Using chest X-ray images and deep learning for automated detection of pathologies

Student Name: Qingsong Tan
Supervisor Name: Katsigiannis, Dr Stamos
Submitted as part of the degree of BSc Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—Medical image classification is a critical task in computer-aided diagnosis. Although CNNs have demonstrated strong performance in this domain, they face inherent limitations in capturing global contextual information. Recently, Vision Transformers (ViTs) have emerged as a promising alternative, leveraging self-attention mechanisms to enhance feature extraction. This study focuses on the task of multi-pathology detection and classification in chest X-ray images, specifically conducting a systematic evaluation of three categories of models on the Chest X-ray 14 dataset: traditional CNN architectures, pure Vision Transformer (ViT) models, and hybrid approaches that integrate CNNs with self-attention mechanisms. Furthermore, traditional attention mechanisms, including the Convolutional Block Attention Module (CBAM) and the Attention Gate (AG), are integrated into CNN architectures to enhance feature representation. To further refine classification decisions, an AUC-based threshold optimization method is employed to adjust the final model predictions. Experimental results demonstrate that both ViT models and hybrid architectures outperform traditional CNNs in overall classification performance. Notably, hybrid models based on ConvNeXt achieve the highest mean AUC of 0.77, representing a 0.04 improvement over the baseline DenseNet121, while also reducing training time by approximately 40%. In contrast, for certain disease categories, CNN models enhanced with attention mechanisms still exhibit competitive performance. For instance, the lightweight combination of MobileNetV3 and CBAM achieves a mean AUC of 0.76 with a significantly reduced parameter size (only 12.4MB), approaching the performance of ViT-based models. This highlights its strong potential for clinical deployment where computational efficiency is critical. In addition, a graphical user interface (GUI) has been developed to allow users to input chest X-ray images for preliminary diagnostic classification. The system also integrates Grad-CAM++ and Self-Attention Attribution to generate model attention heatmaps, thereby enhancing interpretability and assisting in the localization of pathological regions.

Index Terms—Attention mechanisms, Convolutional Neural Network(CNN), Deep learning, Disease detection, Grad-CAM++, Medical imaging, Self-Attention Attribution, Vision Transformer(ViT)

1 Introduction

Medical image analysis plays a pivotal role in the modern healthcare system, especially in the early diagnosis and accurate treatment of diseases. Traditional medical image analysis methods mainly rely on the professional knowledge and rich experience of clinicians for manual interpretation, which can meet clinical needs to some extent, but there are still many limitations in terms of efficiency and precision. Common lung diseases such as COVID-19, tuberculosis, and pneumonia, for example, remain a global public health problem that threatens the lives of millions of people [1]. In traditional chest radiographic examinations, the identification of lung abnormalities often relies on subjective judgment of radiologists, which is not only time-consuming and cumbersome, but also susceptible to differences in individual experience, resulting in subjective and inconsistent diagnostic results. Moreover, prolonged workloads can lead radiologists to fatigue-induced diagnostic errors, further compromising the accuracy and consistency of their assessments.

1.1 Study Background

In recent years, with the development of artificial intelligence technology, especially deep learning, new solutions have been provided for intelligent analysis of medical images. Among them, Convolutional Neural Networks, as an important architecture of deep learning, have demonstrated excellent feature extraction and classification capabilities in image processing tasks, have been widely used in the automated detection and assisted diagnosis of lung diseases, and have achieved preliminary and encouraging results in several studies [2].

However, CNNs exhibit significant limitations in modeling long-range contextual dependencies due to their inherently local receptive fields [3]. In recent years, Vision Transformers (ViTs) and hybrid architectures such as ConvNeXt models have achieved remarkable success in various computer vision tasks. Nevertheless, within the domain of medical image analysis, systematic and comprehensive comparisons among these three classes of models—namely, conventional CNNs, pure ViT architectures, and hybrid CNN-transformer models—remain relatively scarce. Therefore, this study is dedicated to thoroughly evaluating and analyzing the performance of these model categories in the context of medical image classification.

Despite significant advances in deep learning methods in medical image analysis, their practical deployment remains challenged by several critical limitations. Firstly, deep neural networks typically require large-scale, high-quality annotated datasets to achieve optimal performance. However, in the medical imaging domain, data acquisition is often constrained by privacy concerns, the high cost and

time intensity of expert annotation, and severe class imbalance. Secondly, most deep learning models require training from scratch, which demands substantial computational resources and high-performance hardware—conditions that are often unavailable in resource-limited clinical settings [4]. Finally, most deep learning models lack adequate interpretability and explainability, making their decision-making processes difficult for clinicians to understand and trust, which further limits their adoption and deployment in real-world clinical environments. Furthermore, deep learning models are susceptible to overfitting, particularly when trained on small or imbalanced datasets, and often exhibit poor generalisability and limited transferability. These problems hinder their ability to deliver consistent and robust performance across different imaging devices, healthcare institutions, and patient populations.

To address these challenges, researchers are increasingly turning to transfer learning (TL), more specifically deep transfer learning (DTL), as a promising alternative [5]. Instead of training models from scratch, DTL leverages prior knowledge acquired from large-scale, general-purpose datasets (e.g., ImageNet) [6] by employing pre-trained models as feature extractors for downstream tasks within the medical imaging domain. This paradigm substantially reduces the dependence on extensive labelled medical datasets, while simultaneously enhancing model generalisation and alleviating overfitting in data-constrained scenarios. Although the application of transfer learning in medical imaging has achieved preliminary results, it is still necessary to pay attention to the semantic differences between the pre-trained model and the target task, and its migration effect may be affected by feature space inconsistency or domain bias. Therefore, how to design more robust and efficient migration strategies remains one of the key issues in current research.

1.2 Research Objectives

In view of the above, this study aims to evaluate the effectiveness of transfer learning for the classification of chest X-ray images using a variety of deep learning architectures. The investigation encompasses several convolutional neural network models, including VGG16 [7], ResNet101 [8], InceptionV3 [9], DenseNet121 [10], and MobileNetV3Large [11], as well as a Vision Transformer-based model [3] and a hybrid model, ConvNeXt [12]. To enhance the feature extraction capabilities of the CNN-based models, two attention mechanisms—Attention Gate (AG) and Convolutional Block Attention Module (CBAM)—are independently integrated and assessed. The main objectives of this study are as follows: (1) to evaluate and compare the performance of different CNN, ViT and hybrid architectures within a transfer learning framework; (2) to investigate the impact of attention mechanisms on classification accuracy and model robustness; (3) to identify the most suitable architecture for multi-label chest disease classification under limited data conditions; and (4) in order to support post-processing and interpretability, a graphical user interface (GUI) for AUC-based model prediction threshold optimisation, and interpretability techniques such as Grad-CAM++ and self-attention attribution were employed to visualise and localise pathological regions in chest X-ray images.

1.3 Project Achievement

Ultimately, this study presents a comprehensive comparative analysis of multiple deep learning models within a unified transfer learning framework. The experimental results yield several key insights. First, older CNN architectures such as VGG16, ResNet101, and InceptionV3 exhibit relatively poor performance compared to more recent models. While state-of-the-art networks such as DenseNet121 and MobileNetV3Large consistently achieve strong evaluation scores, their performance remains inferior to that of Vision Transformer (ViT)-based and hybrid (ConvNeXt) models. Among all evaluated architectures, the ViT-based and ConvNeXt models demonstrated the highest overall classification performance, in addition to exhibiting the shortest training times. Notably, the MobileNetV3Large model, when augmented with attention mechanisms, achieved performance comparable to ViT and ConvNeXt, while maintaining a significantly smaller parameter size and lower computational complexity. This finding suggests that attention-enhanced lightweight CNNs offer a compelling alternative to more complex models, particularly in resource-constrained medical imaging environments. Nevertheless, it is important to note that the training time for such models—especially those integrated with attention modules—remains relatively long, which may pose limitations for time-sensitive clinical applications or iterative development workflows.

This project conducts a comprehensive evaluation of both conventional and state-of-the-art deep learning architectures for multi-label classification of chest X-ray images. By systematically integrating and analysing attention mechanisms such as Attention Gate (AG) and Convolutional Block Attention Module (CBAM), the study not only enhances model performance and interpretability but also highlights the potential of lightweight CNNs as viable alternatives to more complex architectures in resource-constrained clinical settings. Moreover, the development of a user-friendly interface for threshold optimisation, combined with visual explanation techniques, further improves the clinical applicability of the proposed models. These contributions are expected to advance the development of accurate, efficient, and interpretable AI-based diagnostic tools for chest disease detection, thereby supporting early diagnosis and clinical decision-making in real-world healthcare environments.

2 Related Work

2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have long been the cornerstone of deep learning in image processing tasks, demonstrating exceptional performance in tasks such as image classification, object detection, and segmentation. By employing convolutional layers, CNNs apply filters (or kernels) to local regions of the input data, enabling them to automatically detect spatial patterns, textures, and hierarchical features within images. Over the years, various architectures, such as VGG, ResNet, DenseNet, and MobileNet, have been developed and widely adopted, particularly in the field of medical image analysis.

2.1.1 VGG [13]

Previous studies have predominantly focused on the application of various models and datasets. The VGG model was proposed in 2015, but it was not until the COVID-19 pandemic in 2021 that Naveen et al. applied it to binary classification of pneumonia versus normal. They utilized the VGG16 architecture along with chest X-ray images. The model achieved a test accuracy of 95.67% on the pneumonia dataset through the application of data augmentation [13]. However, this study primarily concentrated on the binary classification task of pneumonia versus normal, without addressing the detection of other potential diseases. Additionally, the dataset used was relatively homogeneous, which may constrain the model's generalizability and limit its performance in broader real-world applications.

2.1.2 ResNet [14]

Notably, back in 2018, Ivo M. Baltruschat et al. used a network architecture that outperforms the VGG model and is powerful, and applied it to the problem of multiple classification of chest X-ray images: ResNet, which employs Residual Learning to solve the common deep neural network problems of gradient vanishing and gradient explosion by introducing skip-connections, allowing the network to go deeper and learn more complex features efficiently [14]. Experimental results show that the ResNet model performs well in a multi-label classification task on the Chest X-ray 14 dataset [15], effectively extracting features from complex X-ray images. In addition, different depths of ResNet architectures, such as ResNet-38 and ResNet-101, were also tried, resulting in an average AUC of 0.75 for all three models on the dataset [16]; however, this study was limited to the use of the ResNet model only, and did not analyse it in comparison with other deep neural networks. Since ResNet has a deeper architecture and a larger number of parameters than other networks, it requires more computational resources. For resource-limited environments (e.g., clinical settings), hardware limitations may be encountered, which may affect inference speed or make deployment difficult. "Moreover, due to their inherent local receptive fields, they struggle to capture global contextual features, thereby limiting their performance in pulmonary disease classification tasks.

