Comparative Analysis of Performance of CNNs and Vision Transformers on Multi-label and Multi-class Classification of Chest X-rays

Project report submitted by

AMAN NASIM 420107
BOBBILI SUNNY RISHY VARDHAN 420117
KALAKONDA PRUDHVI RAJ 420145

Under the supervision of

Dr. Srilatha Chebrolu

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH
TADEPALLIGUDEM-534101, INDIA

May 2024

Comparative Analysis of Performance of CNNs and Vision Transformers on Multi-label and Multi-class Classification of Chest X-rays

Project report submitted by

AMAN NASIM 420107
BOBBILI SUNNY RISHY VARDHAN 420117
KALAKONDA PRUDHVI RAJ 420145

Under the supervision of

Dr. Srilatha Chebrolu

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
NATIONAL INSTITUTE OF TECHNOLOGY ANDHRA PRADESH
TADEPALLIGUDEM-534101, INDIA

May 2024

PROJECT WORK APPROVAL

This project work entitled "Comparative Analysis of Performance of CNNs and Vision Transformers on Multi-label and Multi-class Classification of Chest X-rays" worked out by AMAN NASIM (420107), BOBBILI SUNNY RISHY VARDHAN (420117), and KALAKONDA PRUDHVI RAJ (420145) is approved for the degree of Bachelor of Technology in Computer Science and Engineering.

Examiners

_____

_____

_____

Supervisor(s)

_____

_____

_____

Chairman

_____

Date : _____

Place : _____

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

AMAN NASIM	BOBBILI SUNNY RISHY VARDHAN	K PRUDHVI RAJ
420107	420117	420145
Date:	Date:	Date:

National Institute of Technology, Andhra Pradesh

Department of Computer Science and Engineering

Certificate

This is to certify that the report entitled **"Comparative Analysis of Performance of CNNs and Vision Transformers on Multi-label and Multi-class Classification of Chest X-rays"** submitted by AMAN NASIM, Roll No: 420107, BOBBILI SUNNY RISHY VARDHAN, Roll No: 420117, and KALAKONDA PRUDHVI RAJ, Roll No: 420145 to National Institute of Technology, Andhra Pradesh is a record of bonafide research work carried out by them under my supervision and guidance. This work has not been submitted elsewhere for the award of any degree.

Dr. Srilatha Chebrolu

(Project Guide)

Place: Tadepalligudem

Date:

ACKNOWLEDGEMENT

We take this opportunity to express our profound gratitude and deep regards to our supervisor Dr. SRILATHA CHEBROLU, Assistant Professor, Department of Computer Science and Engineering, NIT Andhra Pradesh for her exemplary guidance, monitoring and constant encouragement throughout the project.

We avail ourselves of this proud privilege to express our regard to all the faculty of the department of Computer Science and Engineering at NIT Andhra Pradesh for empathizing and providing all the necessary facilities throughout the work.

I thank all my friends and batch mates for being a significant part of my student life and making me ready for the future. They will always stay with me in the form of the conversations we shared and memories we had. Finally, I thank my parents for being a constant support in my life and helping me through my highs and lows.

AMAN NASIM	BOBBILI SUNNY RISHY VARDHAN	K PRUDHVI RAJ
420107	420117	420145
Date:	Date:	Date:

May 8, 2024

Table of Contents

List of Figures

Chapter 1

Abstract

Compared to the traditional computational imaging algorithms, the utilization of deep learning methodologies, particularly CNNs and the Transformers, has significantly increased the efficacy of medical imaging and enhanced its classification, mainly of chest X-rays. This paper examines the prescribing power of CNNs and Vision Transformers for this task. We conducted an extensive analysis of the advanced deep learning networks using the VinDr-CXR and VinDr-PCXR datasets. Our experiments encompassed both multi-class and multi-label classification, with evaluation metrics spanning accuracy, F1-score, precision, recall, and AUC. Our study provides insights into the strengths and potential applications of these deep learning models in medical image analysis. These findings are crucial for researchers aiming to develop methods that improve diagnostic accuracy in interpreting chest X-rays.

Keywords : VinDr-PCXR, VinDr-CXR, deep learning, Multi-class classification, Multi-label classification, ResNet, EfficientNet, ConvNeXt, Swin Transformer, DaViT, CoAtNet

Chapter 2

Introduction

Medical imaging [2, 6, 53], which is made possible by rapid advances in technologies, plays a pivotal role in diagnosing diseases and is also used to monitor patient health. A radiologist performs detection, which is the first of his tasks of marking any abnormal local finding in the chest radiograph. Once the local abnormality is marked, the next task is to diagnose and give a comprehensive medical report. However, it was gathered that one of the main problematic issues in building automated diagnostic systems is the sparsity of big annotated datasets. This is where the notated data plays a significant role, especially in medical imaging cases.

To bring in the effect of Deep Learning (DL) approaches into medical image analysis, Convolutional Neural Networks (CNNs) [41, 5, 28] have achieved very good results in this field with automation capability. The deficit of carefully enriched datasets, which consist of classes of diseases and their anatomical position, creates a major hindrance in training advanced pathological models. CNNs, being powerful tools are essential for medical image analysis. They can not only be good at capturing small details in images but also have a strong disentanglement ability. As a result, traditional CNNs may encounter challenges in capturing the complex interplay and distant relationships inherent in medical images, particularly in contexts such as the interpretation of chest X-rays (CXRs) by medical professionals.

The breakthroughs in the deep learning field, specifically Vision Transformers (ViTs) [13, 24, 27], is a candidate that shows a bright perspective. Attention mechanisms, viewed as a distinctive feature of ViTs, often prove to be quite successful in extracting the whole content of data, which is also aimed at long-range relationship discovery. The fact is that the transformation functions dis-

play no equivariance of the translation property present in CNN and these lack better inductive bias compared to CNN. Nevertheless, Transformers with the most versatile and adaptable architecture enable the modeling of many different factors and connections to have an edge in various DL tasks.

In our study, we examine chest radiographs, which can be classified as a domain where several comorbidities might manifest as a result of different diseases. The utilized datasets, VinDr-PCXR [42, 17] and VinDr-CXR [40, 39, 17] contain diverse kinds of diseases whose individual imaging and diagnostic properties may be distinctive. The main groups of diseases are connected to pneumonia, bronchitis, and other lung and heart problems that would typically show as subtle marginal cases under a microscope. Further, we aim to conduct a systematic comparison between CNNs and Vision Transformers on the task of chest X-ray (CXR) classification. By assessing their performance across various disease categories, we aim to identify the strengths and limitations inherent in each approach. Through this comparative examination, we aim to offer insights into the appropriateness of CNNs and Vision Transformers for automated diagnosis in CXRs, addressing the critical need for precise and efficient disease detection and categorization in patients.

Chapter 3

Literature Review

Kim et al.(2022) [29] proposed a technique using Transfer Learning with EfficientNet v2-M to classify lung diseases on chest X-ray (CXR) images, obtaining validation scores of 82.15% for normal, pneumonia and pneumothorax classes on the NIH dataset [51] and 82.20%, respectively on the SCH dataset [20]. These authors' study unveils the critical role that data augmentation and transfer learning play for the precise multi-class lung disease diagnosis on the basis of the CXR images. Ibrahim et al. [23] proposed an approach based on deep learning that depends on having a pre-trained AlexNet [31] model for classifying the possible categories namely, COVID-19 [36], non-COVID-19 viral pneumonia, bacterial pneumonia and normal CXR scans. The fact that the models can be highly precise, sensitive, and precise in both binary and multi-class classifications highlights that even computer-aided diagnoses are helpful in diagnosing COVID-19.

A neural network design tailored for multi-class classification of pneumonia, lung cancer, TB, lung opacity, and COVID-19 available from CXRs is done by Goram et al. [4]. In their study, the VGG19 + CNN model of their proposed architecture showed exceptional performance with 99.82% (Area Under Curve) AUC. Furthermore, they recommend for some future studies that in addition to CT scan images, these images should include region of interest (ROI) identification for better disease severity classification. Emtiaz et al. [22] introduced the CoroDet CNN model designed to identify COVID-19 from CXRs and CT scan images. High performance manifested through results showing an accuracy of 99.1%, 94.2%, and 91.2% for the respective classes. CoroDet surpasses the existing methods by offering possible solutions in an event of testing kit shortage.

Asmaa et al. [1] presented DeTraC, a deep CNN model specifically designed for classifying

CXRs with COVID-19 and have an accuracy of 93.1% and sensitivity of 100%. One of the principal strengths of DeTraC is that it can work with real-world dataset irregularities as it involves a class decomposition technique. It can also demonstrate strong performance on real-world data. Allaouzi et al. [3] presented a unique approach merging CNN models and transformation methods to classify thorax diseases on CXR images, e.g. multiple labels. An approach leveraging a pre-trained DenseNet-121 model alongside various transformation methods yielding model performance better than the current state-of- the-art on ChestX-ray14 dataset [16]. This work has proven that recent advances in CNN and multi-label classifier could be applied effectively to Computer-Aided Diagnosis (CAD) systems and outperform thorax disease diagnosis.

Jin et al. [25] demonstrates the breakthrough in the classification of CXR images into multiple labels among current research. They introduced a new approach that combines visual and semantic vectors by adapting the ConvNeXt [35] and applying BioBert [32] encoding. By projecting features into a shared metric space and integrating a dual-weighted metric loss function, it manages to obtain superior capability with an average AUC value of 0.826 that exceeds the current state-of-art approaches. Umar Marikkar et al. [37] proposed LT-ViT, which is a transformer-based network architecture for multi-label classification of CXRs, that outperforms current methods by not using any decoder layers. LT-ViT can be trained from scratch on several available datasets of CXRs and it demonstrates a high level of generalizability across pre-training methods. Besides, LT-ViT offers model interpretability without any dependence on Grad-CAM [44] and its types.

Sina Taslimi et al. [48] proposed a multi-label classification model where the Swin-Transformer backbone is presented for the diagnosis of diseases from CXR images. They conducted significant experiments on the Chest X-ray14 dataset [16], where a three layer-headed model was used which outperformed previous methods by achieving the state-of-the-art performance, getting an average AUC score of 0.810. The report also provides the foundation for a fair benchmarking and confirms the model focusing on the diseases-related areas of the lungs. Eff-CTM [33], a hybrid medical image classification model integrating CNN and Transformer achieves a superior balance between efficiency and performance. Its multi-branch CNN module efficiently extracts local features, while the Transformer components capture global features effectively. Extensive experiments on various medical image datasets demonstrate Eff-CTM's superiority over CNN and Transformer methods in terms of efficiency and performance.

The CTransCNN [52], a powerful merger of CNN and transformer architectures, exhibits extreme efficiency in the categorization of multilabel medical images. CTransCNN which consist of modules, target the pairwise label correlation and model parameter improving to beat their counterparts on the ChestX-ray11, NIH ChestX-ray14 [16] and the TCMTD dataset [8]. Ahmed F. Mohamed et al. [38] proposed a novel classifier of breast cancer supported by the Chaos Game Optimization (CGO), the Nutcracker optimizer (NO), and the Cross Vision Transformer as the feature extractor. Thus the model optimization through CGO explore NO which gives an enhanced tumor detection and identification of best solution. The results show a promising perspective to be utilized in the diagnostics field, with potential for upgrading by means of implementing opposite learning and evolutionary methods.

