NTNU

Kunnskap for en bedre verden

Department of Computer Science

IMT4894 - Advanced Project Work

Transformers In Medical Domain: Swin Transformer as a Binary Classification Model for Pneumonia

Author:
Alen Bhandari

Dec, 2021

Chapter 1

Abstract

In the field of computer vision, Machine Learning (ML) Models has taken a huge leap compared to what we could do some years before. And, due to the digital age,we cannot comprehend the amount of data that are growing day by day due to which we have huge computer vision datasets like ImageNet,Cifar and other many more. But unlikely, for the medical domain finding dataset that have enough training and testing dataset is still a difficult task. Due to which researchers developed a concept of transfer learning which is the state-of-the-art model for computer vision tasks. Here, the model is trained in a large dataset and then it is implemented by hyper tuning the parameters for our own problem domain.

In the midst of emerging technology, there was particularly one specific machine learning model that caught the attention of researchers: 'Transformers'. They are attention based model which was able to beat the state-of-the-art model for computer vision tasks but it is still unexplored properly in Medical domains. Due to which in this projects we are mainly focusing on the use of transformer as a binary classification model for Chest X-ray images. Throughout this project we will be going more in-depth about the architecture of transformers and their implementation in computer vision task.However, this project does not focus on producing a new findings, rather, on testing the performance of transformers for chest X-ray images compared to different State-of-the-art Convolution Neural Networks (CNNs), Deep Neural Networks(DNN) models.

Contents

Figures

Chapter 2

Introduction

Pneumonia is a medical condition caused by bacteria or viruses which mainly affects individuals and infants. It is an acute pulmonary infection that affects lungs, which causes the inflammation of the air sacs and pleural effusion, situation where lungs are filled with fluid. According to WHO (World Health Organization), it accounts for 14% of all deaths in children under the age of 5, killing almost 740,180 children's world wide in 2019 [1]. Pneumonia are most common in the underdeveloped countries and developing countries where there exist problems such as overpopulation, air pollution's, unhygienic environmental conditions and specially in the area where medical resources are minimal. Hence, early detection of pneumonia is essentially important for preventing it from being fatal. Usually, pneumonia are diagnosed interpreting different radiology examinations of lung such as computed tomography(CT),radiography(X-ray) or Magnetic Resonance imaging(MRI). In this project we are concerned with the Chest X=ray images.

The advancement of technology has made it possible for researchers to build complex systems due to the availability of huge computational power and storage unit. In recent years, Artificial Intelligence (AI), has been affecting almost every sectors, and medical domain is no different from that. The amount of investments, research that has been put in the AI domain is enormous also the advancement that we have seen compared to past years is huge [2]. Different applications use a form of AI one way or another, be it a RPA(Robot Process Automation), chatbots, or different healthcare analytics, AI is every where. But specifically in medical domain, problem domain such as classification and segmentation is highly influenced by AI. Application of AI in the medical domain is basically reading either numerical data(for example,blood pressure or heart rate) or image-based (which our project is based on) as a input. After that, the algorithm learn from the input data and give the output either a probability or a classification(in some case followed by segmentation of infected portion).

Different tools of AI has been extensively explored in the field of computer vision problems(Medical problems) namely Deep learning Networks(DLM) and Convolution Neural Networks(CNNs) [3], [4]. However, such models perform with good accuracy only only when provided with large amount of database. In case of biomedical image classification problem, such vast amount of labeled database if really difficult as the expert doctors needs to classify each image which is an expensive and time consuming task. Due to which the concept of transfer learning was introduced. Transfer learning is basically training the CNNs model in a huge dataset such as ImageNet [5] (consisting of more than 14 million images) and then use the pre-trained model for the smaller dataset by tuning the parameter of the pre-trained model.

Transformers [6], introduced by Google in June 2017, that uses seq2seq model where it takes a sequence as input and returns a sequence as output. It was originally introduced for the task of Natural Language Processing (NLP) [7] where it astounding performance beating NLP the state=of-the-art model [8]. After the success in the NLP domain researchers were focused on using it's architecture for the computer vision task. That's when they introduced ViT (Vision Transformer) [9], the first computer vision transformer. The reason for transformer being so popular is the model encoder, where the encoding is done for the whole sequence(sentence) compared to traditional CNNs where encoding is done word by word. Additionally, transformer revolutionized the traditional use of attention mechanism.

In this project we are curious about testing the performance of transformers for medical images specially radiology images. Also, since the experiment was conducted in the local machine so we have reduced the amount of dataset, also the size of the images. Due to which our experimental setup is different from the general machine learning setup for computer vision problem where they exploit the model with huge dataset and large image size. Also, to test and evaluate the performance of the models we will be using different evaluation metrics such as accuracy, precision, recall, Area under the curve (AUC), confusion matrix and some others.

Chapter 3

Related Work

In medical domain, transformers being developed for NLP(Natural Language Processing) task first, it has been used extensively for problems such as radiology report generation except for computer vision task. Since, the findings of the report is crucial so it might be critical if left overlooked by respective clinician [10]. But finding relevant results from the radiology report was a difficult task on it's own [11], [12]. However, the advancement of NLP technologies made it possible to detect the actionable findings as well as various other task using radiology reports [13]. In 2021, [14], implemented a bidirectional encoder representation (BERT), a transformer based model and compared it with different state-of-the-art models for radiology reports and the results from the paper shows that without the order information, it achieved the highest area under the precision-recall curve .

After the success in the NLP domain, researchers were curious to see it's performance in the computer vision tasks. The main reason for them to move to transformers leaving behind the traditional CNNs was it's independence of inductive biases. However, implementing transformers to vision task was not easier due to the typical structure of the visual data. Hence, it required novel model designs and training schemes, which as a result transformer models and their variants are been extensively used for image recognition [9], [15], object detection [16], [17], segmentation [18], image generation and many other use cases.

Transformers are based on attention mechanism, but it is not the first time it has been implemented. It has been extensively used in feed-forward and recurrent network [19], [20], but transformers are completely based on the attention model which allows completely new implementation called multi-head attention, optimized for parallelization. The main advantage of using multi-head attention is their ability to form highly-complex models and large datasets. Transformers requires minimal prior knowledge about the structure of the problem compared to Convolution and recurrent networks [21], [22], due to which they are usually pre-trained on a large scale unlabelled dataset. And then they are fine tuned in a supervised manner to obtain the results.