2.1.3 DensenNet [17]

Subsequently, Dipkamal Bhusal et al. in 2022 proposed another methodology that attempts to use different convolutional neural networks applied to the task of multi-classification of chest X-ray images, specifically DenseNet (Dense Convolutional Neural Network). Their proposed model achieved the highest AUC score of 0.896 and accuracy of 0.826 in Cardiomegaly, while in Nodule classification, the model had the lowest AUC score of 0.655 and accuracy of 0.66 [17]. Nevertheless, the average AUC value of the model was still high at 0.75. However, it has the same limitations as the previous two studies in that they both used only one model and did not analyse the strengths and weaknesses of the model in comparison with other models, and this study did not have the assessment and validation of the model by a healthcare professional, so the usefulness of the model is unknown. Unlike previous studies, the heatmap technique (specifically, Grad-CAM) was used to visualise the area of interest of the model on the X-ray images, which helped to reveal the feature areas that the model relied on to make certain predictions. This provides further support for the interpretability of the model. Similarly, DenseNet shares a common limitation inherent to ResNet and other CNN-based models—namely, the inability to capture global contextual information due to their inherently local receptive fields.

2.1.4 MobileNet [18]

Due to the deep architecture of ResNet and DenseNet, these models typically have a large number of parameters and high computational complexity. As a result, they may face limitations in resource-constrained environments. To address this issue, Reshan et al. in 2023 proposed the MobileNet model, which is optimized for environments with limited computational resources. The study also evaluated the performance of eight pre-trained models, including ResNet50, ResNet152V2, DenseNet121, DenseNet201, Xception, VGG16, EfficientNet, and MobileNet. The results indicated that the MobileNet model outperformed the other models, achieving the highest accuracy of 94.23% and 93.75% on two different datasets, which contain 5,856 and 112,120 chest X-ray images, respectively [18]. However, it is worth noting that the study focused solely on a binary classification task distinguishing between normal and severe pneumonia. Therefore, there is still a significant gap in research on multi-class classification tasks for chest X-ray images.

2.1.5 Multi-model comparison

With the continuous improvement and development of various neural network models, researchers have gradually recognized the limitations of individual models and have begun to focus on comparative analysis of different models. To gain a more comprehensive understanding of the generalization ability of these models in real-world applications, Md Abu Sufian et al. in 2024 analyzed the applicability of DenseNet121, ResNet50, and CheXNeXt models using the NIH Chest X-ray 8 dataset. The study found that the CheXNeXt model outperforms both DenseNet121 and ResNet50 in most pathological tasks [19]. (It is worth noting that CheXNeXt, proposed by Pranav Rajpurkar et al. in 2017, is based on a dense convolutional neural network and incorporates advanced techniques such as self-training and ensemble learning, distinguishing it from traditional transfer learning approaches [20].) Additionally, the study compared the performance of DenseNet121 with radiologists' evaluations, revealing that deep learning model achieves a higher AUC value in certain diseases than the radiologists' assessments. For instance, in the detection of Cardiomegaly, the DenseNet121 model achieved an AUC of 0.888, surpassing the radiologists' AUC of 0.831. These findings highlight the potential of deep learning models, particularly the CheXNeXt model, in providing accurate diagnostic support, potentially surpassing traditional radiologists' assessments in certain areas of medical image analysis.

Although convolutional neural network (CNN)-based models have made significant advances and achieved outstanding performance in chest X-ray image classification

tasks across multiple studies, they still exhibit inherent limitations due to their reliance on local receptive fields. In particular, CNNs struggle to capture long-range dependencies and model global contextual information effectively. These shortcomings become especially pronounced when dealing with complex lesions and multiple coexisting pathological features in medical imaging.

TABLE 1
Related Research of CNN Models

Model	Year	Key Features	Task	Best Performance (AUC / Accuracy)
ResNet	2018	Residual connections, solves gradient vanishing/explosion	Multi-label classification	Average AUC: 0.75
VGG16	2021	Simple stacked convolutional layers	Binary classification	Accuracy: 95.67%
DenseNet	2022	Dense connections, improves information flow	Multi-label classification	Average AUC: 0.75
MobileNet	2023	Depthwise separable convolution, lightweight design	Binary classification	Accuracy: 94.23%
CheXNeXt	2024	Dense network with self-training and ensemble learning	Multi-label classification	AUC: Higher than DenseNet121, ResNet50

Table 1 summarizes the related research work of all the convolutional neural network models discussed above.

2.2 Transformer-based models

As the limitations of traditional CNNs in capturing global contextual information have become increasingly apparent, researchers have begun to explore alternative architectures that better address these shortcomings. Transformer-based models have emerged as a promising class of models with great potential in recent years. Unlike CNNs, which heavily rely on local receptive fields, Transformers leverage self-attention mechanisms to capture long-range dependencies and global context, making them particularly effective for tasks that require a more comprehensive understanding of the data. Transformer-based models have shown significant improvements not only in natural language processing but also in recent advancements in computer vision tasks, including image classification, detection, and segmentation.

In the domain of chest X-ray analysis, Transformer-based architectures, such as ViT [3] and so on, have started to outperform traditional CNNs in certain aspects, presenting opportunities to further enhance the accuracy and efficiency of medical image classification. The following sections will explore the related research on Transformer-based models, particularly in the area of multi-label classification tasks for medical images, and examine their potential advantages over traditional CNN architectures.

2.2.1 ViT

The Visual Transformer model, originally proposed by Dosovitskiy et al., replaces the traditional backbone of CNNs with a transformer-based architecture specifically designed for the task of image classification.ViT is able to capture long-range dependencies and global contexts by dividing an image into fixed-size patches and spreading these patches into sequences for input to the transformer. The core innovation of ViT is the use of self-attention, a mechanism that allows the model to process information globally and learn complex features of an image efficiently [3].

2.2.2 Swin Transformer

Building on this, Sina Taslimi et al. proposed a multi-label classification deep model with Swin Transformer as the backbone in 2022. Unlike ViT, the Swin Transformer divides the image into non-fixed-size sliding windows and applies the self-attention mechanism at different scales, enabling it to extract features at multiple levels. The model employs multiple MLP layers in the head configuration, and each layer achieved highly competitive AUC scores across all categories. Comprehensive experiments on the Chest X-ray 14 dataset demonstrated that the model with a 3-layer head configuration achieved state-of-the-art performance, with an average AUC score of 0.810 [21].

2.2.3 LT-ViT

In the following year, Umar Marikkar et al. went on to develop LT-ViT based on ViT, which is also a transformer architecture that exploits the combined attention between image tokens and randomly initialized auxiliary tokens that represent labels. LT-ViT was evaluated on two publicly available CXR datasets (NIH Chest X-ray 14 and CheXpert-13), both of which perform much better than using a pure ViT model, with AUCs as high as 0.81 and 0.72, respectively [22]. Moreover, the study conducted separate evaluation analyses on two different datasets, which significantly improved the generalisation performance of the model. By testing the LT-ViT on both the NIH Chest X-ray 14 and CheXpert-13 datasets, the researchers ensured that the model was not overfitting to a single dataset, thereby validating its robustness across diverse clinical settings and varying data distributions.

2.2.4 ViT (Enhanced)

Similarly, in 2024, Lan Huang et al. made three major improvements to the ViT model. These included enhancing the recognition of small lesions through the sliding window method, incorporating an attention region selection module into the ViT encoder to focus on key areas, and constructing a parallel patient metadata feature extraction network to integrate multimodal information. These enhancements collectively improved the accuracy of image classification and the overall performance of the model [23]. The final model achieved an average AUC of 0.831 on the Chest X-ray 14 dataset, making it the best-performing model in related studies.

TABLE 2
Related Research of Transformer-based models

Model	Year	Key Features	Task	Best Performance (AUC / Accuracy)
ViT	2021	Fixed-size patches, self-attention for global context	Image classification	N/A
Swin Transformer	2022	Sliding windows, multi-scale self-attention, MLP layers in head	Multi-label classification	Average AUC: 0.81
LT-ViT	2023	Combined attention between image tokens and auxiliary tokens representing labels	Multi-label classification	Average AUC: 0.81
ViT (Enhanced)	2024	Sliding window, attention region selection, parallel patient metadata network	Multi-label classification	Average AUC: 0.83

Table 2 summarizes the related research work of all the Transformer-based models discussed above. It is evident that their overall performance is superior to that of CNN models.

In recent years, an increasing number of studies have shifted towards Transformer-based models, primarily due to their performance being significantly superior to traditional CNN models. However, it is important to note that, compared to CNNs, Transformer models have more demanding

computational requirements. This is because they rely on self-attention mechanisms to capture global context, which necessitates processing a large number of parameters and matrix operations. Additionally, the training and inference processes of Transformer models often require substantial memory and computational resources, which may result in suboptimal performance in hardware-constrained environments. Therefore, balancing the strengths and weaknesses of CNNs and Transformer models remains a key focus of current research. It is anticipated that future studies will increasingly explore hybrid models that combine the advantages of both CNNs and Transformer architectures.

3 METHODOLOGY

3.1 Dataset Overview

The dataset used in this project is the Chest X-ray 14 dataset [15], a publicly available large-scale chest X-ray image dataset widely used for multi-class and multi-label classification tasks in medical image analysis. This dataset was provided by the Clinical Center of the National Institutes of Health (NIH) and contains 112,120 X-ray images from 30,805 patients, each labeled with annotations for the corresponding diseases [15]. The dataset is a publicly available de-identified dataset released by the NIH, and its use does not require additional ethical approval. The dataset includes 15 label categories: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, Pneumothorax, and No Finding. The first 14 labels represent common chest diseases, and each X-ray image may be annotated with one or more of these conditions. The No Finding label indicates that none of the 14 diseases are present in the image.

As shown in Figure 1 below, The number of no finding images is significantly larger than that of other diseases, which may lead to an imbalanced dataset. This situation often affects the model's training effectiveness and performance. To address this issue, the study uses focal loss as the loss function, which will be explained in detail in the evaluation factors section.