The PEFMed [10] method presented by Zhiyong Dai et al. employs a novel model with dual encoder and prior-driven Variational Autoencoder (VAE) module to perform few-shot classification of medical images. By capturing deeper semantic information via massive experiments on a variety of medical image datasets, PFEMed proves to produce superior performance over the baseline state-of-the-art few shot framework known as MetaMed, with better performance by up to 2.63% on the Pap smear dataset [26]. Hengde Zhu et al. [54] had worked on providing a synergetic evolution process of DenseNet towards higher frequency and precision in classifying of medical images, as well as an evolution-driven ensemble learning method. The MEEDNets model have several DenseNet-121s that allowed to surpass publicly available approaches on both SARS-CoV-2 CT scan dataset [46] and the brain tumour dataset [9] . These running schemes give hope that they can be used in upgrading performance and efficiency of medical image systems.

Chapter 4

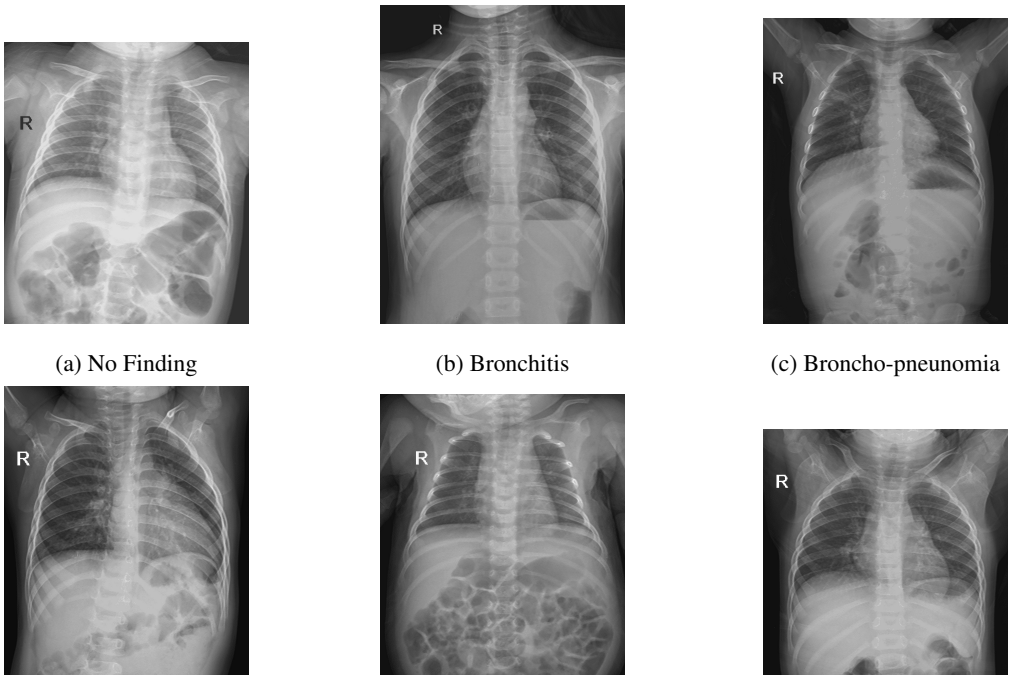
Dataset

4.1 VinDr-PCXR

VinDr-PCXR [42, 17] dataset contains pediatric CXR images. It also includes labels for images containing the disease associated with each image. Every scan underwent manual annotation by a pediatric radiologist. Additionally, each image was labeled for the presence of 36 specific findings or local abnormalities, along with 15 diseases, which are also known as global findings. Figure 4.1 displays sample images extracted from this dataset, each accompanied by its corresponding disease labels. Figure 4.2 shows the count of CXRs for VinDr-PCXR having the presence of the disease labels highlighting a significant issue of class imbalance. 'No finding' label is present in around 6000 images, which makes up a significant portion of the dataset.

4.2 VinDr-CXR

The VinDr-CXR [40, 39, 17] is sourced from the Kaggle competition titled 'VinBigData Chest X-ray Abnormalities Detection' [14], aimed at automating the localization and classification of CXR abnormalities. It comprises of 15,000 scans annotated with bounding boxes, facilitating the automatic localization and classification of 14 types of thoracic abnormalities. These abnormalities cover a spectrum of pathological conditions commonly encountered in chest radiographs, including pneumothorax, pulmonary fibrosis, and cardiomegaly among others. Each scan is annotated



(a) No Finding (b) Bronchitis (c) Broncho-pneunomia

(d) Other disease (e) Bronchiolitis (f) Pneumonia

Figure 4.1: Sample images from VinDr-PCXR dataset with their corresponding disease labels

with multiple labels corresponding to the presence of different abnormalities, enabling multi-label classification task. Figure 4.3 depicts several images sourced from this dataset, with each image annotated to highlight its local abnormalities. Figure 4.4 shows the count of the various thoracic abnormalities in the VinDr-CXR dataset, again showing a high class imbalance among the abnormalities, with the 'No finding' label present in around 10000 images.

4.3 Dataset Preprocessing

In the preprocessing phase, we applied different sort of techniques to have better and consistent quality of CXRs. The following steps were undertaken.

4.3.1 VOI LUT

The preprocessing stage consisted of applying the Value of Interest (VOI), Look-up table (LUT) to relevel and rescale the images. It implements this method by choosing the most appropriate contrast

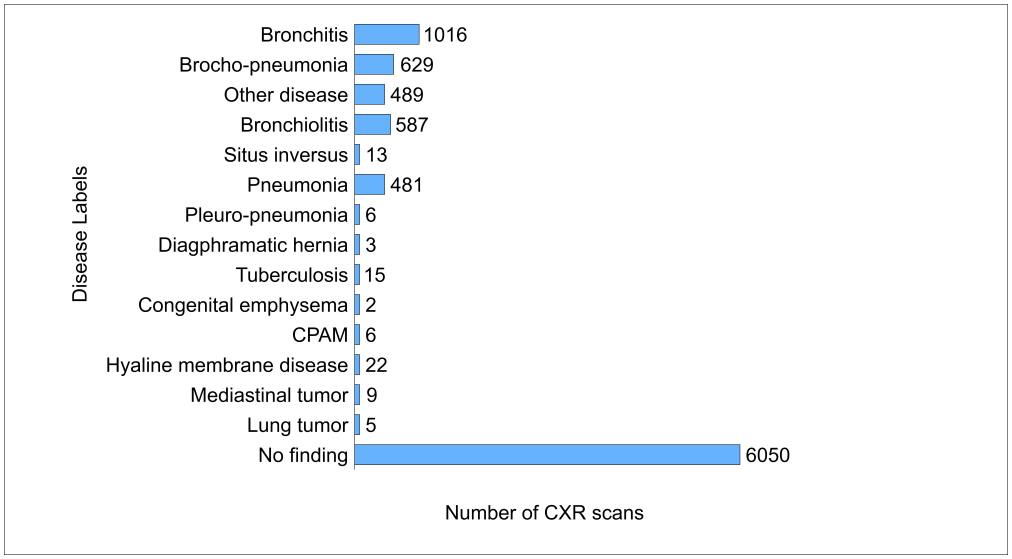


Figure 4.2: Distribution of CXRs across each disease label in the VinDr-PCXR dataset.

and brightness for the image. This is done by selecting one out of the various transformation functions that depend on the header information carried by the image. This enables efficient visual presentation of the anatomic structures as well as the abnormalities on the CXR.

4.3.2 Min-Max Normalization

The pixel intensity values of the original images were normalized using the Min-Max method to scale the values between 0 and 1. This normalization technique helps to keep the data uniform, enabling proper model training without changing the relative comparisons and the position between the CXRs.

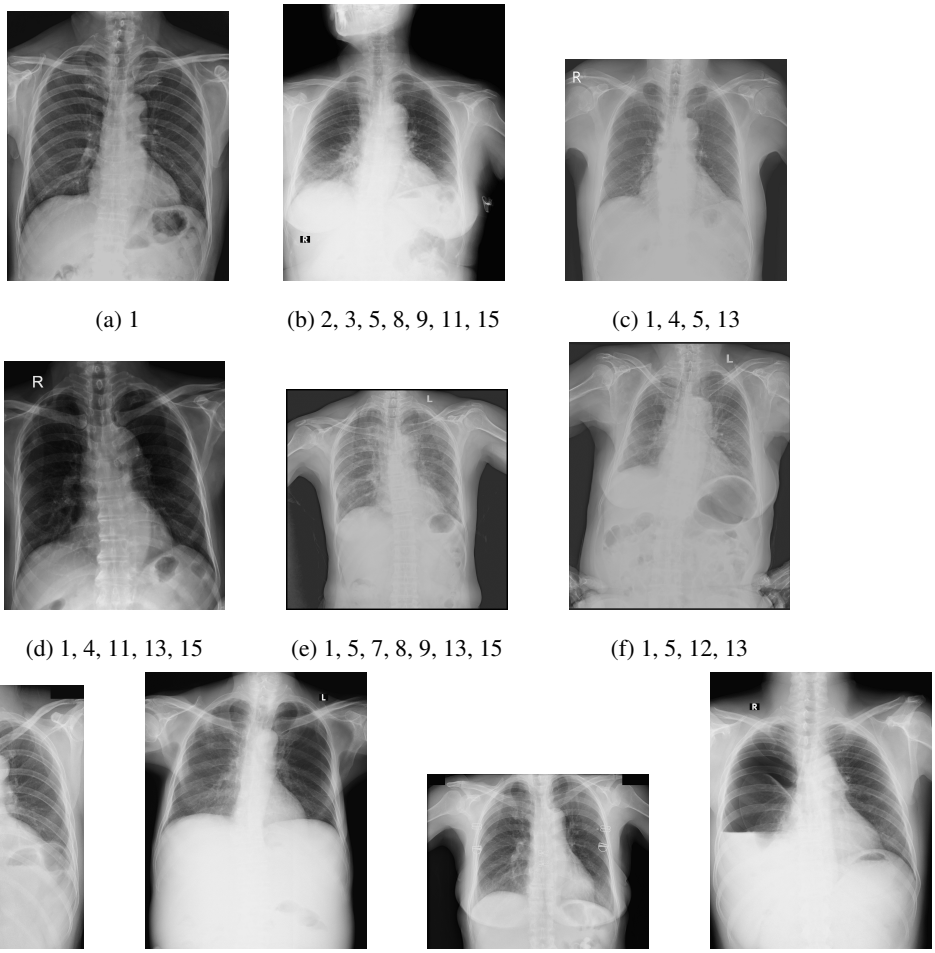
4.3.3 CLAHE

Contrast Limited Adaptive Histogram Equalization (CLAHE) [43] was used with parameters tuned to amplify the local contrast and provide more details for features of interest within the images. The clip limit was set to 2 and tile size to 8x8 pixels to prevent the noise from being amplified too much while also making the image histogram look even.

After performing these tasks, the CXRs were resized to 512x512 pixels. While resizing, we used the area interpolation that solves the problem of oversampling image data and hence preserve

the important spatial details while retaining the image size. Now, the resized version of the images were converted and saved for better and more efficient retrieval of images during model training.

Figure 4.5 shows a sample CXR from the VinDr-CXR dataset before and after applying the above mentioned preprocessing steps. Visually, it clearly demonstrates the effectiveness of the employed techniques to bring out the contrast from the CXRs.