Similar to computer vision task, different use case of transformers are extensively used for Medical domain problems but it is still in the earliest stage. Therefore, finding relevant literature in the subfield is limited which allows the flexibility to explore the potentiality of transformers in the medical domain.

In 2021, [23], used Vision Transformers(ViT) on diseased lungs, namely COVID-19 infected lungs and normal lungs. To test the performance of ViT they segregated the images into different size patches and used different metrics to measure the performance. Based on their result, the best accuracy they got for ViT was 95.36%.

In 2020, [24], used a hybrid Transformer model called GANBERT model which explores the possibility of synthesising PET images from MRI T1-weighted image, removing the burden for multiple scans. The author here tries to combine the part of CNN model inspired by Generative Adversarial Network (GAN) where as the Discriminator from BERT model. This kind of models where they combine transformer with CNN is called the Hybrid model. Their main approach was to generate PET images from MRI images in wide intensity range with no manipulation in pre-post processing.

In 2021, [25], proposed a generative adversarial approach for medical image synthesis called ResViT where they combine local precision of CNN operator with contextual sensitivity of vision transformers. Previously, used GAN models based on CNNs performed local processing with compact filters which compromised the learning of contextual features. But the proposed model combined both the convolutional and transformer model giving rise to ResVit. According to their findings, ResViT showed superiority against the traditional methods in term of qualitative observations and quantitative metrics.

In 2020, [26], introduced a new hybrid transformer called HATNet focusing on the Histopathology images which are basically large even on the gigapixel range. Conventional approach to address histopathology is to use patch-level CNNs for splitting the images into smaller patches and then process each patch independently. The problem with patch-processing is the algorithms looses information from the distant patches. Due to which the authors came up with a new solution called HATNet, which is a hybrid approach of self-attention and CNNs to address the problem. HATNeT basically solved the problem by processing the image at three levels: words, bags and images(Similar to NLP approach).

Chapter 4

Background

4.1 Transformers

The Introduction of transformers was a revolution in it's own way. Researchers were trying to move the dependencies of machine learning models from inductive bias. That's what gave arise to attention based model called transformers by Google in 2017. The transformers basically implement three types of attention mechanism: Self-attention in the case of encoder and decoder and attention in encoder-decoder. A brief description of different attention mechanism are given below:

- Self-attention mechanism is used to evaluate the link between the element in a sentence. For, example noun and a pronoun.
- The attention mechanism is used to evaluate the link between the encoded and decoded elements.
- Transformers also have several attention mechanism (namely multi-head attention) in the same sentence.

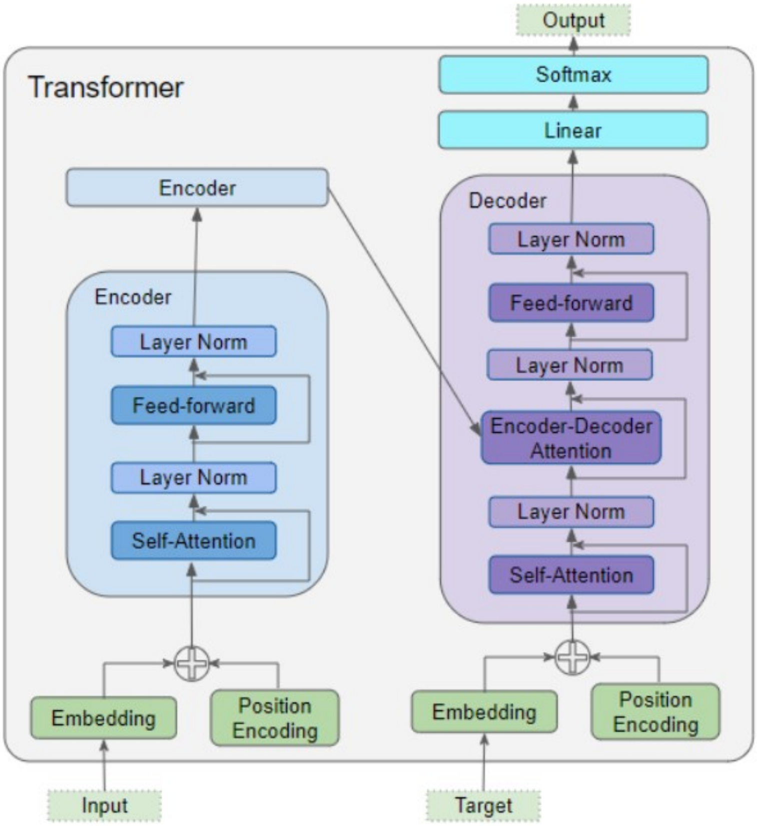
4.1.1 Architecture Overview:

Transformer are composed of multiple components as we can see in the figure 4.2, Different layers of transformer along with the brief description are explained below:

1. We need to pre-process the input before we feed it to encoder and decoder. The major pre-processing step involves:
   - **Embedding Layer**, it encodes the meaning of the word.
   - **Position Encoding Layer**, it represents the position of the word.
2. The encoder of the transformer is the stack of different layers. Each encoder in the transformer consists of:
   - **Multi-Head self-Attention Layer**, here the input sequence pays attention to itself.

- t**Feed-forward layer**

3. The decoder stack also contains number of layers. Each decoder consists of:
    - **Two Multi-Head self-Attention Layers**, here the target sequence pays attention to itself.
    - **Feed-forward layer**, the self-attention output is then passed to Feed-forward layer which is then send the output to the next decoder.
4. The output generates the final output which mainly contains:
    - **Linear Layer**, it consists of three separate linear layers for the Query, Key and Value. The inputs are passed through those Layers to produce Q,K and V matrices.
    - **Softmax Layer**, it is the activation function in the output layer which predicts the multi-nominal probability distribution.