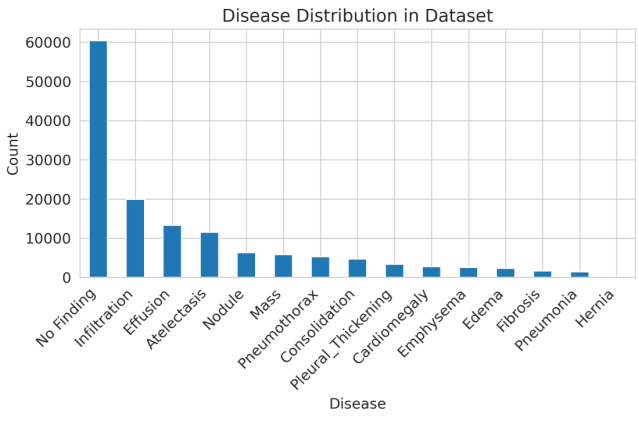

Fig. 1. Distribution of No Finding and the 14 diseases

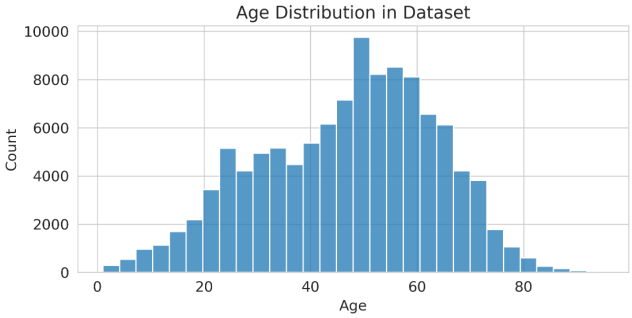

Fig. 2. Distribution of Patient Ages

3.1.1 Data splitting

In this project, the Chest X-ray 14 dataset is divided into three subsets: the training set, validation set, and test set, which are used for training, validation, and final evaluation. The test set is extracted from the official test list provided by the dataset, containing approximately 25,592 images. The remaining 80% of the data is allocated to the training set, which consists of approximately 69,209 images, while 20% is allocated to the validation set, which contains approximately 17,303 images.

3.1.2 Data Pre-processing

Before training the model, several preprocessing steps are performed to ensure that the input data is suitable for the deep learning model. First, it is ensured that all images are resized to a uniform dimension of 224×224 pixels. Next, each model (such as VGG16, ResNet, DenseNet, etc.) requires different preprocessing methods, which typically include pixel normalization, color channel adjustments, and so on. The code uses the preprocess_input function in Keras for each model to process the images. For example, VGG16 uses `vgg16_preprocess_input`, ResNet uses `resnet_preprocess_input`, and so on.

3.1.3 Data Augmentation

In medical image classification tasks, deep learning models are prone to overfitting due to the high cost of data acquisition and the imbalanced distribution of available samples, which ultimately leads to reduced generalization performance. To address this issue, data augmentation has been widely adopted as an effective strategy to expand the size of the training dataset and enhance both the robustness and generalization ability of models. It is important to note that data augmentation is applied only to the training set, while the validation and test sets remain unaltered to ensure the fairness and reliability of model evaluation.

In this project, data augmentation was applied by introducing a series of random transformations to the original images, thereby increasing the diversity of the training data and mitigating the risk of overfitting. Specifically, the augmentation process involved random horizontal flipping, random brightness adjustment, random contrast adjustment, random cropping, and scaling operations as shown in Figure 3. Through these random operations, the model was able to learn a wider range of image features, thus improving its robustness and accuracy in real-world applications. Moreover, data augmentation also significantly

enhanced the model's adaptability to variations in data distribution, such as chest X-ray images acquired from different hospitals or imaging devices. By learning more diverse feature representations, the model was able to maintain high predictive performance even when exposed to out-of-distribution samples.

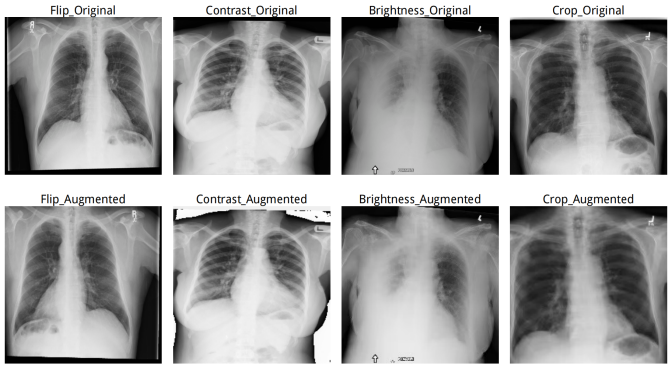

Fig. 3. Examples of augmented images from the Chest X-ray 14 dataset after applying various transformations, such as random flipping, brightness adjustment, contrast adjustment, and cropping.

3.2 Model Selection

In this study, model selection constitutes a critical component of the methodology. To thoroughly evaluate the task of multi-label classification for chest X-ray images, a diverse set of deep learning models were chosen for comparative analysis. These models span a range of widely utilized architectures, including Convolutional Neural Networks (CNNs), Transformer-based models, and hybrid architectures. While some of the currently popular models were briefly discussed in the related work section, this study also includes models that were not previously covered (such as InceptionV3). The models assessed in this study are as follows:

3.2.1 CNNs

The first and foremost models under consideration are those based on CNNs, which typically consist of several key components: input images, convolutional layers, pooling layers, and fully connected layers. The following figure 4 illustrates the general structure of a CNN. Although different models may introduce innovations or incorporate unique components based on this framework, their designs are predominantly modifications or extensions of this fundamental architecture. The subsequent sections will provide a detailed overview of the distinctive features of each model, along with their optimizations and improvements in practical applications.

1) **VGG16 [7]**: VGG16 is a deep convolutional neural network architecture initially proposed by the Visual Geometry Group (VGG) at the University of Oxford in 2014. The VGG16 network consists of 16 trainable weight layers, including 13 convolutional layers and 3 fully connected layers (i.e., the top layers of the network). In VGG16, the convolutional layers utilize very small 3x3 filters, and each convolutional layer is followed by a max-pooling layer (2x2 pooling) to reduce the spatial dimensions of the feature maps, thereby decreasing the computational load [7].

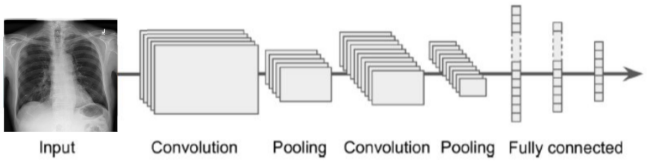

Fig. 4. The general architecture of CNN models: consisting of input, convolutional layers, pooling layers, and fully connected layers for feature extraction and classification.

2) **ResNet101 [8]**: ResNet101 is a deep residual learning architecture specifically designed for image recognition tasks, particularly effective for training very deep networks. It was introduced by Kaiming He et al. in 2015. The key innovation of ResNet lies in its use of residual connections, which help address the common issues of vanishing and exploding gradients in very deep neural networks. By introducing skip connections, ResNet allows for the direct transmission of gradient information, enabling the network to be deeper without a decline in performance [8]. As the name suggests, ResNet101 consists of 101 layers, including convolutional layers, pooling layers, and fully connected layers. The residual connections allow information to pass across layers, making the network easier to train and enhancing its overall performance.
3) **InceptionV3 [9]**: InceptionV3 is a deep CNN architecture initially proposed by researchers at Google, and was detailed in the 2015 paper "Rethinking the Inception Architecture for Computer Vision." The core innovation of InceptionV3 lies in the so-called "Inception module." This module processes the input image in parallel using convolution filters of different sizes (such as 1×1, 3×3, and 5×5) along with pooling operations, and then concatenates their outputs. This structure enables the model to extract features at different scales at each stage, thereby enhancing its expressive power [9].
4) **DenseNet121 [10]**: DenseNet121 was originally proposed by Gao Huang et al. in 2017, with the aim of addressing the challenges of deep network training through dense connections: each layer receives input from all preceding layers, which helps mitigate the vanishing gradient problem and facilitates feature reuse, thereby enhancing the model's representational power and efficiency [10]. DenseNet121 is a member of the DenseNet family, consisting of 121 layers.
5) **MobileNetV3Large [11]**: Among all the models, MobileNetV3 is the only lightweight CNN architecture, initially proposed by Google Research in 2019, designed specifically for mobile and edge devices in resource-constrained environments. The version used in this project is MobileNetV3Large, which outperforms MobileNetV2 by 3.2% in ImageNet classification accuracy [11]. The key innovations of

MobileNetV3 include depthwise separable convolutions and the hard-swish activation function, which together successfully balance performance and computational efficiency.

3.2.2 ViT [3]

The next model under consideration is the Vision Transformer (ViT). As illustrated in Figure 5, the basic structure of ViT is presented, with each module representing the core components of the ViT architecture: image patching, positional encoding, and self-attention layers. Unlike CNN-based models, ViT employs a Transformer architecture, which has demonstrated exceptional performance, particularly in image classification tasks. In this study, four versions of ViT are evaluated: Base 16, Base 32, Large 16, and Large 32.

ViT-B16 and ViT-B32 correspond to models that divide the input image into 16 patches (16x16 pixels) and 32 patches (32x32 pixels), respectively. This division impacts the level of detail captured in each patch. Additionally, the terms Base and Large refer to models of different sizes, with the Large model containing more parameters and layers than the Base model, thereby possessing enhanced representational capacity. However, this increase in capacity also leads to higher computational complexity and longer training times.

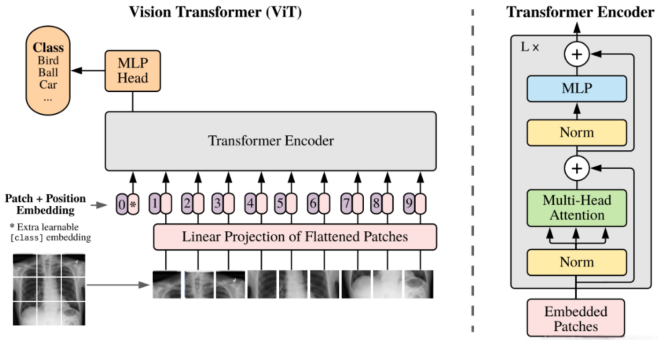

Fig. 5. The architecture of ViT: The ViT model partitions the image into patches, adds positional encodings, and processes them through multiple Transformer encoder layers.

3.2.3 ConvNeXt [12]

ConvNeXt is a hybrid architecture introduced by Zhuang Liu et al. in 2022. It is entirely constructed from standard convolutional neural network (CNN) modules, while also incorporating design ideas from transformers, such as combining local and global feature attention. ConvNeXt achieves competitive performance in terms of both accuracy and scalability, comparable to transformers. On the ImageNet dataset, it achieved a top-1 accuracy of 87.8% [12]. In this study, three versions of ConvNeXt were evaluated: Small, Base, and Large. These versions differ in terms of the number of layers, the number of parameters, and computational complexity, making them suitable for varying computational resources and performance requirements. The figure 6 below illustrates the architecture of the ConvNeXt model.