(a) 1 (b) 2, 3, 5, 8, 9, 11, 15 (c) 1, 4, 5, 13

(d) 1, 4, 11, 13, 15 (e) 1, 5, 7, 8, 9, 13, 15 (f) 1, 5, 12, 13

(g) 1, 6, 9, 12 (h) 1, 8, 9, 13, 15 (i) 1, 10 (j) 12, 14

Figure 4.3: Sample images from VinDr-CXR dataset where each image is labelled with its abnormalities. Label mapping - 1 : No finding, 2 : Aortic enlargement, 3 : Atelectasis, 4 : Calcification, 5 : Cardiomegaly, 6 : Consolidation, 7 : ILD, 8 : Infiltration, 9 : Lung Opacity, 10 : Nodule/Mass, 11 : Other lesion, 12 : Pleural effusion, 13 : Pleural thickening, 14 : Pneumothorax, 15 : Pulmonary fibrosis

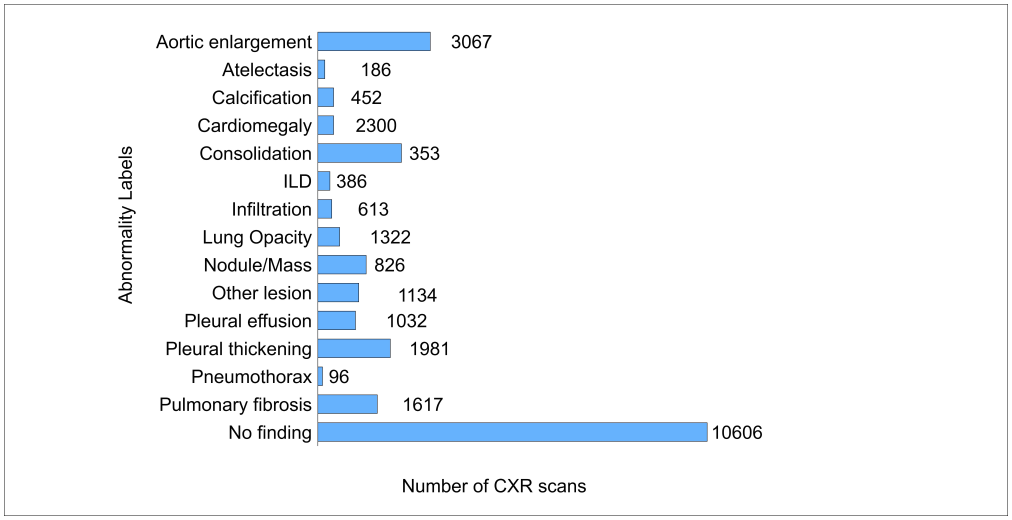


Figure 4.4: Distribution of CXRs across each abnormality label in the VinDr-CXR dataset.

(a) Original image (b) After VOI LUT (c) After Min-Max Norm and CLAHE

Figure 4.5: Preprocessing stages for a sample image from VinDr-CXR

Chapter 5

Overview of Existing DL Models

In this section, we provide an overview of the deep learning architectures utilized for the task of CXR classification. CNNs have been the cornerstone in image classification tasks for their ability to capture spatial hierarchies of features. More recently, ViTs have emerged as a novel architecture, demonstrating competitive performance in various visual recognition tasks by leveraging self-attention mechanisms.

5.1 ResNet

Kaiming et al. [19] introduced a deep learning paradigm named resdiual network that eased the procedure of learning the networks that are much deeper from earlier ones. Not only did this sped up training time but also improved the classification performance. A residual building block can be represented by Equation 5.1.

$$y = F(x, \{W_i\}) + x \tag{5.1}$$

where x and y are the input and output vectors, whereas the function F represents residual mapping, W_i represents the weights to be learnt for the i^{th} layer and $+$ represents element-wise addition operation. The function F can represent either Linear or Convolutional layers. The ResNet architecture has been implemented in various depths, including variants with depths of 18, 34, 50, 101 and 152 layers. The 152-layer deep ResNet that outperforms its other variants on ImageNet

dataset still had lower complexity in terms of number of parameters than VGG-16 and VGG-19 [45].

5.2 EfficientNet

Mingxing et al. [47] proposed a compound scaling method that uniformly scales width, resolution and depth of the ConvNet. At first they built a good baseline ConvNet by solving the optimization objectve as specified by Equation 5.2.

$$\max_{\text{model}} \left[\text{Accuracy}(\text{model}) \times \left(\frac{\text{FLOPS}(\text{model})}{T} \right)^{w} \right] \tag{5.2}$$

where Accuracy(model) and FLOPS(model) denote the accuracy and Floating point operations of the model respectively, T is the target FLOPS and w is a hyperparameter that controls the trade-off between accuracy and FLOPS. This produces a good baseline model which was named as EfficientNet-B0. Subsequently, this model was scaled up using the compound scaling method that maximized the model's accuracy given the resource constraints. This can be specified by Equation 5.3.

$$\begin{aligned} &depth = \alpha^{\phi} \\ &width = \beta^{\phi} \\ &resolution = \gamma^{\phi} \end{aligned} \qquad \begin{aligned} &\text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \\ &\alpha \geq 1, \quad \beta \geq 1, \quad \gamma \geq 1 \end{aligned} \tag{5.3}$$

where α, β, γ are constants and ϕ is a user-specified coefficient that controls how many more resources are available for model scaling. This led to the develpment of many variants of EfficientNet from B1 to B7 that used different values of the compound scaling coefficient ϕ, with increasing number of parameters and improved performance.

5.3 ConvNeXt

Zhuang et al. [35] in their work, aimed for the construction of a pure ConvNet that severely outperforms the ViTs in terms of accuracy and scalability and is much simpler in design. They begin

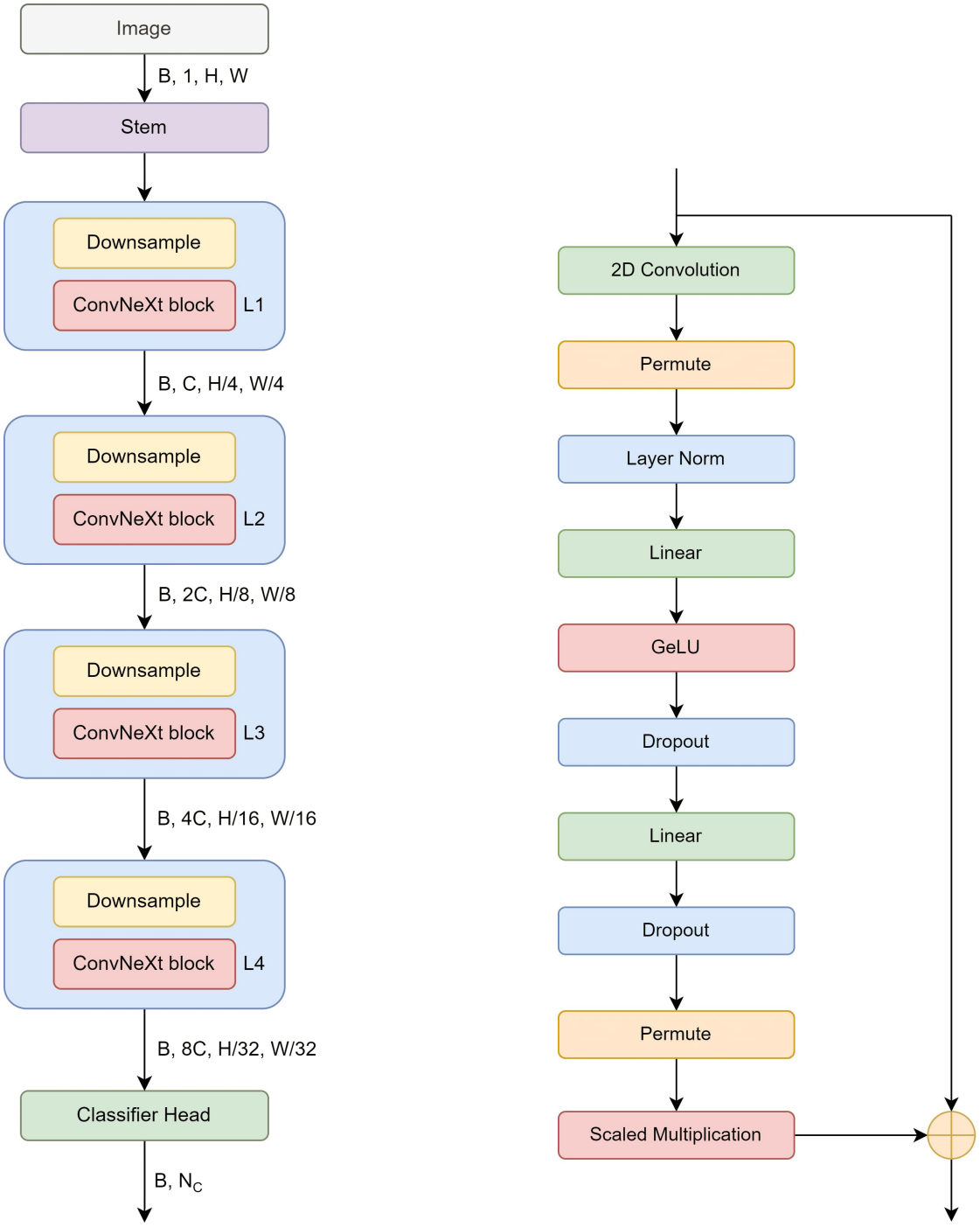
with the base model of ResNet-50 and successively apply appropriate training strategies in phases. There were several design choices adopted that were inspired by ViTs that helped develop a ConvNet collectively referred to as ConvNeXt. This outperformed the even the best of the Hierarchical Transformers like Swin Transformers [34] over the ImageNet-1K classification.

Figure 5.1a shows the overall architecture of ConvNeXt. It starts with the image being passed to the Stem layer which applies 2D convolution operation and prepares the image to be passed on to the 4 successive stages. Each stage comprises of an initial Downsample block that reduces the spatial dimensions of the image. Following that the image is passed on to L_i number of ConvNeXt blocks in the i^{th} stage sequentially. For this specific architecture, the values of L1, L2, L3 and L4 are 3, 3, 9, 3 respectively. Figure 5.1b shows an abstraction of the ConvNeXt block having a combination of Convolution, Layer Normalization, Linear and Dropout layers with GeLU activation function with a residual connection towards the end. After all the 4 stages, the rich image representation so obtained is passed to a Classifier Head that predicts individual class probabilities.

5.4 Swin Transformer

The Swin Transformer [34] is a notable advancement in the field of ViTs that can be used as a backbone in various vision tasks. It is a Hierarchical Transformer and is favoured over other Transformer variants due its unique Shifted Window Attention Mechanism which allows it to capture long range dependencies within the image in linear time. Figure 5.2 shows the architecture of the Swin Transformer. Like the original ViT, it first splits an image into patches of size 4x4 using the initial Patch Embedding layer which is then projected to to a given dimension C. The attention mechanisms are then applied taking each of those patches as a single unit. The value of C is 192 and the number of blocks in each stage viz., L1, L2, L3 and L4 are 2, 2, 18 and 2 respectively, for the Swin-L variant which we have utilized.