(Image by Author)

Figure 4.1: Basic Architecture Of Transformers, Source: [27]

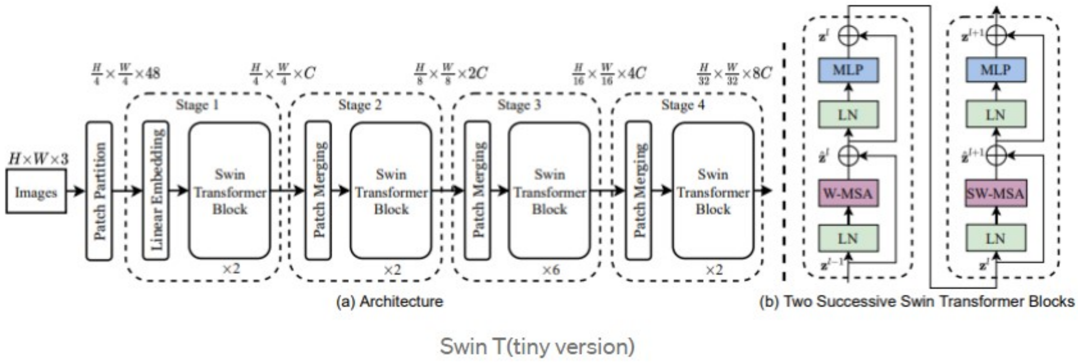
4.2 Swin Transformer

The Swin Transformer is one of the variation of Vision Transformer. In the field of computer vision, they have been able to show excellent performance in different vision use cases such as object detection, image classification, semantic segmentation and most probably any vision task. So, the problems with initial Vision Transformer (ViT) were its ability for adaptation to fully supplement convolutions. Swin transformer has the ability to model the difference between two separate domains like variations in scale of objects and the high resolution of pixels in image more efficiently. Due to which swin transformers can be served as general-purpose pipeline for vision.

According to the original paper of Swin Transformer [28], it describes swin transformer as a hierarchical Transformer that is computed using shifted windows. The main highlight of the swin Transformer as suggested by the paper are:

- Hierarchical representation, starting from a small-sized patches while gradually increasing the size through merging to get scale invariance,
- With the help of shifted window it achieves efficient, linear computational complexity with the help of self-attention locally,

4.2.1 Network Architecture



(a) Architecture

(b) Two Successive Swin Transformer Blocks

Figure 4.2: Basic Architecture Of Transformers, Source: [29]

Based on the network architecture we can see 4 unique building blocks in the diagram above. The transformer basically splits the RGB images into patches by the patch partition layer where each patch is 4 * 4 * 3 (3 being the RGB channels) which is considered as ’token’. Then a linear embedding layer is applied on the raw-valued feature to convert the patch into an arbitrary dimension denoted by c. The transformer block is developed by replacing mostly used multi-head attention

(MSA) module with a module based on the shifted windows. The main architecture of swin transformer is is composed of multiple stages (Mainly 4 stages for Swin-T):

1. **Stage 1:** Swin Transformer block containing linear embedding with Swin Transformer where the length of token is lessened using patch merging layers,
2. **Stage 2:** the patch merging layer then concatenates the feature of each group,
3. **Stage 3 & 4:** the patch merging in stages 2,3 and 4 jointly are responsible for producing the hierarchical representation which aligns to the feature map generated in CNNs, which allows swin transformer ready for the general vision tasks .

Hence, if we talk briefly about the entire architecture of swin transformers, it consists of a shifted window-based MSA along with a 2=layer Multi-Layer Perceptron (MLP) with Gaussian Error Linear Unit (GELU) as a activation function. GELU is basically, a high performing neural network activation function that assign weights to the input based on their values rather than gates inputs by their sign as RELUs. Finally, a layer Norm(LN) is applied before MSA module and each MLP, and a residual connection follows after each module.

Chapter 5

Methodology

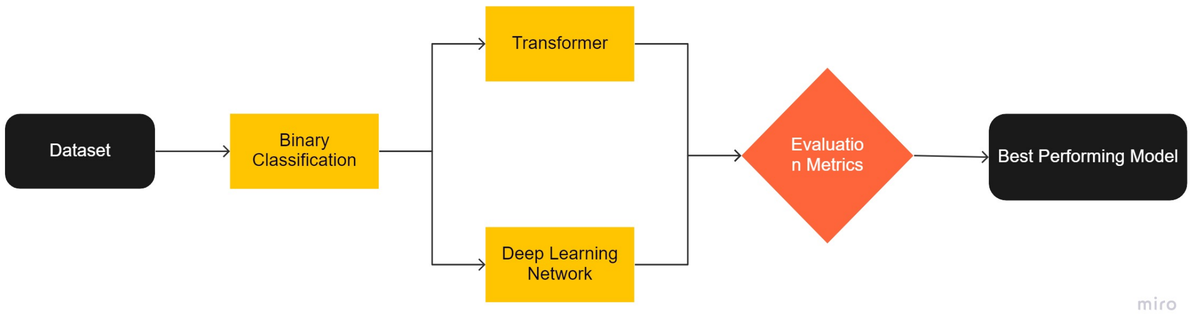


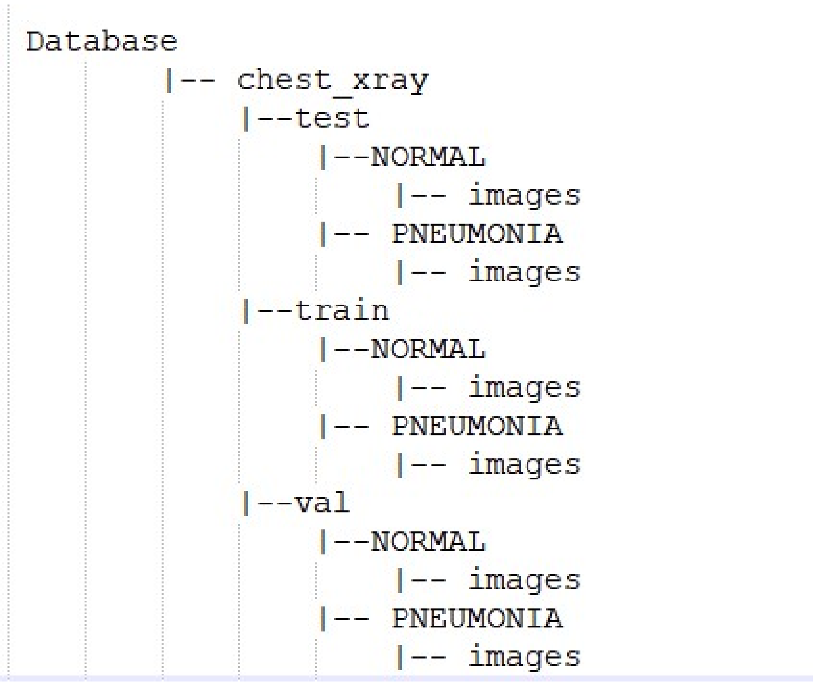
Figure 5.1: Proposed System Flow Chart

The proposed system pipeline is as shown in the figure. Since, we are interested to test the performance of Transformers in the chosen dataset, so we will be using different machine learning evaluation metrics as our main methodology. Also, to create the baseline for comparing the transformer performance we will be implementing state-of-the-art Deep Learning model. Major components of the proposed system are:

1. Dataset
2. Binary Classification
    - Transformer
    - Deep Learning Model
3. Evaluation Metrics

5.1 Dataset

The dataset that was selected for the project was Chest X-Ray images(pneumonia) from the official site of Kaggle [30]. The main reason to choose this dataset was it's gray scale nature where the information of pixels in the image is limited compared to the color images. Also, due to the reason that the chosen dataset being another domain of medical field compared to the Skin lesion domain used in IMT4895- Specialization In Colour Imaging (Experiment conducted using Vision Transformer(ViT) for color images). The folder organization for the performed experiment is shown in the figure below.



```
Database
        |-- chest_xray
            |--test
                |--NORMAL
                    |-- images
                |-- PNEUMONIA
                    |-- images
            |--train
                |--NORMAL
                    |-- images
                |-- PNEUMONIA
                    |-- images
            |--val
                |--NORMAL
                    |-- images
                |-- PNEUMONIA
                    |-- images
```

Figure 5.2: Folder Organization

A total of 1316 images was used for training the models where 641 images were Normal and 675 images were images of pneumonia. For testing the model the total of 624 images were used where 234 images were normal and 390 images were pneumonia. Finally, for validation we only had 16 images where 8 images were selected for normal and pneumonia cases respectively.

5.2 Binary Classification

Binary classification is the process of classifying the elements of the set into two groups on the basis of classification rules [31]. Typical problems related to binary classification includes:

- Medical classification, where the model predicts whether the patient has certain disease or not,
- in the industry for quality control, to predict whether the product meets a

particular specification or not,
- in the case of information retrieval, where the model predicts whether the particular page is relevant or not in the result set of a search.

5.2.1 Vision Transformer

The vision transformer that we have selected for this project is swin Transformer. The author in [28], claims that swin transformer can be used as a general purpose transformer. Also, from the literature research we know that transformers usually require huge amount of dataset to be able to perform accurately. Due to these reasons we are interested to test the performance of swin transformer for the medical domain where we have fixed the number of train,test and validation dataset.

The swin transformer was implemented directly from the paper [28], due to which the transformer implemented is in it's pure form without any modification. To match the input format I had to reduce the size of the chest x-ray images of my dataset which was 32 * 32 * 3 along with the patch size of 2. The dropout rate of the model was 0.03, with embedding dimension of 64. The attention head was of size 8, attention window of size 2 and the shifting window of size 1. Finally, the number of Multi-Layer Perceptron of the model was 256 with output dense layer having only 2 classes for predicting NORMAL and PNEUMONIA(Binary Classification).

5.2.2 Deep Learning Model

The deep learning model that I have selected for this experiment is Efficient Net which was first introduced in 2019 by Tan and Le [32]. After the release it was able to reach the state-of-the-Art performance on both imagenet [33](one of the biggest dataset available for computer vision task, also used as a benchmark dataset for the evaluation of the models) and different image classification transfer learning task.

The model implementation used in the project is referred from the official Keras website [34]. According to the website there are two versions of EfficientNet model available using Keras library, for this project we are implementing pretrained EfficientNetB0 model trained in the imagenet dataset. The input image shape of the model is (224,224,3) where the pixel value in the input image are in the range [0,255]. During the implementation we freeze all the layers and train only the top layers so that we could take advantage of the pre-trained weights. The learning rate used for the training is 1e-2 while the epoch is limited to 25.

5.3 Evaluation Metrics

For any machine learning model accuracy only is not enough for evaluating the model, for example the model might be having good accuracy but on the other side it might have poor results when evaluated with other metrics such as logarithmic_loss or any other such metrics. Due to which for this projects I have selected 7 metrics which are:

1. Classification Accuracy
2. Logarithmic Loss
3. Confusion Matrix
4. Area under Curve (AUC)
5. Precision
6. Recall
7. ROC Curve

5.3.1 Classification Accuracy

It is the ratio of total number of correct prediction to total number of input samples. It is more favourable when the input samples are balanced. Classification accuracy of the models are good but it gives a false hope of achieving good results.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Figure 5.3: Classification Accuracy

5.3.2 Logarithmic Loss

Logarithmic loss is the negative value given as a penalty wherever the model gives false classification. It is the probability value given to each class for all the samples. Let us suppose there are N samples that belongs to M classes, then the Log Loss is given by the below formula:

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

Figure 5.4: Logarithmic Loss

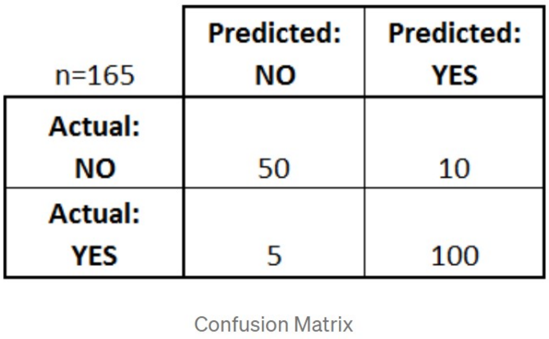
where,

- y_ij means whether the sample i belong to the class j pr not,
- p_ij means the probability of samples i belonging to class j.

Basically, minimising Log Loss means higher accuracy for the classifier model.