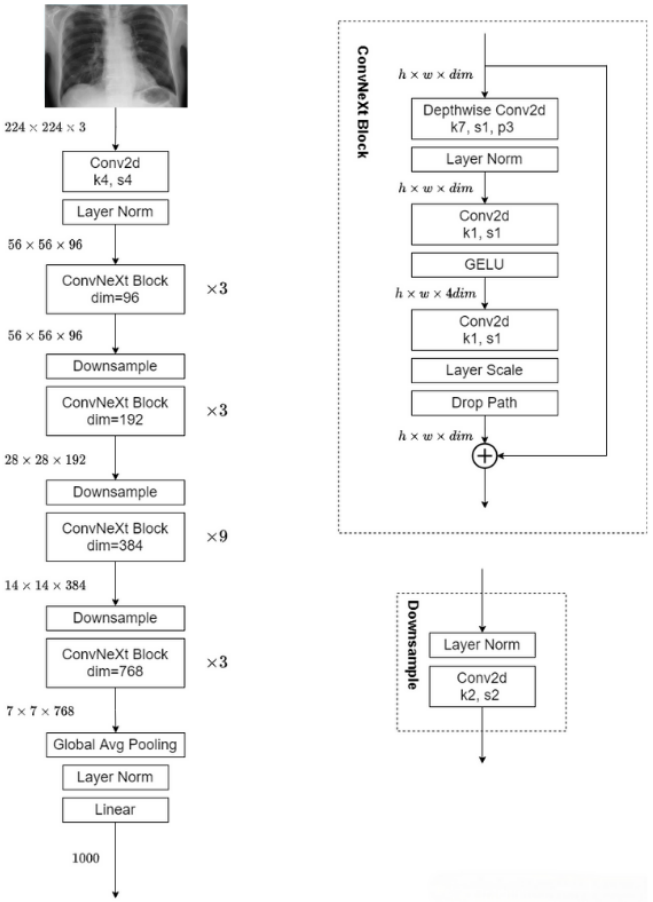

Fig. 6. The architecture of ConvNeXt: The ConvNeXt block is the core building unit of the ConvNeXt architecture, comprising depthwise separable convolutions, layer normalization, GELU activation function, layer scaling, and Drop Path.

3.2.4 Summary

Each model was trained using transfer learning, leveraging pre-trained weights from the ImageNet dataset, followed by fine-tuning on the Chest X-ray 14 dataset. This approach significantly reduces the reliance on large labeled datasets and accelerates convergence during the training process.

This study evaluates a variety of deep learning architectures, including CNN, ViT, and hybrid architectures (ConvNeXt), with a comparative analysis. The selection of these models was based on their versatility, performance in prior studies, and ability to address the complexities of multi-label classification in medical image analysis. By assessing a diverse set of models, this study aims to identify the most suitable architecture for the given task, balancing accuracy, efficiency, and generalization performance.

The models selected in this study encompass some of the mainstream architectures in the field of deep learning, particularly in the application of medical image analysis. However, with the rapid advancement of deep learning technologies, there are still numerous other models or emerging architectures that may be better suited to address specific challenges in medical image analysis, particularly in handling more complex multi-label classification tasks. For instance, models based on self-supervised learning or graph neural networks (GNNs) could offer new perspectives for

medical image analysis. Therefore, exploring the applicability of alternative model architectures and investigating how to integrate the strengths of different models remains a promising avenue for further research.

3.3 Attention Mechanism

In deep learning, attention mechanisms are widely used for their ability to significantly enhance model performance. The core idea is to focus on the most relevant regions of the input data while suppressing irrelevant areas, thereby improving the model's feature extraction capabilities. For instance, in the case of chest X-ray images in this project, the attention mechanism allows the model to focus more on regions of interest, such as the heart and the central areas of the lungs, while suppressing irrelevant regions, such as the background or areas like the arms. This improves the overall performance of the network. This section will explore two commonly used attention mechanisms: Attention Gate (AG) and Convolutional Block Attention Module (CBAM).

3.3.1 Attention Gate: AG

The Attention Gate operates by applying weights to the input feature map. Its core idea is to compute an importance coefficient for each pixel location based on a specific "gating signal." This coefficient determines whether the pixel is crucial for the final output [24]. For example, in chest X-ray images, the gating signal computes lower coefficients for irrelevant background pixels compared to important areas such as the heart. As a result, these coefficients are used to selectively suppress or enhance corresponding regions of the feature map, thus enabling the model to focus more on the relevant areas.

The specific operation of the Attention Gate module begins by applying a linear transformation to the input feature map and the gating signal, typically achieved through a 1×1 convolutional layer for dimensionality reduction. The transformed features are then combined and passed through a ReLU activation function to introduce non-linearity. This step helps the model learn complex patterns in the data. Subsequently, a Sigmoid activation function is applied to generate an attention mask, which has values between 0 and 1, representing the importance of different regions in the feature map. The attention mask is element-wise multiplied with the input feature map, emphasizing relevant regions and suppressing irrelevant ones. To prevent the loss of important features and maintain information, the output of the AG operation is added back to the original input feature map via a skip connection. This process helps retain the key information filtered by the attention mechanism. As shown in the figure 7, the schematic diagram of the entire AG operation is presented.

3.3.2 Convolutional Block Attention Module: CBAM

Another attention module is the Convolutional Block Attention Module (CBAM), which is a simple yet effective feed-forward convolutional neural network attention module. Given an intermediate feature map, the module sequentially infers attention maps along two distinct dimensions: channel and spatial. These attention maps are then multiplied

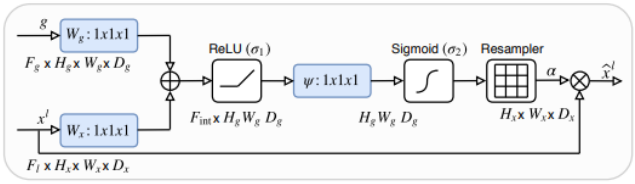

Fig. 7. The AG schematic diagram includes key components such as the input features (x^l), gating signal (g), ReLU activation function (σ_1), Sigmoid activation function (σ_2), and other important elements.

element-wise with the input feature map to perform adaptive feature refinement [25]. Unlike the AG, the operation of CBAM consists of two main steps: the channel attention module and the spatial attention module.

The channel attention module first extracts the inter-channel dependencies by compressing the spatial dimensions of the feature map. Specifically, the channel attention module aggregates each channel of the feature map using Global Average Pooling (GAP) and Global Max Pooling (GMP). Then, these two pooling results are concatenated and processed through a series of fully connected layers, ultimately generating a channel attention map. The values of this attention map lie between 0 and 1, representing the relative importance of each channel. The channel attention map is then element-wise multiplied with the input feature map, enhancing the features of important channels and suppressing less important ones.

The feature map is then passed into the spatial attention module. The spatial attention module generates a spatial importance map by considering the spatial information of the feature map. It first applies Global Average Pooling (GAP) and Global Max Pooling (GMP) across the channels, producing two spatial description maps. These two maps are then concatenated and processed through a convolutional layer to generate a spatial attention map. The spatial attention map reflects the importance of different spatial locations and is element-wise multiplied with the input feature map to emphasize critical spatial regions and suppress irrelevant ones.

Throughout the process, the CBAM module sequentially integrates channel attention and spatial attention, which not only focuses on feature selection along the channel dimension but also enhances feature representation in the spatial dimension as shown in the figure 8. In this way, the model can more adaptively adjust which features are most critical for the task, thereby improving overall performance.

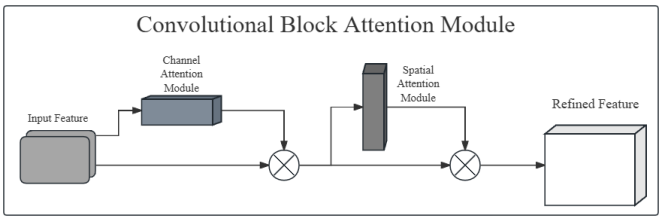

Fig. 8. The workflow diagram of CBAM: It shows how the feature map is processed through the channel attention module, then the spatial attention module, and finally results in the enhanced feature map.

3.3.3 Integration of Attention Mechanism into the Model

Both the Attention Gate (AG) and the Convolutional Block Attention Module (CBAM) are lightweight and architecture-agnostic, making them well-suited for integration into Convolutional Neural Network (CNN)-based models. In this study, all CNN backbone architectures (including DenseNet121, ResNet101, MobileNetV3, etc.) were modified by inserting either an AG or CBAM module immediately after the final major convolutional block and before the global pooling layer. By incorporating attention mechanisms at this stage, the models are able to more effectively highlight salient feature regions while suppressing irrelevant or redundant information, thereby optimizing feature representations. Owing to the modular nature of AG and CBAM, their integration does not require significant alterations to the original network architectures. Overall, this end-stage integration strategy enhances the feature extraction capacity and classification performance of the models, while also improving their robustness and generalization across different tasks and datasets.

3.4 Evaluation Factors

After selecting the models to be studied, the next critical step is the performance evaluation of these models. In the context of multi-label multi-class medical image tasks, selecting appropriate evaluation metrics is crucial. To ensure the effectiveness of the models in complex tasks, this study primarily employs the loss function and AUC (Area Under the Curve) as the main evaluation metrics.

3.4.1 Focal Loss

In most cases, binary cross-entropy loss is commonly used as the evaluation metric for multilabel classification tasks. However, in this study, due to the severe class imbalance in the dataset (as mentioned earlier, the "No Finding" class contains a significantly larger number of samples compared to all other disease categories combined), the conventional cross-entropy loss is not suitable for handling this imbalance effectively. Therefore, to better address this issue, the Focal Loss function was employed as an alternative. Focal Loss mitigates the impact of easily classified examples (such as the "No Finding" class) and places greater emphasis on hard-to-classify examples (such as the other disease classes), which helps improve the model's learning performance, especially for minority classes.

Focal Loss [26] was specifically developed to address the issue of extreme class imbalance. It is derived from the binary cross-entropy (CE) loss, serving as the foundation for its formulation. The following is the formula for cross-entropy (CE) loss:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{if } y = otherwise \end{cases} \tag{1}$$

(Where p is the model's predicted probability for the class $y = 1$, and y is the ground truth label (with values of +1 or -1), and $p \in [0, 1]$.)

On this basis, a new probability p_t is introduced:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = otherwise \end{cases} \tag{2}$$

(Where p_t is the predicted probability for the correct class.)