Each stage helps in producing a hierarchical representation of the image by decreasing the spatial dimensions in each layer by 4. It also involves the passage of the intermediate features from atleast one pair of Swin Transformer (ST) blocks as shown in Figure 5.3, followed by the alternating Window Attention and Shifted Window Attention operations. Equations 5.4, 5.5, 5.6 and 5.7



(a) ConvNeXt architecture

(b) ConvNeXt block

Figure 5.1: ConvNext architecture and ConvNext block subcomponents

represents two successive ST blocks, outlining this straightforward yet distinctive process.

$$\hat{z}_l = \text{W-MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{5.4}$$

$$z_l = \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l \tag{5.5}$$

$$\hat{z}_{l+1} = \text{SW-MSA}(\text{LN}(z_l)) + z_l \tag{5.6}$$

$$z_{l+1} = \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1} \tag{5.7}$$

where z_{l-1} and z_l are the input and output to the l^{th} ST block and $\hat{z}_l$ is the intermediate output obtained after element wise addition of z_{l-1} and after applying Window Attention on it for the l^{th} layer. W-MSA, SW-MSA , MLP and LN denote Window Mulitihead Self Attention, Shifted Window Multihead Self Attention, Multilayer Perceptron layer and Layer Normalization layer respectively. Figure 5.4 shows the Attention block for Swin Transformer. During Shifted Window Attention the window boundaries are shifted and then Window Attention is applied locally within the new windows. By doing so, Shifted Window Attention propagates the information captured by the local windows to the rest of the image. In Window Attention, as depicted by Figure 5.5, self attention is applied locally in each of the image windows having multiple patches each. This enables the attention to be limited in the local neighborhood. Hence for a window we have a common Key matrix for different Queries inside the window. The special window shifting mechanism of the Swin Transformer allows it to capture both local and global dependencies efficiently.

5.5 DaViT

DaViT [12], is a variant of the Vision Transformer which is able to capture global context information while maintaining computational efficiency. Figure 5.6 shows the overall setup of DaViT. The entire model consists of 4 stages, with each stage having a Downsample embedding layer accompanied by the Dual-Attention (DA) block as shown in Figure 5.7. It alternatively applies Spatial

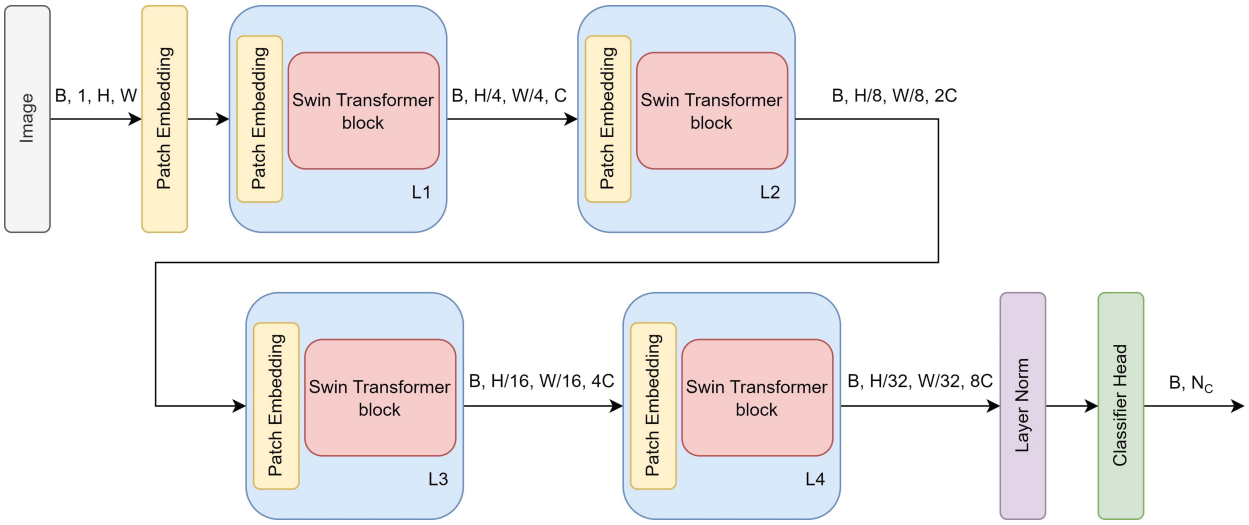


Figure 5.2: Swin Transformer Architecture

Window Multihead attention and Channel Group Attention to the image features. The Spatial Window attention simply applies attention locally within various windows in the image. This plays a similar role to the convolution operation by capturing spatial relationships inside the windows.

Figures 5.8 and 5.9 illustrate the two kinds of attention applied on an image feature of shape P x C, where P is the image's spatial dimension's abstraction. For the DaViT-Base variant the value of C is 128 and number of blocks L1, L2, L3 and L4 are 1, 1, 9, 1 for each successive stage respectively. The formulation for Spatial Window Attention is given by the Equation 5.8.

$$A_{\text{window}}(Q,K,V) = \{A(Q_i,K_i,V_i)\}_{i=0}^{N_w} \tag{5.8}$$

where N_w is the total number of windows, A defines the Self Attention operation and Q_i, K_i and V_i are the Query, Key and Value matrices in the i^{th} window respectively. These matrices are passed on to N_h number of self attention heads for all the N_w windows.

The Channel Group Attention applies attention on the transpose of the patch level image units, with number of heads N_h = 1. This fuses multiple global views of the image dynamically and in linear spatial time. The formulation for Channel Group Attention is given by Equation 5.9.

$$\begin{aligned} A_{\text{group}}(Q_i,K_i,V_i) &= \text{softmax}(\frac{Q_i^T K_i}{\sqrt{C_g}})V_i^T \\ A_{\text{channel}}(Q,K,V) &= \{A_{\text{group}}(Q_i,K_i,V_i)^T\}_{i=0}^{N_g} \end{aligned} \tag{5.9}$$