5.3.3 Confusion Matrix

It gives the result in a form of matrix which describes the overall performance of the model. To explain it a bit further, lets take an example of a binary classification problem. In a binary classification there are only two classes one true one false. Lets say we have some samples belonging to two classes: YES or NO. Also, lets say we have an sample of 165 images and the results that we get is as shown below:



n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Confusion Matrix

Figure 5.5: Confusion Matrix

The main important terms that we need to take into account while calculating the confusion Matrix are:

- **True Positive:** The total number of corrected prediction i.e. the model predicted YES when the actual output was also YES.
- **True Negatives:** The total number of corrected prediction i.e. the model predicted NO when the actual output was also NO.
- **FALSE Positive:** The total number of wrong prediction i.e. the model predicted YES when the actual output was also NO.
- **FALSE Negative:** The total number of wrong prediction i.e. the model predicted NO when the actual output was also YES.

5.3.4 Area Under Curve (AUC)

It is one of the widely used binary classification metrics for evaluating the model. It is an aggregated value of a performance of a binary classifier model on all possible threshold values. It basically calculates the ROC curve due to which it's value lies between 0 and 1. Generally, higher the AUC of a model the better it is.

5.3.5 Receiver Operating Charactier (ROC) Curve

It is a plot showing the performance of a binary classifier as a function of it's cut-off threshold. Basically, ROC curve gives the True Positive rate (TPR) against the False Positive Rate (FPR) for different threshold value.

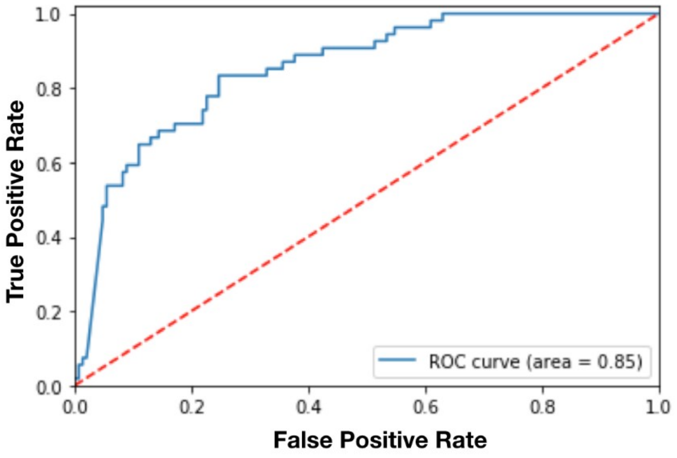


Figure 2. A sample ROC curve.

Figure 5.6: Area Under Curve

ROC Curve are generally used to look at the overall performance of the model and help select the right cut-off threshold for the model.

5.3.6 Precision

It is simply the number of correct positive results divided by the total number of positive results predicted by the classifier.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Precision

Figure 5.7: Precision

5.3.7 Recall

Recall calculates how many of the Actual positives our model captured through labeling it as positive (TP). For example, in case of medical domain, if a sick patient (True Positive) goes through the test and predicted as not sick (False Negative).

Due to which the cost associated with False Negative will be really high if the disease is contagious. Hence, for a model associated with the medical domain high recall value is more favourable.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Figure 5.8: Recall

Chapter 6

Results

As discussed in the chapter5, I have used number of metrics for evaluating the model. The model performance based on the different metrics are reported in this chapter. We will be discussing what those value actually mean in the discussion section in chapter 7.

6.1 Classification Accuracy

Based on the experiment performed model the reported accuracy train accuracy for swin transformer was 88% and the test accuracy was 98% which is a higher value.

In case of EfficientNetB0 the model train accuracy was 85% and the test accuracy was 73% which was low compared to the Swin Transformer.

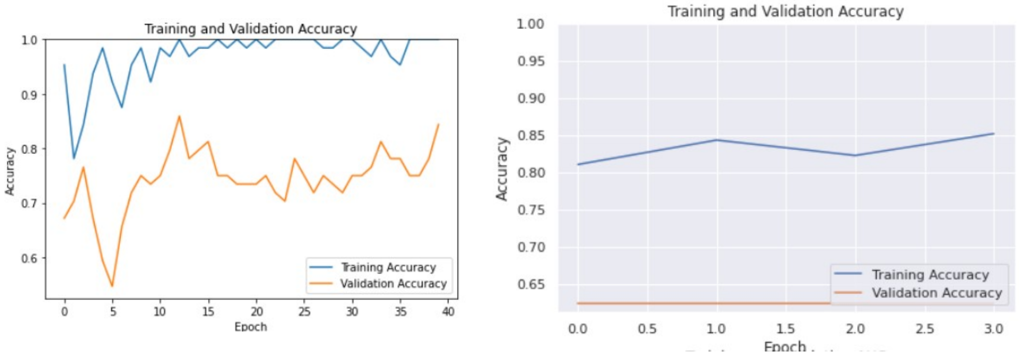


Figure 6.1: The Left Plot is the training and testing (validation) accuracy for Swin Transformer while the Right Plot is the training and testing (validation) accuracy for the EfficientNetB0

6.2 Logarithmic Loss

As mentioned in the Methodology chapter, only testing the accuracy is not enough for evaluating the model due to which we need some more metrics for the evaluation of the model. Based on the experimental setup the training loss for the Swin Transformer was recorded 48% while the testing loss was 24% which is less.

For EfficientNetB0, the training loss was recorded 53% and the testing loss was recorded 4.5 i.e 450% which is not possible. So, there must be some outliers in the dataset that caused the anomaly. Due to the time constraint the outliers were not identified and since we are using other metrics for the evaluation and comparison. So, I am just reporting the findings.