Combining the above two formulas, the cross-entropy loss can be rewritten as:

$$CE(p, y) = CE(p_t) = -\log(p_t) \tag{3}$$

The Focal Loss introduces a modulation factor $(1 - p_t)^\gamma$ based on the cross-entropy loss, which dynamically adjusts the loss according to the model's predicted probability. This factor increases the contribution of difficult-to-classify examples (low-probability samples), guiding the model to focus more on these challenging samples. The definition of Focal Loss is as follows:

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \tag{4}$$

As shown in the formula above, when p_t is small, the modulation factor approaches 1, and the loss is large. When p_t is large, the modulation factor approaches 0, and the loss is small. In this way, the loss function reduces the impact of easily classified samples (such as the background class). By adjusting γ, the strength of the modulation factor can be controlled. For example, when $\gamma = 2$, the loss for misclassified samples increases, while the loss for easily classified samples decreases. This makes the model focus more on hard-to-classify samples.

In all the experimental results of the models in this project, the aforementioned definition of Focal Loss was employed. Specifically, the focusing parameter γ was set to 2, following the original paper [26]. However, the exact formulation is not critical to the overall performance.

3.4.2 Area Under the Curve: AUC

The second metric used to evaluate model performance is the Area Under the Curve (AUC). AUC is derived from the Receiver Operating Characteristic (ROC) curve, which illustrates the relationship between the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings [27].

Specifically, the True Positive Rate (TPR, sensitivity) and the False Positive Rate (FPR, 1-specificity) are computed as follows:

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

where:

- TP is the number of true positives,
- FN is the number of false negatives,
- FP is the number of false positives,
- TN is the number of true negatives.

TPR represents the proportion of actual positive samples correctly identified by the model, while FPR represents the proportion of actual negative samples incorrectly classified as positive by the model. By plotting the ROC curve, we can visually evaluate the model's performance at various decision thresholds, with AUC providing a quantitative assessment of overall classifier performance.

Furthermore, one advantage of the ROC curve is its ability to visualize and organize classifier performance in

multi-label classification tasks, without considering the distribution of labels or the cost of errors. This capability is crucial when handling multi-label classification tasks, as it allows researchers to evaluate the model's performance on each individual label. By plotting the performance of a set of classifiers, researchers can ensure that the ROC curve remains consistent across different evaluation conditions for various labels. Although the regions of interest may change with varying label distributions or error costs, the structure of the ROC curve itself remains unaffected.

For multi-class AUC, Provost and Domingos proposed generating a reference ROC curve for each class [28], measuring the area under the curve, and then summing the AUCs while weighting them according to the prevalence of the reference class in the data. More precisely, they define:

$$AUC_{\text{total}} = \sum_{i=1}^{n} AUC(c_i) \cdot p(c_i) \tag{7}$$

where:

- n is the total number of labels,
- $AUC(c_i)$ is the AUC value for the label c_i,
- $p(c_i)$ is the sample proportion (or weight) for the label c_i.

In conclusion, the Area Under the Curve provides a robust metric for evaluating model performance in multi-label classification tasks. By combining AUC with the ROC curve, it allows for a more comprehensive assessment of the classifier's ability to differentiate between classes at various decision thresholds, especially when dealing with minority or hard-to-classify samples. Additionally, AUC can be used to evaluate the performance of different models across various disease classification tasks, offering a more comprehensive analysis compared to accuracy. Compared to accuracy, AUC-ROC is a better representation of the model's true performance, particularly in cases of class imbalance, where accuracy may fail to fully capture the model's ability to identify minority class samples. It is worth noting that while metrics such as recall, precision, and F1-score can also assess model performance, they are similarly based on the relationship between the True Positive Rate and False Positive Rate. Therefore, to avoid redundancy, these metrics are not used in this analysis.

3.5 Model Interpretability

Finally, to ensure that the model's predictions are understandable and trustworthy, two interpretability methods have been integrated into the graphical user interface (GUI): Grad-CAM++ [29] (for CNN models) and self-attention attribution [30] (for ViT models). These methods allow users to gain a more intuitive understanding of the model's decision-making process. The results of these methods are then superimposed onto heatmaps, revealing the model's attention on different regions and features during prediction. Ultimately, this visualization of the model's attention distribution helps users or clinicians more clearly identify the key areas the model focuses on in the image, thereby enhancing the model's transparency and reliability, and further improving its effectiveness and applicability in real-world scenarios. A brief introduction to these two methods will be provided below.

3.5.1 Grad-CAM++

Grad-CAM++ improves upon the original Grad-CAM method by providing a more accurate pixel-wise weighting of the convolutional feature maps. It integrates higher-order gradient information to precisely highlight the regions of the image that significantly contribute to the CNN's decision [29]. The final weighting formula used in Grad-CAM++ is defined as follows:

$$w_k^c = \sum_i \sum_j \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_{a,b} A_{ab}^k \frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}} \cdot \text{ReLU}\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{8}$$

where:

- w_k^c denotes the weight for feature map k with respect to class c.
- Y^c represents the prediction score for class c.
- A_{ij}^k is the pixel at location (i, j) in the convolutional activation map A^k.
- $\text{ReLU}(\cdot)$ is the rectified linear unit activation, retaining only gradients that positively influence the class prediction.

Grad-CAM++ thus allows for finer-grained visualization of CNN predictions, enhancing the interpretability and transparency of deep neural networks.

3.5.2 Self-Attention Attribution

Self-Attention Attribution provides a method for interpreting the internal decision-making processes of Transformer-based models, by explicitly attributing model predictions to their self-attention mechanisms. Specifically, this method quantifies the contribution of attention heads within Transformer layers by calculating an attribution score matrix [30]. The attribution score for the h-th attention head is formally defined as follows:

$$\overline{\text{Attr}}_h(A) = \frac{A_h}{m} \odot \sum_{k=1}^{m} \frac{\partial F\left(\frac{k}{m} A\right)}{\partial A_h} \tag{9}$$

where:

- $\overline{\text{Attr}}_h(A)$ denotes the attribution score matrix for the h-th attention head.
- A_h represents the attention weight matrix corresponding to the h-th head.
- $F(\cdot)$ is the forward function of the Transformer model.
- m is the number of discrete steps used in the integrated gradients approximation.
- k indexes each interpolation step from 1 to m for the discrete approximation of the integral.
- $\odot$ denotes element-wise multiplication.

By utilizing integrated gradients to approximate the importance of attention weights, Self-Attention Attribution captures detailed interactions among input features. Aggregating these scores across multiple heads and layers produces a comprehensive and intuitive visualization of internal information flow, substantially enhancing the transparency and interpretability of Transformer-based models.

TABLE 3
Experimental Environment Configuration

Component	Configuration
Operating System	Ubuntu 24.04
Python Version	3.9
TensorFlow	2.13.0
TensorFlow Addons	0.23.0
NumPy	1.24.3
Pandas	2.2.3
Matplotlib	3.5.3
Seaborn	0.13.2
Scikit-learn	1.6.0
OpenCV-Python	Installed
vit-keras	Installed
Hardware	NVIDIA A100 80GB PCIe GPU
Framework	TensorFlow with XLA and TF32 enabled

4 Results

4.1 Experimental Environment and Parameter

The deep learning framework adopted in this study is TensorFlow. All experiments were conducted on a high-performance computing server provided by the university, equipped with an NVIDIA A100 80GB PCIe GPU. This hardware setup supports TensorFloat-32 computations and XLA compilation optimizations, which significantly accelerate large-scale model training. A detailed summary of the experimental environment is presented in Table 3.

During model training, a consistent training strategy and set of hyperparameters were applied across all experiments. Prior to training, input images were resized to 224×224. The hyperparameter configuration was as follows: the batch size was set to 64, the Adam optimizer was employed with an initial learning rate of 1×10^{-6}, and the maximum number of training epochs was set to 300. An early stopping mechanism was incorporated to prevent overfitting, which monitored the performance on the validation set and automatically terminated training if no improvement was observed over 20 consecutive epochs. In addition, certain models were further enhanced by introducing Dropout (with a rate of 0.5) and Batch Normalization to improve generalization.

In addition to the training configurations, this study also recorded the number of trainable parameters for each model to assess the architectural complexity of their network structures. (As shown in Table 5 of the Evaluation section, there is a significant variation in model sizes.) For instance, MobileNetV3Large (11.49 MB) and ConvNeXt-Small (188.70 MB) represent lightweight architectures with relatively low memory, making them particularly well-suited for deployment in resource-constrained or edge-computing environments. In contrast, ViT-l16 and ViT-l32 both exceed 1.1 GB in parameter size, while ConvNeXt-Large reaches 748.65 MB, indicating stronger representational capacity at the cost of substantially higher computational and deployment demands.

4.2 Training and Testing Result

After completing the environment setup and hyperparameter configuration, all selected models were successfully trained and evaluated. Due to space limitations, only the ROC curve and training process of the ConvNeXt_base model are presented (as shown in the figure 9 and 10). The overall results of all models are summarized in the table 4, which reports the AUC scores for each architecture across 15 chest disease categories.

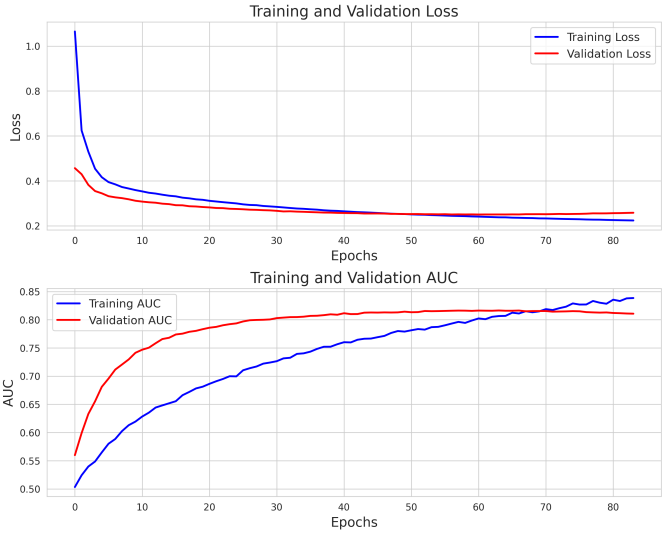

Fig. 9. Training and validation loss and AUC curves of the ConvNeXt-Base model.