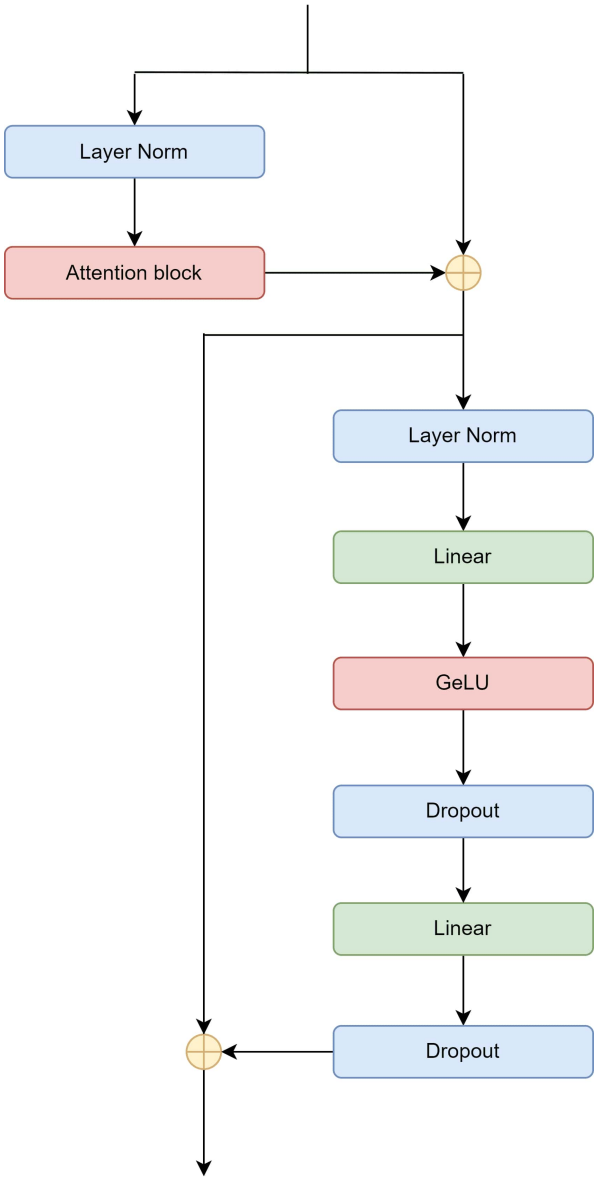


Figure 5.3: Swin Transformer block

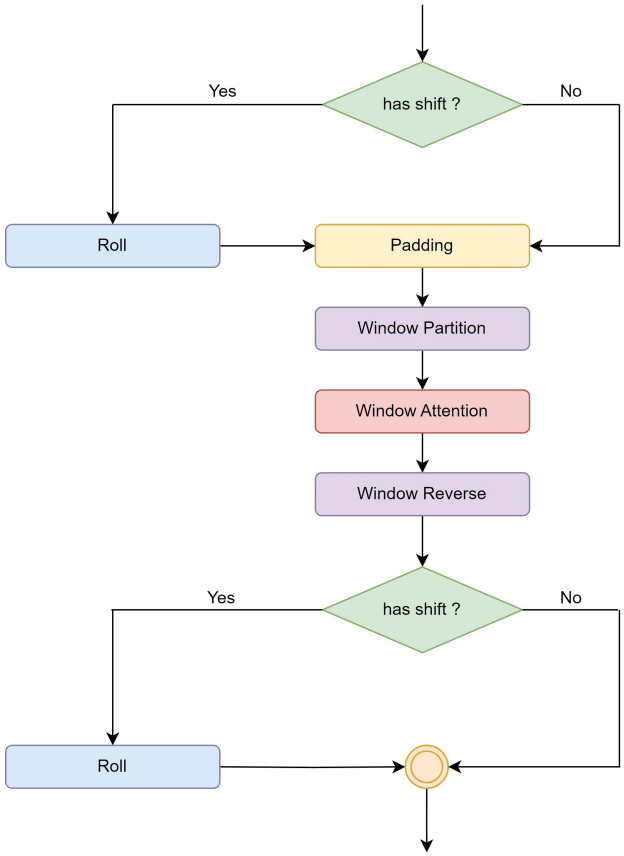


Figure 5.4: Swin Transformer Attention block

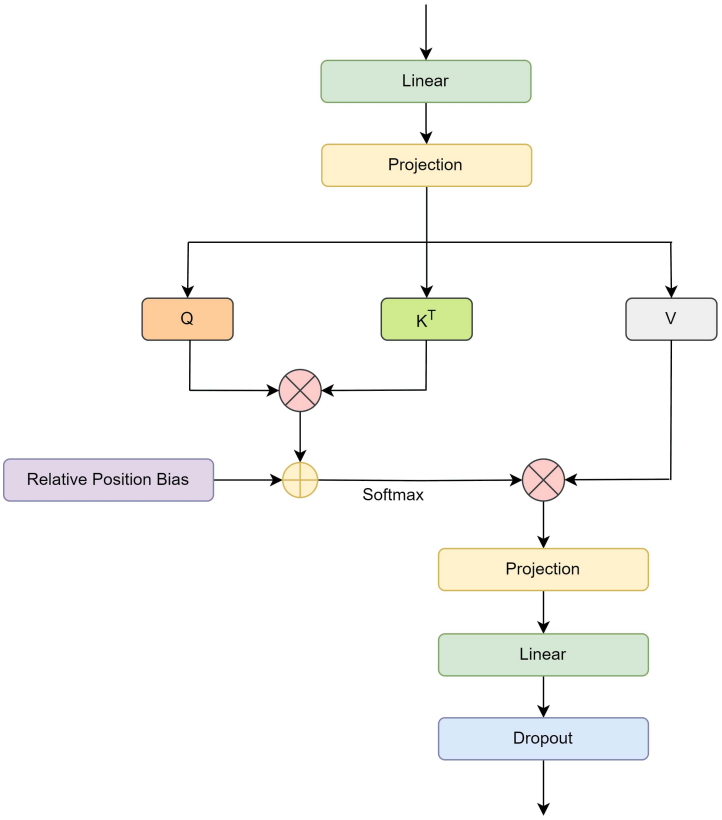


Figure 5.5: Swin Transformer Window Attention

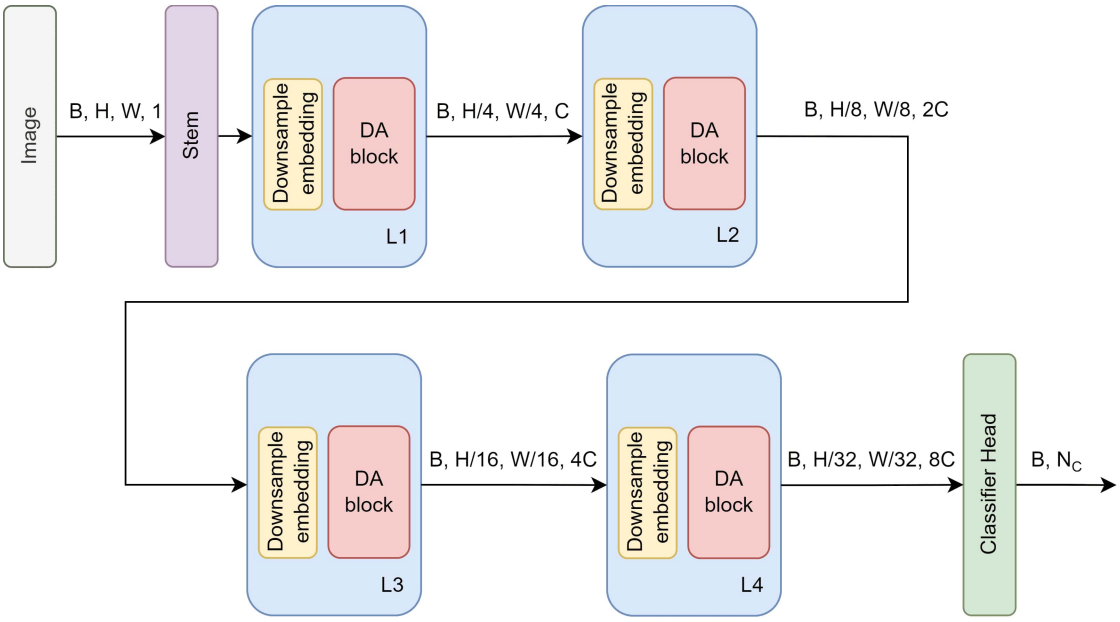


Figure 5.6: DaViT Architecture

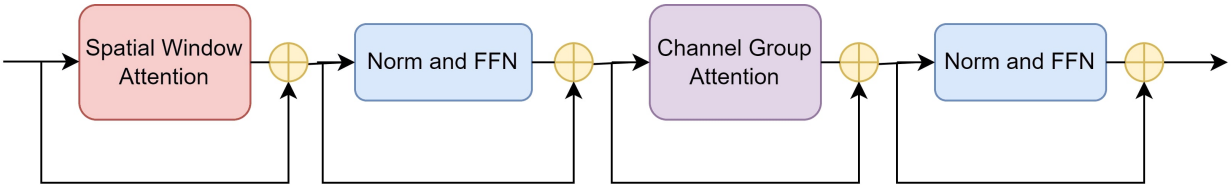


Figure 5.7: DaViT DA block

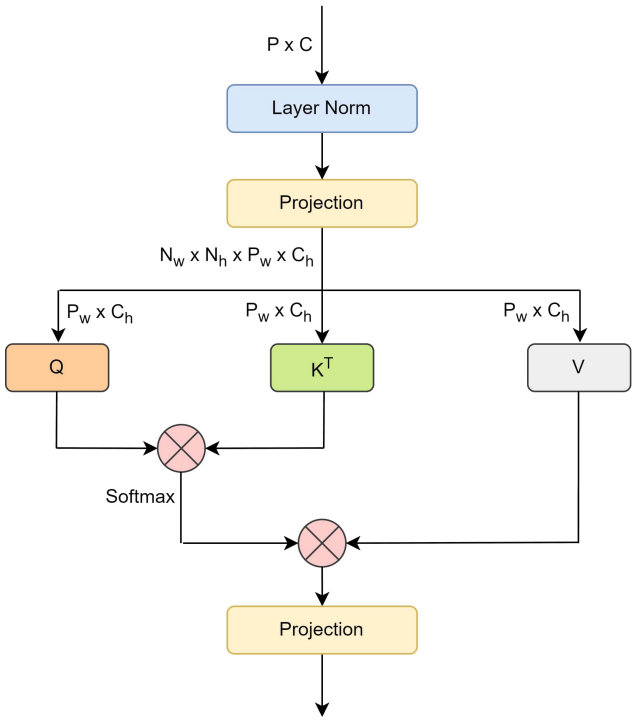


Figure 5.8: DaViT Spatial Window Attention

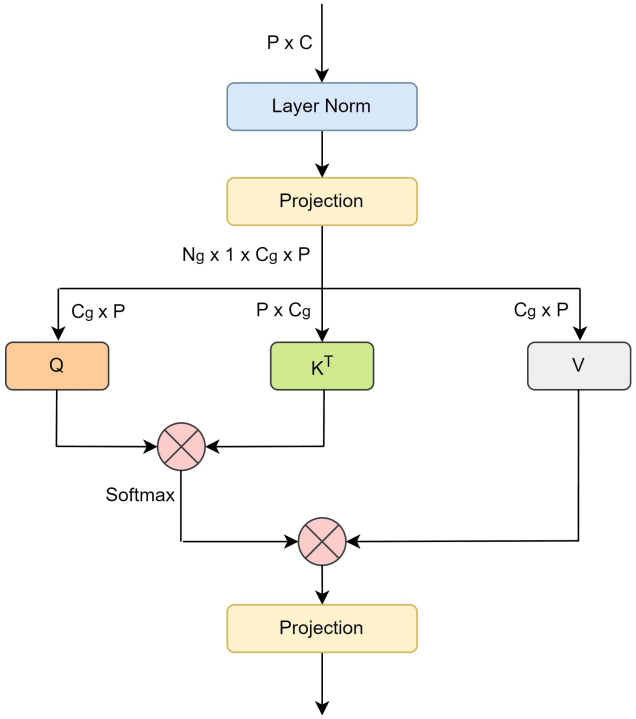


Figure 5.9: DaViT Channel Group Attention

where N_g is the total number of channel groups, C_g is the number of channels per group and Q_i, K_i and V_i are the Query, Key and Value matrices in the i^{th} group of channels respectively. These channel level matrices form an abstraction of the entire image interacting globally with a group of other channels, effectively capturing relationships across various image features.

5.6 CoAtNet

CoAtNet [11] introduces a very effective way to combine convolution and attention inheriting the best properties of both. The paper suggests that Transformers have a large model capacity than ConvNets but low generalizability due to lack of right inductive bias. CoAtNet combines depth-wise convolution and self attention using relative attention. It vertically stacks the convolution and attention layers in principled manner with attention layers always following the convolution layers. The pre-normalized relative attention used in CoAtNet's Transformer block is given by the Equation 5.10.

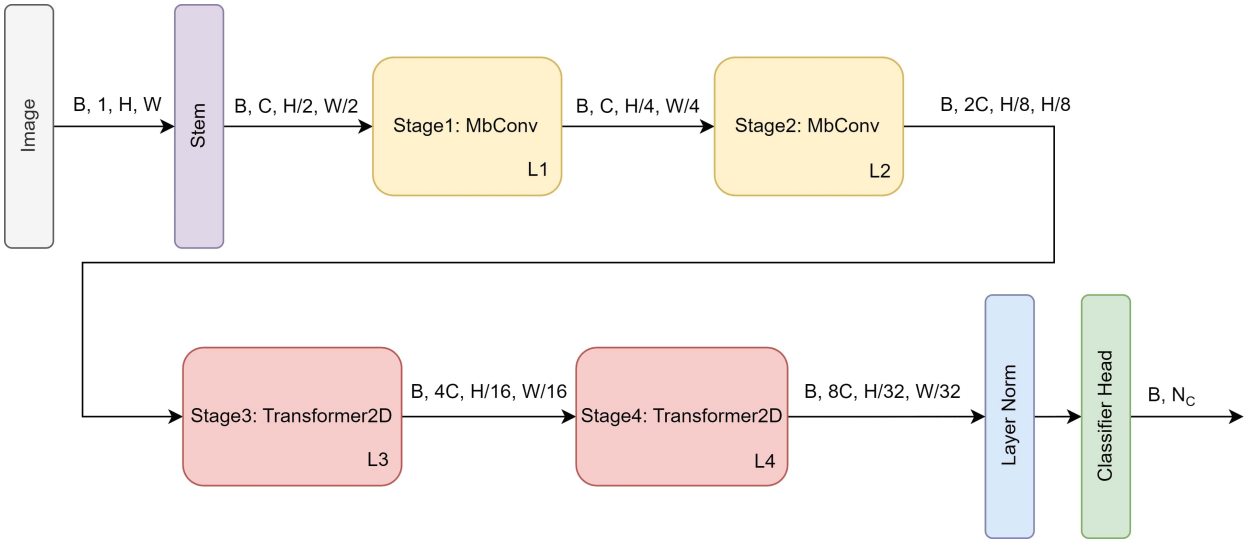


Figure 5.10: CoAtNet architecture

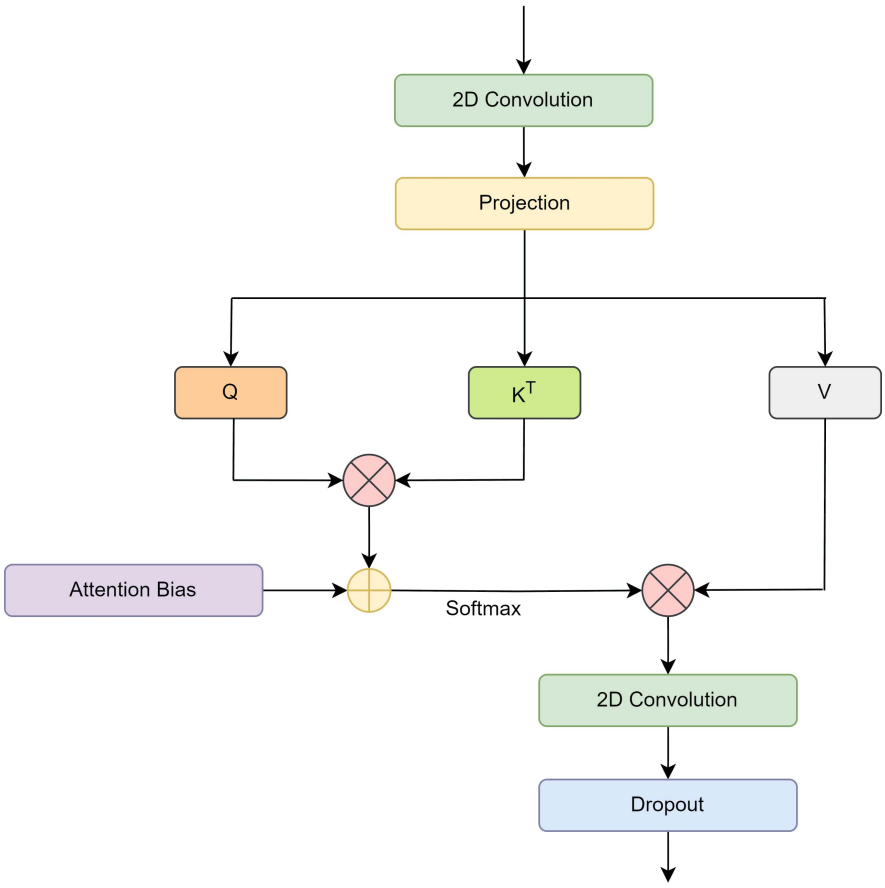


Figure 5.11: CoAtNet's Attention module

$$y_i^{\text{pre}} = \sum_{j \in G} \frac{\exp(x_i^T x_j + w_{i-j})}{\sum_{k \in G} \exp(x_i^T x_k + w_{i-k})} x_j \tag{5.10}$$