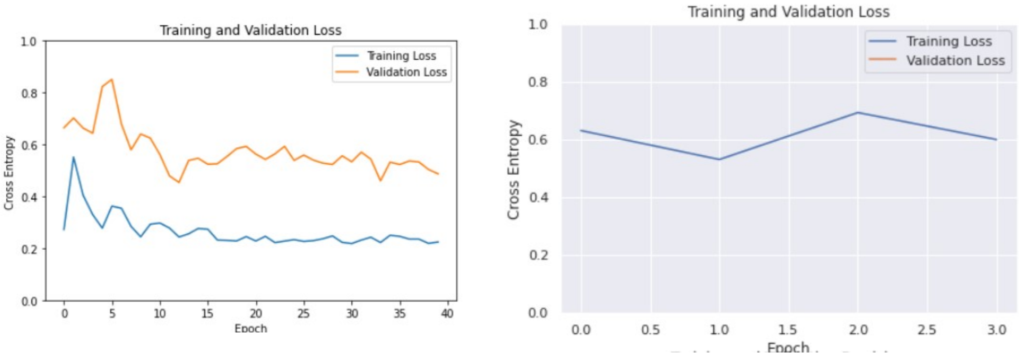


Figure 6.2: The Left Plot is the training and testing (validation) loss for Swin Transformer while the Right Plot is the training and testing (validation) loss for the EfficientNetB0

6.3 Confusion Matrix

It is one of the important metrics which give the overall performance of the binary classification metrics. It gives the total value of True Positive, True Negative, False Positive and False Negative predicted by the model.

For Swin Transformer, in a sample of 128 records where 64 of them were Pneumonia. The total number of True Negatives and True Positives were 52 while the total number of False Positives and False Negatives 12.

For EfficientNetB0, in a sample of 128 records where 64 of them were Pneumonia. The total number of True Negatives and True Positives were 47 while the total number of False Positives and False Negatives 17.

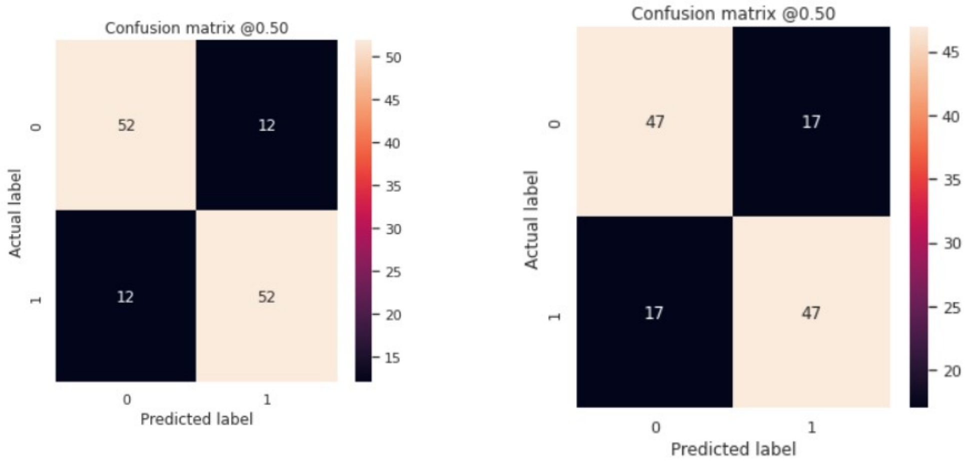


Figure 6.3: The Left figure is the confusion Matrix for Swin Transformer while the Right figure is the confusion Matrix the EfficientNetB0

6.4 Area Under the Curve (AUC)

Area Under the Curve (AUC) basically measures the ability of a classifier to distinguish between the classed and is also used as a summary of the ROC curve.

Based on the experimental setup, the AUC value for training dataset for swin transformer was recorded 0.86 and for validation dataset it was recorded 0.98.

For EfficientNetB0, the AUC value for training dataset was 0.92 while for validation dataset it was recorded 0.73.

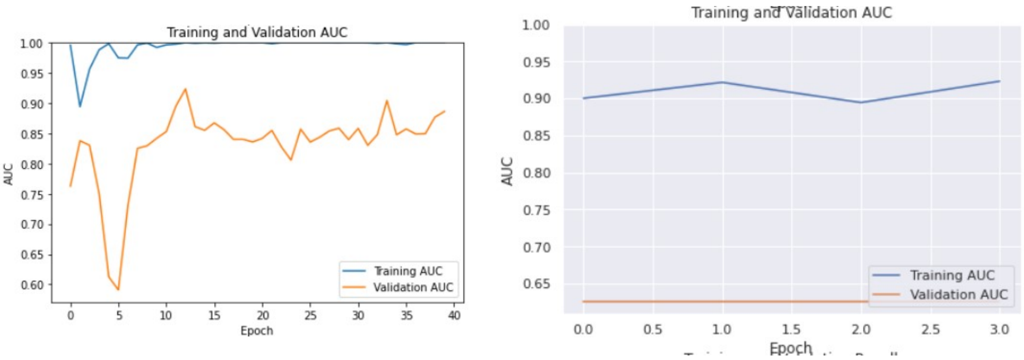


Figure 6.4: The Left Plot is the training and testing (validation) AUC for Swin Transformer while the Right Plot is the training and testing (validation) AUC for the EfficientNetB0

6.5 Receiver Operating Charactier (ROC) Curve

ROC curve are usually implemented for binary classifier. They shows the trade-off between the sensitivity and specificity. Usually, the model that produce curves closer to the top-left corner are considered to be better performing model. Based on the experimental setup the ROC curve of both the models are shown below:

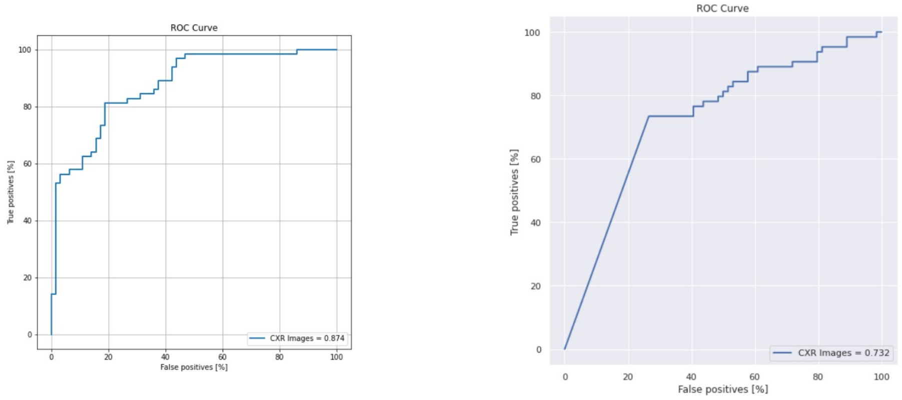


Figure 6.5: The Left figure is the ROC curve for Swin Transformer while the Right figure is the ROC curve for the EfficientNetB0

6.6 Precision

Precision is usually used to quantify the total number of correct positive prediction made by the model. So, based on the experimental setup the precision for training dataset of swin transformer was 81% while for the testing dataset it was 98%.