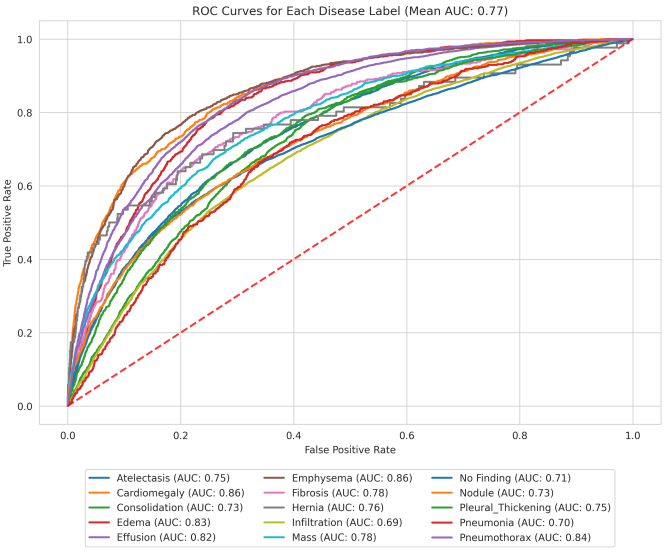

Fig. 10. ROC curves of the ConvNeXt-Base model on the test set across 15 chest diseases.

The experimental results indicate that the hybrid ConvNeXt models consistently achieve the most stable and highest overall performance, with all variants reaching an average AUC of 0.77. Closely following are the four Vision Transformer (ViT) models, which achieve average AUC values ranging between 0.75 and 0.77. Among them, models that divide the input into 16 patches (e.g., ViT-b16 and ViT-

TABLE 4
AUC Performance of Different Models on 15 Chest Diseases

Type	Model	Ate.	Car.	Con.	Ede.	Eff.	Emp.	Fib.	Her.	Inf.	Mass	NoF.	Nod.	PleT.	Pneu.	PneT.	Avg.
CNN	VGG16	0.73	0.79	0.69	0.82	0.79	0.74	0.72	0.69	0.68	0.72	0.71	0.69	0.69	0.64	0.79	**0.73**
CNN	VGG16+BN+AG	0.72	0.79	0.69	0.81	0.79	0.81	0.71	0.69	0.67	0.72	0.70	0.68	0.71	0.67	0.81	**0.73**
CNN	VGG16+BN+CBAM	0.72	0.79	0.70	0.80	0.79	0.79	0.74	0.75	0.66	0.71	0.70	0.69	0.70	0.65	0.80	**0.73**
CNN	ResNet101	0.73	0.81	0.70	0.82	0.79	0.75	0.72	0.66	0.68	0.71	0.70	0.67	0.71	0.66	0.79	**0.73**
CNN	ResNet101+BN+AG	0.72	0.80	0.70	0.80	0.79	0.78	0.73	0.59	0.66	0.71	0.70	0.66	0.71	0.65	0.80	**0.72**
CNN	ResNet101+BN+CBAM	0.73	0.82	0.71	0.81	0.79	0.75	0.73	0.64	0.66	0.69	0.70	0.67	0.69	0.65	0.77	**0.72**
CNN	InceptionV3	0.71	0.76	0.68	0.81	0.77	0.78	0.74	0.72	0.67	0.70	0.69	0.68	0.70	0.65	0.79	**0.72**
CNN	InceptionV3+BN+AG	0.72	0.75	0.68	0.79	0.78	0.77	0.73	0.66	0.66	0.72	0.69	0.68	0.70	0.65	0.79	**0.72**
CNN	InceptionV3+BN+CBAM	0.71	0.76	0.68	0.80	0.77	0.76	0.72	0.68	0.66	0.71	0.68	0.67	0.68	0.66	0.77	**0.71**
CNN	DenseNet121	0.73	0.84	0.72	0.81	0.81	0.83	0.75	0.75	0.68	0.74	0.71	0.69	0.72	0.68	0.82	**0.75**
CNN	DenseNet121+BN+AG	0.74	0.84	0.71	0.82	0.81	0.80	0.73	0.74	0.67	0.76	0.71	0.69	0.73	0.68	0.82	**0.75**
CNN	DenseNet121+BN+CBAM	0.74	0.84	0.71	0.82	0.80	0.82	0.77	0.74	0.68	0.75	0.71	0.71	0.73	0.67	0.83	**0.75**
CNN	MobileNetV3	0.72	0.82	0.70	0.81	0.80	0.79	0.75	0.78	0.67	0.73	0.70	0.68	0.72	0.65	0.81	**0.74**
CNN	MobileNetV3+BN+AG	0.73	0.83	0.70	0.82	0.80	0.84	0.75	0.78	0.67	0.74	0.70	0.69	0.72	0.67	0.83	**0.75**
CNN	MobileNetV3+BN+CBAM	0.74	0.83	0.70	0.82	0.80	0.85	0.77	0.78	0.67	0.74	0.70	0.69	0.72	0.66	0.84	**0.76**
Trans	ViT-b16	0.75	0.87	0.74	0.84	0.81	0.84	0.78	0.74	0.68	0.78	0.71	0.72	0.74	0.70	0.83	**0.77**
Trans	ViT-b32	0.74	0.86	0.72	0.84	0.80	0.81	0.77	0.74	0.67	0.75	0.70	0.68	0.73	0.68	0.83	**0.75**
Trans	ViT-l16	0.75	0.87	0.73	0.83	0.81	0.82	0.77	0.76	0.68	0.78	0.71	0.70	0.74	0.69	0.83	**0.76**
Trans	ViT-l32	0.74	0.85	0.72	0.83	0.80	0.79	0.75	0.72	0.68	0.74	0.71	0.68	0.73	0.66	0.82	**0.75**
Hybrid	ConvNeXt-S	0.75	0.85	0.72	0.82	0.82	0.87	0.77	0.81	0.67	0.76	0.71	0.72	0.74	0.69	0.86	**0.77**
Hybrid	ConvNeXt-B	0.75	0.86	0.73	0.83	0.82	0.86	0.78	0.76	0.69	0.78	0.71	0.73	0.75	0.70	0.84	**0.77**
Hybrid	ConvNeXt-L	0.75	0.86	0.72	0.82	0.81	0.85	0.78	0.81	0.68	0.77	0.71	0.72	0.72	0.68	0.84	**0.77**

l16) clearly outperform their 32-patch counterparts. Notably, ViT-b16 and ViT-l16 achieve the highest AUC of 0.87 in the "Cardiomegaly" category across all models.

In contrast, CNN-based models generally exhibit lower average AUC scores compared to ViT and hybrid architectures. Traditional CNN models such as VGG16 and ResNet101 only achieve an average AUC of approximately 0.73, and the incorporation of attention mechanisms yields minimal improvements. DenseNet121, however, performs relatively well among CNNs, achieving an average AUC of 0.75.

Compared to Vision Transformers (ViTs) and hybrid ConvNeXt architectures, Liu et al. (2022) pointed out that traditional convolutional neural networks are inherently constrained by their local receptive fields, limiting their ability to model global dependencies [12]. This limitation motivated the development of architectures such as ConvNeXt, which aim to bridge this gap by incorporating design elements that enable more effective global context modeling. Consequently, the local nature of CNNs becomes particularly problematic in chest X-ray classification tasks, where many pathological features span large spatial regions and require holistic contextual understanding.

But, an interesting observation arises from the MobileNetV3Large model. While its baseline average AUC is 0.74, the inclusion of the AG module improves the score by 0.01, and the further integration of the CBAM module results in an additional increase of 0.02—achieving an average AUC that matches or slightly surpasses the best performance achieved by ViT models. Given its significantly smaller number of parameters compared to ViTs, MobileNetV3Large presents a highly competitive and efficient solution for deployment in resource-constrained environments.

4.3 GUI

After the training and evaluation of all models were completed, a graphical user interface (GUI) was developed using the Streamlit framework to facilitate interactive model interpretation and prediction visualization. This GUI integrates two interpretability techniques introduced in the methodology section: Grad-CAM++ for convolutional neural network models and Self-Attention Attribution for Transformer-based models.

The interface provides two key functionalities: model selection and threshold customization. Users can flexibly choose among all trained models and specify which disease categories they are interested in. For each selected disease, the GUI allows users to either adopt the default threshold, which is automatically computed based on the optimal value derived from the ROC curve, or manually adjust it according to task-specific requirements.

The threshold is derived using Youden's J statistic [31], a widely adopted metric in medical statistics for identifying the optimal cutoff point on the ROC curve. This approach is designed to balance sensitivity (true positive rate, TPR) and specificity, thereby maximizing overall diagnostic effectiveness. The index is formally defined as:

$$J = \text{Sensitivity} + \text{Specificity} - 1 = \text{TPR} - \text{FPR} \tag{10}$$

where:

- **TPR (True Positive Rate)**: Also known as Sensitivity; the proportion of actual positives correctly identified by the model.
- **FPR (False Positive Rate)**: Equal to $1-$Specificity; the proportion of actual negatives incorrectly classified as positive.

By maximizing the J value, i.e., $J = \text{TPR} - \text{FPR}$, the threshold that provides the optimal balance between sensitivity and specificity can be selected. As show in Figure 11, when the DenseNet121 model is selected, the corresponding

prediction thresholds are automatically derived based on its average AUC score of 0.75, ensuring optimal decision boundaries for each disease category.

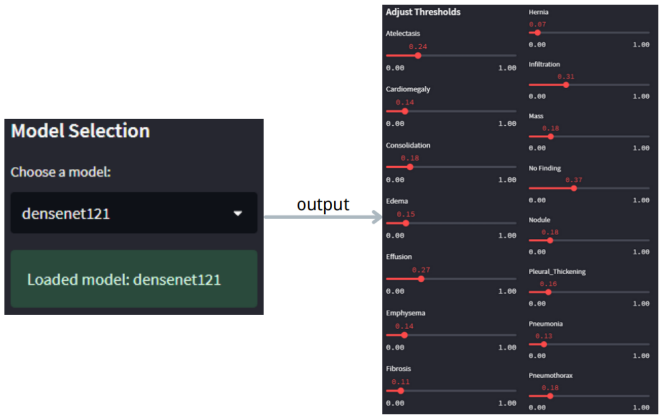

Fig. 11. The disease prediction thresholds used for each selected model are determined based on the optimal cutoff values calculated from the corresponding ROC curves.

Once a model and threshold configuration are selected, users can upload a chest X-ray image to perform inference. The system outputs the predicted disease labels exceeding the specified thresholds, along with corresponding interpretability heatmaps. These heatmaps highlight the most salient image regions contributing to the prediction. Additionally, the GUI automatically identifies and marks the region of highest activation using a bounding box placed in the center of the most informative area, enhancing spatial localization and interpretability. This GUI not only improves transparency and trust in the model's decision-making process but also offers a practical and user-friendly tool for potential clinical integration, enabling medical professionals to visualize and validate model predictions with minimal technical effort.

As illustrated in Figure 12, the input is a chest X-ray image with a ground-truth label of Cardiomegaly, and the selected model is DenseNet121. Under the currently configured thresholds, the model predicts both Cardiomegaly and Fibrosis as positive, with prediction scores exceeding their respective decision boundaries. Among them, Cardiomegaly yields a higher probability of 0.3411, indicating a more confident prediction.