where y_i^{pre} is the i^{th} pixel value got after applying the relative attention mechanism on pixel x_i, G denotes the global spatial space, exp is the exponentiation operation and w_{i-j} represents the weights of the convolution kernel that are treated as a scalar to reduce the computational complexity. Figure 5.11 shows the Attention module of CoAtNet, where convolution operation is performed before and after the scaled dot product attention. Doing this allows it to infuse the spatial relationships captured by convolution inside the attention mechanism.

The variant of CoAtNet that we have utilized is CoAtNet-3, which is of 4 stages as specified by Figure 5.10, viz., *C-C-T-T* where *C* is a Convolution stage having MbConv blocks and *T* is a Transformer stage consisting of Transformer2D blocks. The number of blocks per stage L1, L2, L3 and L4 are 2, 6, 14 and 2 respectively. Prior to the stages the image passes through the Stem block, where it undergoes convolution twice. The embed dimensions or the number of channels produced by each stage are 192, 384, 768 and 1536 respectively, i.e., C = 192. The MbConv block performs sequenced convolution operations with residual connections, whereas the Transformer2D block applies the relative attention mechanism. This hybrid architecture enhances the performance not only with large sized datasets but also with small datasets.

Chapter 6

Experimentation

All the experiments were conducted in the Kaggle environment using its GPU T4 x2 accelerator. The experimenation phase was aimed to achieve a good performance for the classifcation tasks on VinDr-PCXR and VinDr-CXR datasets by making use of the already existing pre-trained models. The experiments were conducted on two versions of the datasets. The 'unbalanced' one indicates the presence of the label 'No finding' which has overwhelmingly large number of image samples in both the datasets. The 'balanced' version contains the dataset without the samples having 'No finding' label. This approach was implemented because the 'No finding' label is exclusive, indicating the absence of any other abnormality or disease in the dataset. In addition to that, for the VinDr-PCXR dataset we removed the disease labels having a count of less than 50. This resulted in the following disease labels 'Bronchitis', 'Brocho-pneumonia', 'Other disease', 'Bronchiolitis', 'Pneumonia' and 'No finding'. Table 6.1 shows the disease distribution for the resulting dataset after following this procedure. However the VinDr-CXR dataset had sufficient number of images per class, so the entire 15 abnormality labels were utilized according to the balanced and unbalanced versions of the dataset as shown in Table 6.2. Furthermore, VinDr-PCXR dataset was having images that had a very sparse number of images having multiple diseases. This made it a suitable candidate for both Multi-class and Multi-label classifcation, for which we carried out the experiments.

Disease Labels	Multi-label		Multi-class	
	Unbalanced	Balanced	Unbalanced	Balanced
Bronchitis	1016	1016	933	933
Brocho-pneumonia	629	629	493	493
Other disease	489	489	415	415
Bronchiolitis	587	587	511	511
Pneumonia	481	481	410	410
No finding	6048	0	6048	0

Table 6.1: Disease Frequency for VinDr-PCXR Milti-label and Multi-class Classification tasks

Disease	Unbalanced Counts	Balanced Counts
Aortic enlargement	3067	3067
Atelectasis	186	186
Calcification	452	452
Cardiomegaly	2300	2300
Consolidation	353	353
ILD	386	386
Infiltration	613	613
Lung Opacity	1322	1322
Nodule/Mass	826	826
Other lesion	1134	1134
Pleural effusion	1032	1032
Pleural thickening	1981	1981
Pneumothorax	96	96
Pulmonary fibrosis	1617	1617
No finding	10606	0

Table 6.2: Disease Frequency for VinDr-CXR Milti-label Classification

6.1 Data Transformations

We leveraged a custom-built data loading mechanism to accommodate our task-specific requirements. It loads the preprocessed CXRs, followed by a series of transformations to prepare the data for model training. These transformations include the following.

- Ensuring that the first diemnsion is holding the image channel. This ensures that the channel dimension is placed as the first dimension of the image.
- Image resizing to convert the spatial dimensions of the images to a uniform size.
- Scaling intensity values of the image pixels to a normalized range, typically between 0 and 1.
- Application of random zoom to augment the data during training. This helps introduce variability in the training data, reducing overfitting and enhancing the generalization capability of the model. This transformation is not applied during testing and validation to ensure consistency in evaluation.

6.2 Hyperparameters

We set the batch size to 8 and fixed the target size of the image resizing transformation to 224 x 224 for our experiments. Additionally we set the number of parallel processes used for data loading to 4. Increasing this number can accelerate data loading by leveraging multi-core processing capabilities but is constrained by the available computational resources.

For the training of pretrained models, we adopted a transfer learning approach. Initially, we loaded the pretrained weights and trained only the last classifier layers for a variable number of epochs, typically ranging from 5 to 30 epochs, for each model. This allows the model to adapt to the specific characteristics of our dataset while leveraging the knowledge learned from the pretrained weights. Subsequently, we reduced the learning rate and trained the entire model end-to-end for a total of 100 epochs. For the Unbalanced version of the datasets we have used learning rates of 1×10^{-4} and 1×10^{-5} with initial training epochs equal to 30. Whereas for the Balanced version, the learning rates used are 3×10^{-5} and 7×10^{-6} with initial training epochs equal to 5. This

gradual fine-tuning process enables the model to learn discriminative features from our dataset while preserving the valuable knowledge encoded in the pretrained weights.

6.3 Loss functions

In our experimentation, we employed specific loss and activation functions suited to the classification tasks at hand, viz., Multi-class and Multi-label classification. For Multi-class classification, where each sample belongs to one of multiple classes, we applied the softmax activation function at the output layer and utilized the Cross Entropy Loss function given by Equation 6.1.

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i}\sum_{c}\left[Y_c^{(i)}\log\left(P_c^{(i)}\right)\right] \tag{6.1}$$

where N is the total number of images, $Y_c^{(i)}$ is the ground truth indicating the presence of class c for the i^{th} image and $P_c^{(i)}$ is the predicted probability for class c for the i^{th} image. On the other hand, in Multi-label classification each sample may belong to multiple classes simultaneously. So for this scenario we used the sigmoid function after the models' output, coupled with a threshold of 0.5 to determine positive predictions. A modified Binary Cross Entropy function was utilized directly over the models' raw output represented by Equation 6.2. This combines sigmoid activation function with binary cross-entropy loss to get individual class probabilities. Throughout all experiments, we maintained the Adam [30] optimizer, an adaptive optimization algorithm widely used for training deep neural networks.

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i}\sum_{c}\left[Y_c^{(i)}\log\left(\left(P(Y_c^{(i)})\right)\right) + (1-Y_c^{(i)})\log\left(1-\left(P(Y_c^{(i)})\right)\right)\right] \tag{6.2}$$

where N is the total number of images, $Y_c^{(i)}$ is the ground truth indicating the presence of class c for the i^{th} image and $P_c^{(i)}$ is the predicted probability for class c for the i^{th} image.

6.4 Performance Metrics

- Accuracy: It is the most widely used indicator by statisticians to assess the model performance. It provides a ratio between the samples that have been correctly identified over all the other samples in the dataset.

- Precision: This metric represents the ratio of the number of correctly-classified positive samples out of the total samples correctly labeled as positive by our model.

- Recall: It represents the proportion of correctly identified positive samples out of all actual positive samples.

- F1-score: It is the harmonic mean of precision and recall.

$$f1\_score_i = \frac{2 \cdot precision_i \cdot recall_i}{precision_i + recall_i} \tag{6.3}$$

where $precision_i$ is the precision for class i and $recall_i$ is the recall for class i.

$$precision_i = \frac{TruePositive_i}{TruePositive_i + FalsePositive_i} \tag{6.4}$$

where $TruePositive_i$ gives the number of true positives for class i and $FalsePositive_i$ gives the number of false positives for class i.

$$Recall_i = \frac{TruePositive_i}{TruePositive_i + FalseNegative_i} \tag{6.5}$$

where $FalseNegative_i$ gives the number of false negatives for class i

- AUC: The ROC curve gives you a picture of the realration between the True Positive Rate (TPR) and False Positive Rate (FPR). TPR is nothing but recall, and FPR is a numeral mea-

sured by the relation between the number of the same data samples that were incorrectly predicted and the total number of data samples that were present to begin with.

The metrics of experiments are macro-averaged for each of the class or the categories. The class averaged macro-averaging operation calculates the metric for each class on a separate basis and then takes the average over all classes. By following this approach, one can have complete assurance that all the classes with very few data samples contribute similarly when it comes to evaluating the total performance, compared to other well represented classes.

6.5 Results

Table 6.3 displays the Floating Point Operations per second (FLOPS) and the number of parameters for each pretrained model. FLOPS range from 27.02 million (M) for EfficientNet to 35.06 billion (G) for Swin Transformer, while parameters vary from 48.42 thousand (K) for EfficientNet to 194.90 M for the Swin Transformer. Notably, the Swin Transformer exhibits the highest computational complexity, whereas EfficientNet demonstrates the lowest.

6.5.0.1 Results on VinDr-CXR Dataset

Table 6.4 shows the performance of various deep learning models on the VinDr-CXR dataset on the unbalanced version of the dataset. The top performers include PVT with an AUC of 0.9388 and ConvNeXt with 0.9267. CoAtNet leads precision at 0.7021, while DenseNet follows at 0.6722. For F1-score, CoAtNet excels at 0.5170, trailed by ConvNext at 0.4816. Notably, ConvNeXt and CoAtNet stands out the best among the CNNs, while PVT and SwinTransformer shines among the Transformers for this set of experiments.