For EfficientNetB0, the precision for training was 84% and the testing it was 73%.

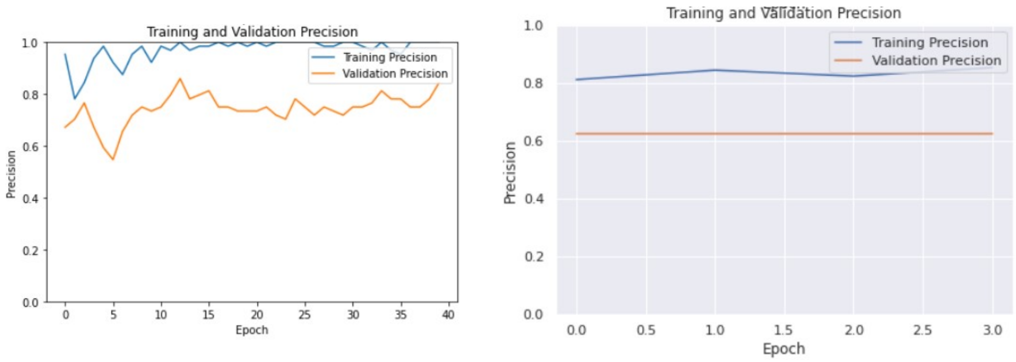


Figure 6.6: The Left Plot is the training and testing (validation) Precision values for Swin Transformer while the Right Plot is the training and testing (validation) precision values for the EfficientNetB0

6.7 Recall

Recall is little different then Precision, unlike precision which considers only correct prediction out of all positive prediction, recall provides and indication of missed positive predictions too. So, based on the experimental setup the recall for training dataset of swin transformer was 81% while for the testing dataset it was 98%.

For EfficientNetB0, the recall for training was 84% and the testing it was 73%.

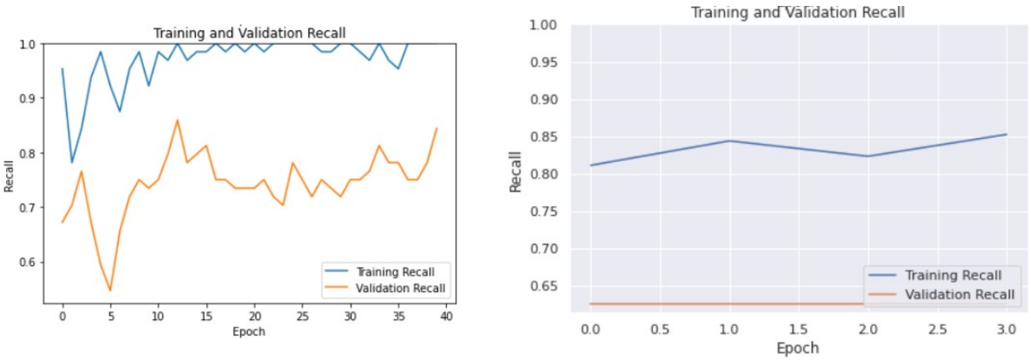


Figure 6.7: The Left Plot is the training and testing (validation) recall values for Swin Transformer while the Right Plot is the training and testing (validation) recall values for the EfficientNetB0

Chapter 7

Discussion

From the experiment performed I can infer that the swin transformer performs better than the traditional state-of-the-art model i.e EfficientNetB0 for Chest X-ray images.

Based on the different metrics values and on the experimental setup we can see that swin transformer beats EfficientNetB0 in every metrics. If we look at the accuracy and loss value then swin Transformer has high accuracy and low loss value compared to EfficientNetB0 which means that the swin transformer was able to give higher correct prediction and the low loss value means it has less false classification.

Since, confusion metrics is one of the major metrics for evaluating the performance of the model. For this metrics also Swin Transformer has higher True Positive i.e. 52 which means it was able to predict 52 correct pneumonia cases out of 64 pneumonia cases compared to EfficientNetB0 which had a value of 47 for True Positive. Also, the swin transformer has less False Positive and less False Negative compared to EfficientNetB0.

If we evaluate the model based on the area under the curve we can test the effectiveness of the model. The higher the AUC value the better is the model. For example, if the model has AUC value of 1 which is the maximum value, it means the model is classifying correctly both the classes. Based on our experiment results, Swin Transformers have high AUC value i.e 0.98 on the test dataset compared to EfficientNetB0 which has the value of 0.73. The value of 0.98 means swin Transformer were able to classify the labels more efficiently compared to EfficientNetB0.

Also, if we look at the Receiver Operator Characteristic(ROC) curve we can see in figure of ROC curve in the result section, that swin transformer has better ROC curve than EfficientNetB0. It's an important binary classifier because it is the probability curve that plots TPR against FPR at different threshold values.

Finally, Precision and recall are also two extremely relevant model evaluation metrics. Since, precision gives the percentage of the result that are relevant while recall gives the total relevant results that are correctly classified by our algorithm. Hence, the model having high precision and recall are more favourable. And based on our experimental setup, swin transformer has the high precision and recall values i.e 84% respectively compared to EfficientNetB0 which has the precision and recall values as 73%.

Chapter 8

Conclusion

The main objective of this report was to test the performance of the swin transformer for the chest X-ray images along with a small dataset scenario. Also, in the literature research I got to know that swin transformers can be implemented for general purpose computer vision task. So, other motive of the experiment was to test the generalizability of the swin transformer in the medical domain which is a separate domain in the computer vision tasks.

Hence, based on the experiment and different metrics value I can conclude that swin transformers are best suited for the medical domain too and it performs well despite having the small dataset. But to fully conclude the experimental results I need to test the performance of swin transformer further on the bigger dataset which will be kept as a next phase of the project.

Bibliography

[1] *Pneumonia*. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/pneumonia`.

[2] NamrataThakur, *Namratathakur/siim-pcr-pneumothorax-segmentation: This repository contains the image classification followed by semantic segmentation of chest x-rays to detect a clinical condition called pneumothorax.* [Online]. Available: `https://github.com/NamrataThakur/SIIM-PCR-Pneumothorax-Segmentation`.