Furthermore, the Grad-CAM++ visualization clearly highlights the most activated region located in the lower central area of the thoracic cavity, which spatially corresponds to the lower part of the heart. This alignment with anatomical priors provides strong evidence supporting the model's interpretability and the clinical plausibility of its decision-making process.

4.4 Limitations and Error Analysis

Despite the strong performance of the proposed framework across various deep learning architectures, several limitations and sources of error were observed during the experimental process and warrant further discussion.

First, certain disease categories—such as Fibrosis, Nodule, and Pleural Thickening—consistently exhibited lower

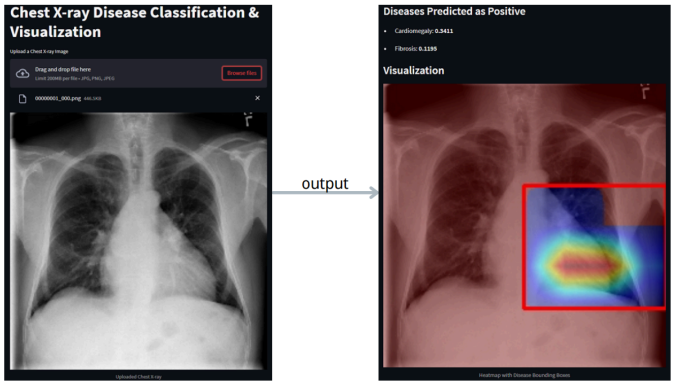

Fig. 12. The figure shows the original input image alongside and output the right label. The visualization generated using the Grad-CAM++ technique.

AUC scores across all models. The primary reason for this performance degradation lies in the severe class imbalance within the dataset, where positive samples for these categories are significantly underrepresented. Although Focal Loss was incorporated to dynamically reweight class contributions and mitigate the impact of this imbalance, experimental results indicate that its effectiveness remains limited in the context of extremely sparse classes. This may be attributed to the insufficient number of training samples required for robust feature learning, compounded by the inherently vague, overlapping, or ill-defined visual characteristics of these pathologies. Consequently, models exhibit reduced generalization capacity and a higher likelihood of false-negative predictions during testing.

For example, as shown in Figure Figure 13 below, the ground-truth label of the image is Nodule. However, under the prediction of the DenseNet121 model, the probability assigned to Nodule is merely 0.2022, while the probability of being classified as No Finding is as high as 0.4428. Moreover, the corresponding heatmap indicates that the model fails to focus on the pulmonary region, instead exhibiting excessive attention to peripheral areas.

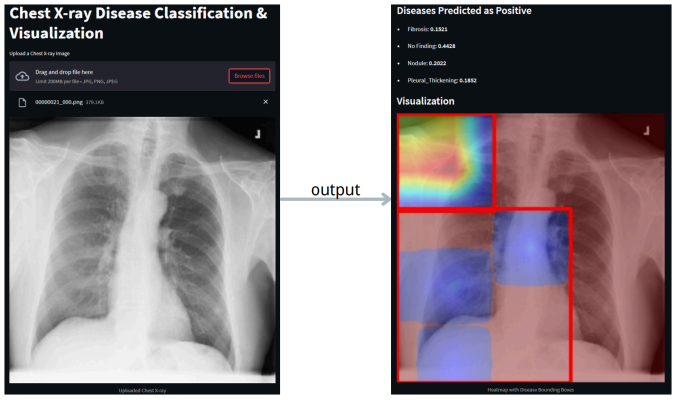

Fig. 13. The figure shows the original input image alongside and output the error label. The visualization generated using the Grad-CAM++ technique.

Second, while Grad-CAM++ and Self-Attention Attribution were employed to enhance the interpretability of CNN and Transformer-based models, these visualization

techniques are essentially heuristic in nature. In certain cases, the generated attention maps may highlight regions that lack clinical diagnostic relevance. This issue becomes particularly pronounced when pathological features are diffuse or span multiple anatomical regions, resulting in less accurate attention localization and diminished clinical trustworthiness.

Finally, one notable limitation arises in the GUI-based disease prediction when applying threshold-based multi-label classification. Although the system can successfully identify the correct disease in many cases, the absence of a strict ground truth constraint on the number of labels (i.e., the true number of conditions per image) introduces challenges. Since the model operates in a multi-label setting without being explicitly trained to predict only one or a fixed number of diseases, it may occasionally produce over-predictions. For example, images that contain only one or two ground-truth conditions may be predicted as having three or more diseases, leading to false positives despite partially correct outputs. In some edge cases, the system may even predict both "No Finding" and one or more disease labels simultaneously, which is logically contradictory and clinically implausible. These issues suggest that while the threshold-based GUI is effective in providing interpretable predictions, additional constraints or post-processing strategies may be needed to improve diagnostic precision and reduce misclassification rates.

5 Evaluation

This study comprises two key components: first, the evaluation of model performance during the training, validation, and testing phases, primarily measured using the Area Under the Curve (AUC) to assess classification effectiveness; and second, the development and application of a graphical user interface (GUI), which integrates visualization techniques and human-computer interaction principles to evaluate the system's usability, functional completeness, and potential for clinical decision support. The former focuses on a comprehensive quantitative analysis of the accuracy, robustness, and generalization capabilities of various deep learning architectures in multi-label chest disease classification tasks. In contrast, the latter relies more heavily on subjective user needs and perceptions, and is thus more appropriately assessed through qualitative analysis.

5.1 Quantitative Evaluation for Models

During the training, validation, and testing phases, model performance was primarily evaluated using the Area Under the Curve (AUC) as the core evaluation factor. The detailed performance of each model has already been presented in the Results section and will not be reiterated here. This section conducts a quantitative evaluation focusing on three key aspects: the number of trainable parameters, the actual number of training epochs, and the final AUC performance. As presented in Table 5, these statistics collectively provide a comprehensive overview of each model's architectural complexity, training efficiency, and classification effectiveness.

In the table, although the introduction of attention mechanisms to VGG16 did not lead to noticeable improvements in AUC performance, it significantly reduced the number of training epochs required for convergence. This suggests that attention modules may accelerate the training process. While the addition of attention mechanisms increases the parameter count, the overall model size remains relatively lightweight, making VGG16 more deployment-friendly compared to larger architectures such as ViT and ConvNeXt.

TABLE 5
Summary of Model Size, Training Epochs and AUC

Model	Size	Epochs	AUC
VGG16	64.4MB	103	0.73
+ AG	128.8MB	81	0.73
+ CBAM	128.7MB	77	0.73
ResNet101	162.9MB	55	0.73
+ AG	164.4MB	67	0.72
+ CBAM	166.9MB	68	0.72
InceptionV3	83.3MB	94	0.72
+ AG	84.8MB	92	0.72
+ CBAM	87.3MB	125	0.71
DenseNet121	26.9MB	167	0.75
+ AG	27.7MB	96	0.75
+ CBAM	27.9MB	124	0.75
MobileNetV3	11.5MB	250	0.74
+ AG	12.2MB	300	0.75
+ CBAM	12.4MB	265	0.76
ViT-b16	327.3MB	120	0.77
ViT-b32	333.7MB	185	0.75
ViT-l16	1.13GB	61	0.76
ViT-l32	1.14GB	111	0.75
ConvNeXt-S	188.7MB	155	0.77
ConvNeXt-B	334.1MB	84	0.77
ConvNeXt-L	748.7MB	63	0.77

In contrast, ResNet101 and InceptionV3 showed limited benefits from the incorporation of attention modules. The performance in terms of AUC, training efficiency, and model size remained largely unchanged, indicating that these architectures may be less responsive to such enhancements.

Notably, DenseNet121 and MobileNetV3, as two representative lightweight CNN architectures, demonstrated competitive performance in multi-label classification tasks, comparable to that of Transformer-based models. In particular, MobileNetV3 with CBAM achieved an average AUC of 0.76, which is close to that of ViT, while maintaining an extremely compact model size of only 12.4MB. However, its main limitation lies in the prolonged training time, as MobileNetV3 required nearly 300 epochs to reach optimal performance, the longest among all evaluated models.

Among Transformer-based and hybrid architectures, both ViT and ConvNeXt consistently delivered high classification performance (AUC $\approx$ 0.77), but with larger parameter sizes. For instance, ConvNeXt-Small, despite having a relatively compact size of 188.7MB, needed 155 epochs to converge. In contrast, ConvNeXt-Large achieved optimal performance in only 63 epochs, offering significant improvements in training efficiency despite its substantial model size of 748.7MB.

In summary, for scenarios with limited computational resources, lightweight models such as DenseNet121 and Mo-

bileNetV3 represent excellent balance between performance and deployability. For applications requiring fast convergence and shorter training times, ViT-l16 and ConvNeXt-Large are preferable choices. If classification performance is the primary concern, the ConvNeXt series offers the best overall balance between effectiveness and efficiency.

5.2 Qualitative Evaluation for GUI

The design and effectiveness of the graphical user interface are largely influenced by users' subjective perceptions and specific usage scenarios. As such, in contrast to the quantitative evaluation applied to model performance, GUI assessment is more appropriately conducted through qualitative analysis. To this end, the evaluation framework is based on Jakob Nielsen's ten usability heuristics for user interface design [32]. This well-established set of principles in the field of human-computer interaction (HCI) facilitates a systematic evaluation of the interface in terms of functional completeness, feedback mechanisms, and overall usability. The heuristic method provides valuable insights into key experiential dimensions encountered during user interaction with the system.

