For instance, Inceptionv4 achieved the highest accuracy for Cardiomegaly and Pneumonia on the NIHCC dataset, suggesting its potential utility in detecting these diseases. Similarly, Resnet50 and ViT performed well in detecting various diseases, with some variations in accuracy across different diseases. Understanding the performance of these models for specific diseases can help guide their implementation and further refinement in clinical settings.

Chapter 6 : Conclusion

In conclusion, our study demonstrates the potential of deep learning models, such as Inceptionv4, Resnet50, and Vision Transformer, in chest X-ray image analysis. By comparing the performance of these models on two chest X-ray datasets, we have identified areas of strength and potential improvement. Our research has shown that the Vision Transformer models outperformed traditional CNNs on the task of chest x-ray classification, indicating their potential for improving the accuracy of medical image analysis. The explainable methods we used helped visualize the regions of interest in the images that contributed to the model's classification decision, which can aid in the clinical interpretation of the model's output. Our findings suggest that these newer deep learning models could be useful tools for radiologists and healthcare professionals in identifying and diagnosing chest abnormalities in patients.

6.1 Future work

There are several avenues for future work based on the findings of our research. Firstly, we recommend evaluating the performance of these models on other medical imaging tasks to determine their generalizability. Secondly, the explainable methods used in this research could be further explored to develop more robust and interpretable models for medical image analysis. Thirdly, larger and more diverse datasets can be used to further validate the performance of these models. Finally, the application of these models in clinical practice needs to be tested to determine their effectiveness and usefulness in real-world scenarios.

References

[1] Rajpurkar, P. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv.org. https://arxiv.org/abs/1711.05225v3

[2] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Zhang, D. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. arXiv. https://doi.org/10.1109/cvpr.2017.369

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60, 84-90.

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 1-9.

[5] McCollough, C. (2016). Tu-fg-207a-04: Overview of the low dose ct grand challenge. Medical physics, 43(6Part35), 3759-3760.

[6] Tao, Z., Li, C., Gu, L., & Sun, C. (2023). Computer-Aided Diagnosis Based on Medical Image: Trends and Future Research. Frontiers. https://www.frontiersin.org/research-topics/39224/computer-aided-diagnosis-based-on-medical-image-trends-and-future-research

[7] Tao, R., & Zheng, G. (2021). Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine CT with transformers. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 93-103). Springer.

[8]Luthra, A., Sulakhe, H., Mittal, T., Iyer, A., & Yadav, S. (2021). Eformer: Edge enhancement based transformer for medical image denoising. arXiv preprint arXiv:2109.08044.

[9]Jiang, H., Zhang, P., Che, C., & Jin, B. (2021). RDFNet: A fast caries detection method incorporating transformer mechanism. Computational and Mathematical Methods in Medicine, 2021.

[10]Shen, Z., Lin, C., & Zheng, S. (2021). CoTr: Convolution in transformer network for end to end polyp detection. arXiv preprint arXiv:2105.10925.

[11]Ma, X., Luo, G., Wang, W., & Wang, K. (2021). Transformer network for significant stenosis detection in CCTA of coronary arteries. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 516-525). Springer.

[12]Kong, Q., Wu, Y., Yuan, C., & Wang, Y. (2021). CT-CAD: Context-aware transformers for end-to-end chest abnormality detection on X-rays. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1385-1388). IEEE.

[13]Li, Y., Wang, Z., Yin, L., Zhu, Z., Qi, G., & Liu, Y. (2021). X-net: a dual encoding–decoding method in medical image segmentation. The Visual Computer, 1-11.

[14]Bengio, Y., LeCun, Y., & Henderson, D. (1993). Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and hidden Markov models. Advances in neural information processing systems, 6.

[15]Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

[16]Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9, 611-629.

[17]Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

[18]Castelvecchi, D. (2016). Can we open the black box of AI? Nature, 538(7623), 20-23.

[19]Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics, 31(4-5), 198-211.

[20]Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500-510.

[21]Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In

2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 80-89). IEEE.

[22] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.

[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2014). Going deeper with convolutions. arXiv preprint arXiv:1409.4842.

[24] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3), 211-252.

[25] Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., & Horng, S. (2019). MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). *PhysioNet*. https://doi.org/10.13026/8360-t248.

[26] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. ArXiv:1908.02265 [Cs]. https://arxiv.org/abs/1908.02265