Table 6.5 presents the performance metrics of various deep learning models on the balanced version of VinDr-CXR for the task of Multi-label classification. SwinTransformer leads with the highest AUC of 0.8390 and precision of 0.6658, followed closely by CoAtNet with an AUC of 0.8044 and precision of 0.6490. SwinTransformer also secures the highest F1-score of 0.5445, highlighting its effectiveness. Among the CNNs, CoAtNet emerges as a top model with competitive

S.No.	Model	FLOPS	Parameters
CNNs			
1	VGG-19 [45]	19.63G	139.59M
2	DenseNet [21]	2.75G	6.87M
3	EfficientNet [47]	**27.02M**	**48.42K**
4	ConvNeXt [35]	4.45G	27.80M
5	CoAtNet [11]	32.49G	163.27M
6	ResNet [19]	11.60G	58.15M
Transformers			
7	SwinTransformer [34]	**34.06G**	**194.90M**
8	MViT [15]	8.83G	50.47M
9	DaViT [12]	15.18G	86.87M
10	PVT [50]	11.33G	81.38M
11	GCViT [18]	13.94G	89.01M
12	EfficientViT [7]	102.45M	2.14M
13	MaxViT [49]	5.33G	30.27M

Table 6.3: FLOPS and Parameters for each model

performance in AUC and precision, whereas among ViTs, SwinTransformer stands out as the top performer across all metrics, demonstrating robustness.

Table 6.6 shows the abnormality wise AUC and Precision scores for CoAtNet and Swin Transformer. For almost all the abnormalities CoAtNet achieves a better AUC and Precision than Swin Transformer.

6.6 Results on VinDr-PCXR Dataset

Table 6.7 presents performance metrics of various models on the VinDr-PCXR dataset for Multi-label classification. CoAtNet achieves the highest AUC of 0.7316, while SwinTransformer follows closely with 0.7259. SwinTransformer leads in precision with 0.5137, while CoAtNet follows with 0.4239. In terms of F1-score, CoAtNet ranks highest with 0.1976, followed by PVT with

S.No.	Model	Accuracy	Precision	Recall	F1-score	AUC
CNNs						
1	VGG-19	0.9341	0.4714	0.3289	0.3573	0.8827
2	DenseNet	0.9511	0.6722	0.3673	0.4394	0.9241
3	EfficientNet	0.9455	0.6080	0.3284	0.3758	0.9135
4	ConvNeXt	**0.9543**	**0.6999**	**0.4273**	**0.4816**	0.9267
5	CoAtNet	0.9509	**0.7021**	**0.4520**	**0.5170**	0.9302
6	ResNet	0.9506	0.5657	0.4007	0.4558	0.8953
Transformers						
7	SwinTransformer	**0.9522**	0.5959	0.3932	0.4483	**0.9328**
8	MViT	0.9469	0.5887	0.3727	0.4436	0.9155
9	DaViT	0.9506	0.6713	0.3655	0.4435	0.9210
10	PVT	0.9507	0.5442	0.2985	0.3342	**0.9388**
11	GCViT	0.9414	0.3662	0.2332	0.2583	0.8909
12	EfficientViT	0.9462	0.5554	0.3113	0.3654	0.8894
13	MaxViT	0.9472	0.5467	0.3321	0.3851	0.8961

Table 6.4: Multi-label classification on VinDr-CXR unbalanced

0.2286.Among the CNNs CoAtNet achieves the highest AUC of 0.7316 and F1-score equal to 0.1976, while ConvNeXt maintains competitive performance across all metrics. Among Transformers, SwinTransformer stands out with the highest precision of 0.5137, followed by PVT with high accuracy of 0.8837 and a competitive F1-score of 0.2286.

Table 6.8 showcases the performance of different models on the VinDr-PCXR dataset for multilabel classification. DenseNet stands out with the highest accuracy of 0.7937 and precision equal to 0.4887, while CoAtNet excels in recall equal to 0.2900 and F1-score of 0.2909. Although SwinTransformer performs well overall, it lags behind in recall and F1-score compared to CoAtNet.

Table 6.9 displays the model performance on the VinDr-PCXR dataset for multiclass classification. EfficientNet leads among CNNs with an accuracy of 0.6758, while PVT tops the Transformer models with an accuracy of 0.6751. CoAtNet excels in recall equal to 0.2398 and AUC of 0.7254, while MaxViT achieves the highest precision of 0.4269.

Table 6.10 displays model performance on the VinDr-PCXR dataset for multiclass classification. SwinTransformer leads with an AUC of 0.6327, while EfficientViT showcases top precision value

S.No.	Model	Accuracy	Precision	Recall	F1-score	AUC
CNNs						
1	VGG-19	0.8058	0.4642	0.3432	0.3617	0.7358
2	DenseNet	0.8274	0.5707	0.3809	0.4180	0.7879
3	EfficientNet	0.8229	0.5318	0.3078	0.3457	0.7469
4	ConvNeXt	0.8341	0.6132	0.4221	0.4675	0.7862
5	CoAtNet	**0.8352**	0.6490	**0.4693**	0.5147	**0.8044**
6	ResNet	0.8304	0.5944	0.4162	**0.4762**	0.7801
Transformers						
7	SwinTransformer	**0.8390**	**0.6658**	**0.4851**	**0.5445**	**0.8041**
8	MViT	0.8208	**0.6625**	0.3848	0.4340	0.7697
9	DaViT	0.8300	0.5410	0.3534	0.3943	0.7812
10	PVT	0.8219	0.5979	0.4133	0.4566	0.7868
11	GCViT	0.3648	0.3415	0.2442	0.1872	0.5836
12	EfficientViT	0.8193	0.5440	0.3329	0.3756	0.7581
13	MaxViT	0.8195	0.4830	0.3206	0.3514	0.7661

Table 6.5: Multi-label classification on VinDr-CXR balanced

of 0.3066, recall equal to 0.3020, and F1-score equal to 0.3015. VGG-19 achieves the highest precision of 0.4023, and EfficientNet attains the highest accuracy of 0.3739. Among the CNNs VGG-19 emerges as the top performer with the highest precision of 0.4023. On the other hand, among the Transformer models, EfficientViT demonstrates superior performance, exhibiting the highest precision of 0.3066, recall equal to 0.3020, and F1-score equal to 0.3015.

Tables 6.11 and 6.12 shows the AUC and Precision values per diseease label for the Multi-label and Multi-class classifcation tasks respectively. For both, CoAtNet achieves superior AUC values and comparable Precision values.

Abnormalities	**AUC**		**Precision**	
	CoAtNet	**Swin**	**CoAtNet**	**Swin**
Aortic enlargement	0.9779	0.9764	0.9437	0.8571
Atelectasis	0.9405	0.9402	0.5714	0.6667
Calcification	0.8811	0.8748	0.4615	0.5000
Cardiomegaly	0.9777	0.9824	0.9724	0.8670
Consolidation	0.9615	0.9375	0.5385	0.4545
ILD	0.9362	0.9529	0.9000	0.5789
Infiltration	0.9440	0.9446	0.5385	0.5172
Lung Opacity	0.9304	0.9242	0.5660	0.6269
Nodule/Mass	0.9280	0.9058	0.4382	0.5556
Other lesion	0.9161	0.9069	0.3537	0.4146
Pleural effusion	0.9787	0.9771	0.8800	0.8429
Pleural thickening	0.9188	0.9208	0.8033	0.5137
Pneumothorax	0.7407	0.8532	1.000	0.0
Pulmonary fibrosis	0.9382	0.9190	0.6312	0.5913
No finding	0.9832	0.9754	0.9330	0.9524

Table 6.6: AUC and Precision values per abnormality label for the Multi-label classification on VinDr-CXR

S.No.	Model	Accuracy	Precision	Recall	F1-score	AUC
CNNs						
1	VGG-19	0.8701	0.3325	**0.2225**	**0.2432**	0.6375
2	DenseNet	0.8759	0.4007	0.1565	0.1958	0.6820
3	EfficientNet	0.8848	0.2469	0.1626	0.1399	0.6403
4	ConvNeXt	0.8905	0.2913	0.1548	0.1572	0.7210
5	CoAtNet	**0.8923**	**0.4239**	0.1809	0.1976	**0.7316**
6	ResNet	0.8795	0.3295	0.1958	0.2120	0.6367
Transformers						
7	SwinTransformer	**0.8915**	**0.5137**	0.1712	0.1906	**0.7259**
8	MViT	0.8695	0.3322	**0.2211**	**0.2324**	0.6537
9	DaViT	0.8816	0.4055	0.1390	0.1490	0.6801
10	PVT	0.8837	0.3695	0.2052	0.2286	0.7010
11	GCViT	0.8779	0.2813	0.1544	0.1606	0.6252
12	EfficientViT	0.8841	0.3317	0.1650	0.1664	0.6587
13	MaxViT	0.8784	0.4076	0.1787	0.1978	0.6450

Table 6.7: Multi-label classification on VinDr-PCXR unbalanced

S.No.	Model	Accuracy	Precision	Recall	F1-score	AUC
CNNs						
1	VGG-19	0.7779	0.4305	0.1728	0.2169	0.6307
2	DenseNet	**0.7937**	**0.4887**	0.0901	0.1275	**0.6506**
3	EfficientNet	0.7720	0.4341	0.1693	0.2219	0.6056
4	ConvNeXt	0.7737	0.3197	0.1379	0.1821	0.6268
5	CoAtNet	0.7528	**0.4540**	**0.2900**	**0.2909**	0.6366
6	ResNet	0.7683	0.3778	0.1361	0.1911	0.6142
Transformers						
7	SwinTransformer	0.7745	0.3973	0.2079	0.2454	0.6430
8	MViT	0.7754	0.2414	0.1141	0.1473	0.6119
9	DaViT	0.7787	0.4455	0.1590	0.1924	0.6072
10	PVT	0.7574	0.3355	**0.2480**	**0.2729**	0.6036
11	GCViT	0.7896	0.3220	0.0311	0.0512	0.6126
12	EfficientViT	0.7891	0.3887	0.0973	0.1299	**0.6479**
13	MaxViT	**0.7921**	0.4493	0.1287	0.1382	0.6182

Table 6.8: Multi-label classification on VinDr-PCXR balanced

S.No.	Model	Accuracy	Precision	Recall	F1-score	AUC
CNNs						
1	VGG-19	0.6092	0.2509	0.2121	0.2146	0.6015
2	DenseNet	0.6432	0.2823	0.2231	0.2290	0.6839
3	EfficientNet	**0.6758**	**0.3579**	0.1885	0.1771	0.6638
4	ConvNeXt	0.6721	0.2511	0.1905	0.1772	**0.7103**
5	CoAtNet	0.6617	0.3305	**0.2398**	**0.2412**	**0.7254**
6	ResNet	0.6151	0.2541	0.2230	0.2268	0.6465
Transformers						
7	SwinTransformer	0.6247	0.2865	**0.2357**	0.2339	0.6610
8	MViT	0.6706	0.2760	0.2137	0.2148	0.6978
9	DaViT	0.6580	0.3276	0.2322	**0.2422**	0.6766
10	PVT	**0.6751**	0.3399	0.1949	0.1887	0.6750
11	GCViT	0.6736	0.2337	0.1778	0.1567	0.6176
12	EfficientViT	0.6706	0.2525	0.1841	0.1694	0.6459
13	MaxViT	0.6743	**0.4269**	0.1784	0.1580	0.6414