Specifically, the evaluation encompasses the following ten aspects:

- **Visibility of system status:** The interface allows users to switch between models and load them in real time, with automatic initialization of the corresponding optimal thresholds upon successful loading. During image upload and prediction, the system provides timely visual feedback (e.g., status messages and progress indicators), ensuring that users are continuously informed of the GUI's operational state and system responsiveness, thereby enhancing transparency and user confidence.
- **Match between system and the real world:** This GUI is designed in alignment with clinical workflows and the logic of medical image interpretation, employing standardized medical terminology (e.g., "Atelectasis," "Cardiomegaly") to ensure that clinical users can operate the system without additional learning costs. At the same time, intuitive visual icons (such as those for model selection and image upload) are used to facilitate natural understanding and ease of use, allowing users to quickly learn how to interact with the system.
- **User control and freedom:** Users are granted full flexibility to switch between different deep learning models, automatically adjust output thresholds to select target diseases, reset uploaded images or prediction inputs at any time, and re-run the inference process as needed. This operational flexibility facilitates iterative testing and comparison in clinical decision-making scenarios, thereby enhancing the system's practical usability and overall user experience.
- **Consistency and standards:** The layout style of the system interface is consistent, with fixed positions for core functionalities such as model selection, image upload, and threshold adjustment. This design enhances operational coherence and reduces the cognitive load on users.
- **Error prevention:** For the image upload functionality, only supported formats such as JPG, PNG, and JPEG are allowed, effectively preventing the submission of unsupported file types. This design choice minimizes the risk of invalid inputs and ensures compatibility with the model's expected input format.
- **Recognition rather than recall:** The system interface is designed with a strong emphasis on information visibility and operational cues. All key functions are implemented through clearly labeled buttons and textual prompts, thereby minimizing the user's reliance on memory regarding operational procedures or system states. In addition, the interface retains essential information such as the history of uploaded images and prediction results, facilitating users' ability to review and comprehend past actions. This design effectively reduces the learning curve and enhances overall operational fluency.
- **Flexibility and efficiency of use:** The interface accommodates a wide spectrum of users. Default threshold configurations enable users with limited domain knowledge to perform disease predictions with minimal effort, facilitating accessibility for general users. Meanwhile, expert users such as healthcare professionals are provided with the flexibility to manually adjust decision thresholds, allowing them to fine-tune predictions based on different diagnostic needs, such as distinguishing between benign and malignant conditions. This dual-level design supports both usability and task efficiency across diverse usage scenarios.
- **Aesthetic and minimalist design:** The interface is characterized by a minimalist design and a well-structured layout that focuses on essential functionalities. The deliberate use of whitespace and typographic hierarchy contributes to an aesthetically pleasing and visually comfortable user experience.
- **Help users recognize, diagnose, and recover from errors:** The system actively informs users when operations fail, such as displaying error messages like "Model not downloaded, please download the model" if a selected model is unavailable. Additionally, loading status indicators are shown during model switching or image upload, helping users understand the system state and respond accordingly. This design improves error transparency and supports effective recovery without user confusion.
- **Help and documentation:** Although a detailed user manual has not been developed, the system demonstrates strong self-explanatory capabilities. Clear textual prompts are embedded within the interface at key operational steps—such as model selection, optional threshold adjustment, and image upload—allowing users to complete the workflow smoothly without the need for external documentation.

In summary, the graphical user interface exhibits a high standard of human-computer interaction design, guided by Jakob Nielsen's ten usability heuristics. It effectively

addresses key aspects such as system feedback, terminology alignment, user control, consistency, and error handling. By combining a clean layout, informative prompts, and flexible interaction mechanisms, the interface supports both medical professionals and lay users, enhancing usability and operational efficiency.

While a comprehensive user manual is not yet provided, the system's strong self-explanatory design enables users to complete essential tasks—such as uploading images, loading models, and making predictions—without external guidance.

Overall, the GUI embodies a user-centered design philosophy and offers a solid foundation for the visualization and deployment of intelligent medical imaging systems, paving the way for future clinical usability testing and feature expansion.

6 Conclusion

6.1 Main Contributions

This study tackles the task of multi-label classification for chest X-ray images by establishing a unified evaluation framework on the ChestX-ray14 dataset. It systematically compares the performance of various mainstream architectures, including CNNs (e.g., VGG16, DenseNet121), Transformers (e.g., ViT), and hybrid models (e.g., ConvNeXt). Attention mechanisms were integrated into CNN backbones to assess their effects on accuracy, training efficiency, and parameter complexity.

The results shows that on the ChestX-ray14 dataset, the ConvNeXt model achieved an average AUC of 0.77, representing an improvement of approximately 4% over DenseNet121, while also reducing training time by about 40%. The Vision Transformer (ViT) models also performed competitively, reaching an average AUC of approximately 0.76. Notably, the MobileNetV3Large model with CBAM integration achieved comparable performance to the larger ViT and ConvNeXt models, attaining an average AUC of 0.76, despite having an exceptionally compact model size of only 12.4 MB. This highlights its excellent trade-off between predictive accuracy and computational efficiency. However, its primary limitation lies in its extended training duration—it required approximately 49.7% more training time compared to DenseNet121.

Additionally, a graphical user interface (GUI) was developed to support model selection, threshold adjustment, image upload, prediction, and visual explanation. The GUI improves usability and interpretability, showcasing its potential for real-world deployment in intelligent medical imaging systems.

6.2 Research Significance

This study focuses on the automatic multi-label classification of chest X-ray images, integrating a variety of deep learning architectures and attention mechanisms to systematically investigate model performance differences and applicability on real-world medical datasets. Under a unified evaluation pipeline, the study provides a comprehensive comparison in terms of classification accuracy, training convergence efficiency, and parameter complexity. Such systematic analysis not only highlights the strengths and limitations of different model architectures in medical image tasks but also offers clear guidance for future research in selecting and designing appropriate models.

Moreover, this work explores the feasibility of integrating attention mechanisms into CNN-based models for chest X-ray classification. Results show that these modules improve model accuracy and training stability with minimal computational overhead, especially in lightweight models like MobileNetV3, which achieved performance comparable to larger architectures. These findings support the effectiveness of attention modules in medical image classification and offer practical guidance for enhancing performance in resource-constrained settings.

Finally, the development of an interactive graphical user interface (GUI) in this study improves both the interpretability and user experience of the AI system. It serves as a practical example of how intelligent diagnostic tools can be visualized and deployed in real-world clinical scenarios, offering valuable reference for future translational research and clinical implementation.

6.3 Future Work

Despite the encouraging results achieved in this study on multi-label classification of chest X-ray images, several directions remain worthy of further investigation in future research.

First, with respect to model selection, the current work primarily evaluates classical CNNs, Transformer-based models, and hybrid architectures. To further improve performance and assess model adaptability, future research could incorporate next-generation lightweight or high-precision architectures such as EfficientNetV2, Swin Transformer, and DeiT3. A systematic comparison of these models on medical image classification tasks would help identify architectures with superior accuracy-to-parameter trade-offs.

Second, regarding model interpretability, while Grad-CAM++ and Self-Attention Attribution offer useful visual cues, their spatial localization accuracy remains insufficient for clinical use. To address this limitation, future studies could explore the integration of object detection models such as YOLOv8 or RetinaNet with classification networks. This approach would enable precise lesion localization and boundary estimation, thereby enhancing the clinical utility and reliability of interpretability outputs.

Finally, considering that the graphical user interface (GUI) developed in this study demonstrates strong interactivity and interpretability, future work could involve usability testing in real-world clinical settings. Incorporating feedback from healthcare professionals would guide iterative refinements to the interface design, workflow, and output presentation, ultimately improving the system's practicality and user trust. Furthermore, the generalizability and transferability of the proposed framework to other imaging modalities—such as CT and MRI—or multi-modal medical data could be investigated to expand the applicability and impact of this research.

REFERENCES

[1] A. Kulkarni, G. Parasnis, H. Balasubramanian, V. Jain, A. Chokshi, and R. Sonkusare, "Advancing diagnostic precision: Leveraging machine learning techniques for accurate detection of covid-19, pneumonia, and tuberculosis in chest x-ray images," 2023. [Online]. Available: https://arxiv.org/abs/2310.06080

[2] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019. [Online]. Available: https://arxiv.org/abs/1901.07031

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[4] I. Mwendo, K. Gikunda, and A. Maina, "Deep transfer learning for detecting covid-19, pneumonia and tuberculosis using cxr images – a review," 2023. [Online]. Available: https://arxiv.org/abs/2303.16754

[5] J. Gupta, S. Pathak, and G. Kumar, "Deep learning (cnn) and transfer learning: A review," *Journal of Physics: Conference Series*, vol. 2273, no. 1, p. 012029, may 2022. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/2273/1/012029

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015. [Online]. Available: https://arxiv.org/abs/1512.00567

[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993

[11] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," 2019. [Online]. Available: https://arxiv.org/abs/1905.02244

[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022. [Online]. Available: https://arxiv.org/abs/2201.03545

[13] P. Naveen and B. Diwan, "Pre-trained vgg-16 with cnn architecture to classify x-rays images into normal or pneumonia," in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021, pp. 102–105.

[14] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," 2020. [Online]. Available: https://arxiv.org/abs/2002.05990

[15] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.

[16] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest x-ray classification," 2019. [Online]. Available: https://arxiv.org/abs/1803.02315

[17] D. Bhusal and S. P. Panday, "Multi-label classification of thoracic diseases using dense convolutional network on chest radiographs," 2024. [Online]. Available: https://arxiv.org/abs/2202.03583

[18] M. S. A. Reshan, K. S. Gill, V. Anand, S. Gupta, H. Alshahrani, A. Sulaiman, and A. Shaikh, "Detection of pneumonia from chest x-ray images utilizing mobilenet model," *Healthcare*, vol. 11, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2227-9032/11/11/1561

[19] M. A. Sufian, W. Hamzi, T. Sharifi, S. Zaman, L. Alsadder, E. Lee, A. Hakim, and B. Hamzi, "Ai-driven thoracic x-ray diagnostics: Transformative transfer learning for clinical validation in pulmonary radiography," *Journal of Personalized Medicine*, vol. 14, no. 8, 2024. [Online]. Available: https://www.mdpi.com/2075-4426/14/8/856

[20] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1711.05225

[21] S. Taslimi, S. Taslimi, N. Fathi, M. Salehi, and M. H. Rohban, "Swinchex: Multi-label classification on chest x-ray images with transformers," 2022. [Online]. Available: https://arxiv.org/abs/2206.04246

[22] U. Marikkar, S. Atito, M. Awais, and A. Mahdi, "Lt-vit: A vision transformer for multi-label chest x-ray classification," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1109/ICIP49359.2023.10222175

[23] L. Huang, J. Ma, H. Yang, and Y. Wang, "Research and implementation of multi-disease diagnosis on chest x-ray based on vision transformer," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 3, 2024. [Online]. Available: https://qims.amegroups.org/article/view/122245

[24] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018. [Online]. Available: https://arxiv.org/abs/1804.03999

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018. [Online]. Available: https://arxiv.org/abs/1708.02002

[27] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, rOC Analysis in Pattern Recognition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X

[28] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[29] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar. 2018. [Online]. Available: http://dx.doi.org/10.1109/WACV.2018.00097

[30] Y. Hao, L. Dong, F. Wei, and K. Xu, "Self-attention attribution: Interpreting information interactions inside transformer," 2021. [Online]. Available: https://arxiv.org/abs/2004.11207

[31] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142(1950)3:1¡32::AID-CNCR2820030106¿3.0.CO;2-3

[32] J. Nielsen, *Usability engineering*. Morgan Kaufmann, 1994.