[3] S. Lal, S. U. Rehman, J. H. Shah, T. Meraj, H. T. Rauf, R. Damaševičius, M. A. Mohammed and K. H. Abdulkareem, 'Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition,' *Sensors*, vol. 21, no. 11, p. 3922, 2021.

[4] H. T. Rauf, M. I. U. Lali, M. A. Khan, S. Kadry, H. Alolaiyan, A. Razaq and R. Irfan, 'Time series forecasting of covid-19 transmission in asia pacific countries using deep neural networks,' *Personal and Ubiquitous Computing*, pp. 1–18, 2021.

[5] 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*, Ieee, 2009, pp. 248–255.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, 'Attention is all you need,' in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[7] M. Ott, S. Edunov, D. Grangier and M. Auli, 'Scaling neural machine translation,' *arXiv preprint arXiv:1806.00187*, 2018.

[8] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding,' *arXiv preprint arXiv:1810.04805*, 2018.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, 'An image is worth 16x16 words: Transformers for image recognition at scale,' *arXiv preprint arXiv:2010.11929*, 2020.

[10] P. A. Larson, L. L. Berland, B. Griffith, C. E. Kahn Jr and L. A. Liebscher, 'Actionable findings and the role of it support: Report of the acr actionable reporting work group,' *Journal of the American College of Radiology*, vol. 11, no. 6, pp. 552–558, 2014.

[11] S. J. Baccei, C. DiRoberto, J. Greene and M. P. Rosen, 'Improving communication of actionable findings in radiology imaging studies and procedures using an emr-independent system,' *Journal of medical systems*, vol. 43, no. 2, pp. 1–6, 2019.

[12] T. S. Cook, D. Lalevic, C. Sloan, S. C. Chadalavada, C. P. Langlotz, M. D. Schnall and H. M. Zafar, 'Implementation of an automated radiology recommendation-tracking engine for abdominal imaging findings of possible cancer,' *Journal of the American College of Radiology*, vol. 14, no. 5, pp. 629–636, 2017.

[13] E. Pons, L. M. Braun, M. M. Hunink and J. A. Kors, 'Natural language processing in radiology: A systematic review,' *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[14] Y. Nakamura, S. Hanaoka, Y. Nomura, T. Nakao, S. Miki, T. Watadani, T. Yoshikawa, N. Hayashi and O. Abe, 'Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers,' *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–19, 2021.

[15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jégou, 'Training data-efficient image transformers & distillation through attention,' in *International Conference on Machine Learning*, PMLR, 2021, pp. 10 347–10 357.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, 'End-to-end object detection with transformers,' in *European Conference on Computer Vision*, Springer, 2020, pp. 213–229.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai, 'Deformable detr: Deformable transformers for end-to-end object detection,' *arXiv preprint arXiv:2010.04159*, 2020.

[18] L. Ye, M. Rochan, Z. Liu and Y. Wang, 'Cross-modal self-attention network for referring image segmentation,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 502–10 511.

[19] S. Chaudhari, V. Mithal, G. Polatkan and R. Ramanath, 'An attentive survey of attention models,' *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021.

[20] A. d. S. Correia and E. L. Colombini, 'Attention, please! a survey of neural attention models in deep learning,' *arXiv preprint arXiv:2103.16775*, 2021.

[21] Y. Bengio, I. Goodfellow and A. Courville, *Deep learning*. MIT press Massachusetts, USA: 2017, vol. 1.

[22] S. Hochreiter and J. Schmidhuber, 'Long short-term memory,' *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] J. C. M. Than, P. L. Thon, O. M. Rijal, R. M. Kassim, A. Yunus, N. M. Noor and P. Then, 'Preliminary study on patch sizes in vision transformers (vit) for covid-19 and diseased lungs classification,' in *2021 IEEE National Biomedical Engineering Conference (NBEC)*, 2021, pp. 146–150. DOI: `10.1109/NBEC53282.2021.9618751`.

[24] H.-C. Shin, A. Ihsani, S. Mandava, S. T. Sreenivas, C. Forster, J. Cha and A. D. N. Initiative, 'Ganbert: Generative adversarial networks with bidirectional encoder representations from transformers for mri to pet synthesis,' *arXiv preprint arXiv:2008.04393*, 2020.

[25] O. Dalmaz, M. Yurt and T. Çukur, 'Resvit: Residual vision transformers for multi-modal medical image synthesis,' *arXiv preprint arXiv:2106.16031*, 2021.

[26] S. Mehta, X. Lu, D. Weaver, J. G. Elmore, H. Hajishirzi and L. Shapiro, 'Hatnet: An end-to-end holistic attention network for diagnosis of breast biopsy images,' *arXiv preprint arXiv:2007.13007*, 2020.

[27] K. Doshi, *Transformers explained visually (part 2): How it works, step-by-step*, Jun. 2021. [Online]. Available: `https://towardsdatascience.com/transformers-explained-visually-part-2-how-it-works-step-by-step-b49fa4a64f34`.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, 'Swin transformer: Hierarchical vision transformer using shifted windows,' *arXiv preprint arXiv:2103.14030*, 2021.

[29] S. Park, *Swin transformers: The most powerful tool in computer vision*, Nov. 2021. [Online]. Available: `https://sieunpark77.medium.com/swin-transformers-the-most-powerful-tool-in-computer-vision-659f78744871`.

[30] Abhishekdhule, *Pneumonia detection(resnet amp;inception)tensorflow*, Aug. 2020. [Online]. Available: `https://www.kaggle.com/abhishekdhule/pneumonia-detection-resnet-inception-tensorflow/notebook`.

[31] *Binary classification*, May 2021. [Online]. Available: `https://en.wikipedia.org/wiki/Binary_classification`.

[32] M. Tan and Q. Le, 'Efficientnet: Rethinking model scaling for convolutional neural networks,' in *International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114.

[33] A. Krizhevsky, I. Sutskever and G. E. Hinton, 'Imagenet classification with deep convolutional neural networks,' *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[34] K. Team, *Keras documentation: Image classification via fine-tuning with efficientnet*. [Online]. Available: `https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/`.