Table 6.9: Multi-class classification on VinDr-PCXR unbalanced

S.No.	Model	Accuracy	Precision	Recall	F1-score	AUC
CNNs						
1	VGG-19	0.3378	**0.4023**	0.2842	0.2182	0.6138
2	DenseNet	0.3423	0.1187	0.2053	0.1287	0.5470
3	EfficientNet	**0.3739**	0.3221	0.2600	0.2194	0.6191
4	ConvNeXt	0.3333	0.2304	0.2160	0.1687	0.5324
5	CoAtNet	0.3536	0.2383	0.2408	0.2067	0.5974
6	ResNet	0.3446	0.3280	**0.2968**	**0.2962**	0.6115
Transformers						
7	SwinTransformer	0.3581	0.2802	0.2518	0.2262	**0.6327**
8	MViT	0.2365	0.2682	0.2459	0.2338	0.5702
9	DaViT	**0.3806**	0.3078	0.2643	0.2055	0.6151
10	PVT	0.3401	0.1008	0.2007	0.1217	0.5422
11	GCViT	0.3649	**0.3415**	0.2442	0.1872	0.5836
12	EfficientViT	0.3468	0.3066	**0.3020**	**0.3015**	**0.6319**
13	MaxViT	0.2927	0.2231	0.2328	0.2260	0.5793

Table 6.10: Multi-class classification on VinDr-PCXR balanced

Abnormalities	AUC		Precision	
	CoAtNet	Swin	CoAtNet	Swin
Bronchitis	0.7166	0.7256	0.5294	0.6364
Brocho-pneumonia	0.7911	0.7667	0.5000	0.5000
Other disease	0.6452	0.6302	0.3333	0.5000
Bronchiolitis	0.7024	0.7053	0.0	0.2222
Pneumonia	0.7856	0.7871	0.4286	0.4615
No finding	0.7488	0.7404	0.7521	0.7620

Table 6.11: AUC and Precision values per disease label for the Multi-label classification on VinDr-PCXR

Abnormalities	**AUC**		**Precision**	
	CoAtNet	**Swin**	**CoAtNet**	**Swin**
Bronchitis	0.7098	0.6467	0.2597	0.3269
Brocho-pneumonia	0.7997	0.6942	0.3750	0.1944
Other disease	0.6364	0.5777	0.2500	0.0
Bronchiolitis	0.6718	0.6416	0.0909	0.1176
Pneumonia	0.7919	0.7200	0.2838	0.3500
No finding	0.7428	0.6858	0.7237	0.7300

Table 6.12: AUC and Precision values per disease label for the Multi-class classification on VinDr-PCXR

Chapter 7

Conclusion

Through rigorous experimentation across diverse datasets and evaluation metrics, we have elucidated the prescribing power of various DL models, shedding light on their efficacy in both Muticlass and Multi-label classification. In our study, we explored various DL paradigms, ranging from residual connections to attention mechanisms to hybrid fusion of convolution and Transformers. It is evident that both CNNs and ViTs possess unique inductive capacities that contribute to their efficacy in CXR classification. CNNs excel in capturing intricate spatial hierarchies and fine-grained details, making them ideal as localized feature extraction tools. Whereas, ViTs leverage attention mechanisms to capture long-range dependencies crucial for disease diagnosis. An effective combination of CNNs and ViTs hence becomes paramount, offering promising avenues for improving clinical decision-making in the interpretation of CXRs. As a final note, our findings underscore the promising potential of leveraging a combination of CNNs and ViTs to enhance the accuracy and efficacy of disease diagnosis from CXRs, marking a significant step forward in the field of medical imaging and clinical decision-making.

Bibliography

[1] A. Abbas, M. M. Abdelsamea, and M. M. Gaber. Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network. *Appl Intell*, 51:854–864, 2021.

[2] Mohammed A. A. Al-qaness, Jie Zhu, Dalal AL-Alimi, Abdelghani Dahou, Saeed Hamood Alsamhi, Mohamed Abd Elaziz, and Ahmed A. Ewees. Chest x-ray images for lung disease detection using deep learning techniques: A comprehensive survey. *Archives of Computational Methods in Engineering*, February 2024.

[3] Imane Allaouzi and Mohamed Ben Ahmed. A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288, 2019.

[4] Goram Mufarah M. Alshmrani, Qiang Ni, Richard Jiang, Haris Pervaiz, and Nada M. Elshennawy. A deep learning architecture for multi-class lung diseases classification using chest x-ray (cxr) images. *Alexandria Engineering Journal*, 64:923–935, 2023.

[5] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, March 2021.

[6] Nigus Wereta Asnake, Ayodeji Olalekan Salau, and Aleka Melese Ayalew. X-ray image-based pneumonia detection and classification using deep learning. *Multimedia Tools and Applications*, January 2024.

[7] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Multi-scale linear attention for high-resolution dense prediction, 2024.

[8] Wen-Hsien Chang, Chih-Chieh Chen, Han-Kuei Wu, Po-Chi Hsu, Lun-Chien Lo, Hsueh-Ting Chu, and Hen-Hong Chang. Tongue feature dataset construction and real-time detection. *PLoS ONE*, 17(4):e0296070, 2022.

[9] Jun Cheng, Wei Huang, Sheng Cao, Rui Yang, Wei Yang, Zhe Yun, Zhen Wang, and Qing Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS One*, 10(10):e0140381, Oct 2015. Erratum in: PLoS One. 2015;10(12):e0144479.

[10] Zhiyong Dai, Jianjun Yi, Lei Yan, Qingwen Xu, Liang Hu, Qi Zhang, Jiahui Li, and Guoqiang Wang. Pfemed: Few-shot medical image classification using prior guided feature enhancement. *Pattern Recognition*, 134:109108, 2023.

[11] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021.

[12] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers, 2022.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[14] DungNB, Ha Q. Nguyen, Julia Elliott, KeepLearning, NguyenThanhNhan, Phil Culliton. Vinbigdata chest x-ray abnormalities detection, 2020.

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers, 2021.

[16] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty. Chest x-rays classification: A multi-label and fine-grained problem, 2018.

[17] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components

of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. [Online].

[18] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers, 2023.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[20] Min Hong, Beanbonyka Rim, Hongchang Lee, Hyeonung Jang, Joonho Oh, and Seongjun Choi. Multi-class classification of lung diseases using cnn models. *Applied Sciences*, 11(19), 2021.

[21] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[22] Emtiaz Hussain, Mahmudul Hasan, Md Anisur Rahman, Ickjai Lee, Tasmi Tamanna, and Mohammad Zavid Parvez. Corodet: A deep learning based classification for covid-19 detection using chest x-ray images. *Chaos, Solitons & Fractals*, 142:110495, 2021.

[23] A. U. Ibrahim, M. Ozsoz, S. Serte, and et al. Pneumonia classification using deep learning from chest x-ray images during covid-19. *Cogn Comput*, 2021.

[24] Khawar Islam. Recent advances in vision transformer: A survey and outlook of recent work, 2023.

[25] Y. Jin, H. Lu, W. Zhu, and W. Huo. Deep learning based classification of multi-label chest x-ray images via dual-weighted metric loss. *Comput Biol Med*, 157:106683, May 2023. Epub 2023 Feb 15.

[26] Yanling Jin, Xiaowei Gai, Jin Wang, Yanhui Chen, Heng Wu, Yanhong Liu, Lei Zhang, Ying Li, Shuai Huang, Xiangyu Zheng, Yingli Guo, Xian Zhao, Yan Zhang, Ling Zeng, Xiaofang Liu, Hong Yang, Xiaoyu Yang, Yuxiang Zhou, Hui Liang, Wei Liu, Kui Chen, and Jie Xu. A saliva-based rt-pcr detection assay for covid-19: Monitoring disease activity and recurrence. *Journal of Clinical Microbiology*, 58(7):e00832–20, 2020.

[27] Asifullah Khan, Zunaira Rauf, Abdul Rehman Khan, Saima Rathore, Saddam Hussain Khan, Najmus Saher Shah, Umair Farooq, Hifsa Asif, Aqsa Asif, Umme Zahoora, Rafi Ullah Khalil, Suleman Qamar, Umme Hani Asif, Faiza Babar Khan, Abdul Majid, and Jeonghwan Gwak. A recent survey of vision transformers for medical image segmentation, 2023.

[28] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, April 2020.

[29] Sungyeup Kim, Beanbonyka Rim, Seongjun Choi, Ahyoung Lee, Sedong Min, and Min Hong. Deep learning in multi-class lung diseases' classification on chest x-ray images. *Diagnostics*, 12(4):915, 2022.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1097–1105. 2012.

[32] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[33] Shiwei Liu, Liejun Wang, and Wenwen Yue. An efficient medical image classification network based on multi-branch cnn, token grouping transformer and mixer mlp. *Applied Soft Computing*, 153:111323, 2024.

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

[36] Meng Lv, Xufei Luo, Janne Estill, Yunlan Liu, Mengjuan Ren, Jianjian Wang, Qi Wang, Siya Zhao, Xiaohui Wang, Shu Yang, Xixi Feng, Weiguo Li, Enmei Liu, Xianzhuo Zhang, Ling Wang, Qi Zhou, Wenbo Meng, Xiaolong Qi, Yangqin Xun, Xuan Yu, Yaolong Chen,

(on behalf of the COVID-19 evidence, and recommendations working group). Coronavirus disease (covid-19): a scoping review. *Eurosurveillance*, 25(15):2000106, 2020.

[37] Umar Marikkar, Sara Atito, Muhammad Awais, and Adam Mahdi. Lt-vit: A vision transformer for multi-label chest x-ray classification. In *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, October 2023.

[38] Ahmed F. Mohamed, Amal Saba, Mohamed K. Hassan, Hamdy.M. Youssef, Abdelghani Dahou, Ammar H. Elsheikh, Alaa A. El-Bary, Mohamed Abd Elaziz, and Rehab Ali Ibrahim. Boosted nutcracker optimizer and chaos game optimization with cross vision transformer for medical image classification. *Egyptian Informatics Journal*, 26:100457, 2024.

[39] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2022.

[40] Ha Quy Nguyen, Hieu Huy Pham, Tuan Linh, Dao Minh Le, and Lam Khanh. VinDr-CXR: An open dataset of chest x-rays with radiologist annotations. PhysioNet, 2021. Version 1.0.0.

[41] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[42] Hieu Huy Pham, Tien Thanh Tran, and Ha Quy Nguyen. VinDr-PCXR: An open, large-scale pediatric chest x-ray dataset for interpretation of common thoracic diseases. PhysioNet, 2022. Version 1.0.0.

[43] S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*, pages 337–345, 1990.

[44] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[46] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *medRxiv*, 2020.

[47] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[48] Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers, 2022.

[49] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer, 2022.

[50] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, 2021.

[51] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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)*. IEEE, July 2017.

[52] Xin Wu, Yue Feng, Hong Xu, Zhuosheng Lin, Tao Chen, Shengke Li, Shihan Qiu, Qichao Liu, Yuangang Ma, and Shuangsheng Zhang. Ctranscnn: Combining transformer and cnn in multilabel medical image classification. *Knowledge-Based Systems*, 281:111030, 2023.

[53] Srinivas Yallapu and Aravind Kumar Madam. A chest x-ray image based model for classification and detection of diseases. In Prakash Pareek, Nishu Gupta, and M. J. C. S. Reis, editors, *Cognitive Computing and Cyber Physical Systems*, pages 422–432, Cham, 2024. Springer Nature Switzerland.

[54] Hengde Zhu, Wei Wang, Irek Ulidowski, Qinghua Zhou, Shuihua Wang, Huafeng Chen, and

Yudong Zhang. Meednets: Medical image classification via ensemble bio-inspired evolutionary densenets. *Knowledge-Based Systems*, 280:111035, 2023.