International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 13, Issue 1 (Jan.-Feb. 2024), PP. 11-21

Common Thorax Disease Classification on NHICC Chest X-Ray Database

Ali Murad
Department of Computer Science and Software Engineering
Auburn University
Auburn, AL 36849

Abstract
The recent trend in the healthcare industry indicates the adoption of machine learning methodologies ranging from, CAD – Computer Aided Diagnosis systems to patient demand prediction for timely allocation of resources. Additionally, in healthcare systems including clinics, hospitals, and laboratories, an immense amount of data is collected daily which makes this industry a prime candidate for building innovative solutions to save and improve all life. Through this paper it is shown that with the availability of deep learning tools such applications can be created and can subsequently supplement existing practices. Different neural network architectures are applied including MobileNet-V2, CNN, and two variants of Vision Transformers, B16 and B8 on the ChestX-ray8 Dataset provided by NIHCC – National Institute of Health Clinical Center. Issues such as class imbalance and sparsity are addressed and finally a comprehensive comparative analysis is done on the efficiency of all the methods used.

Date of Submission: 01-01-2024 Date of acceptance: 10-01-2024

I. Introduction

Deep learning methodologies in computer vision have been studied extensively in the healthcare settings. Many of the areas explored in healthcare computer vision include tumor detection, medical imaging, cancer detection, medical training, health monitoring, machine assisted diagnosis, remote patient monitoring and more [1]. It has been shown that such applications can prove to be quite effective when used with knowledge inherited from other subfields of artificial intelligence such as Natural Language Processing [3]. However, widespread adoption of these methodologies by large healthcare companies is still light years behind other industries, largely in part due to the lack of knowledge base of front-line healthcare workers who directly provide care to the patients. This lack of knowledge and understanding further builds a lack of trust in applications which automate patient diagnosis - ultimately causing little adoption. Companies which will be the early adopters of such automated tools and applications will capture market share and dominate the healthcare space by providing better and timely care to their patients.

In this paper the goal is show that using the vast amount of data, that is available in the healthcare industry, predictive applications can be built which facilitate and supplement the workflow of physicians. More specifically multiple predictive models are built on chest x-rays of patients to identify the presence of thoracic diseases. Furthermore, optimization of these models is explored as it pertains to the overall distribution of different classes of diseases being diagnosed and lastly all models are compared and their effectiveness in classifying the different disease classifications is analyzed on a sub-sample of the entire dataset.

1.1 ChestX-ray8 Dataset

The dataset used in this paper comes from NIHCC – The National Institute of Health Clinical Center and is formally known as The ChestX-ray8 dataset [4]. This is one of the largest publicly available datasets on chest x-rays and contains 108,948 images of chest x-rays for 32,717 patients. Working with images is a computationally expensive task, therefore working with available resources and for the purpose of this paper - to demonstrate that predictive tools can be built using open-source technologies - a sub-sample of 6000 images is used while preserving the class distribution.

This is a multi-class and multi-label dataset where in many instances, patients are diagnosed with multiple thoracic diseases. If all the patients are binned into two groups, no finding (presence of no disease) and thoracic disease (presence of disease) then the classes are nearly balanced as shown in the table and the pie chart below.

Binned Class	% Total
No Finding	53.8%
Thoracic Disease	46.2%

Table 1

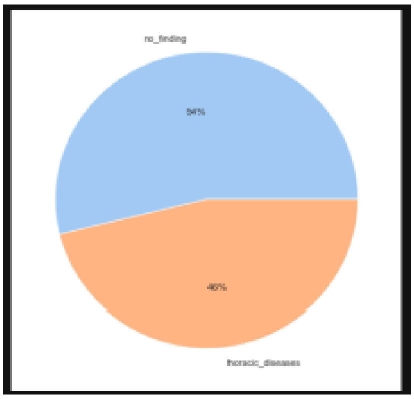


Figure 1 Binned Classes

However, if the dataset is left with the existing class distributions without binning, then there is a presence of 15 diagnoses including no finding. These diagnoses are presented along with their relative class frequencies in the histogram in Figure 2. From the distributions, it can be observed that the dataset is severely class imbalanced and patients without a disease diagnosis constitute most of the dataset. Some of the minority classes in the dataset include hernia, cardiomegaly, edema, emphysema, fibrosis, and pneumonia.

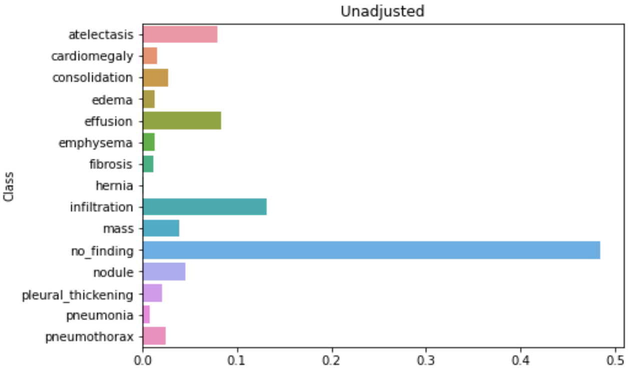


Figure 1 - Actual Class Distributions

1.2 Sample X-ray and Disease Localization

Developing a system to detect the presence of a thoracic disease is a particularly challenging task since many thoracic diseases are localized in certain regions of a person's chest with some diseases also overlapping others. Figure 3 shows a patient diagnosed with Effusion and Infiltration, whereas Figure 4 shows a patient diagnosed with Mass and Pleural Thickening. The presence of tumor in each case can be seen on the lower right side of each x-ray image.

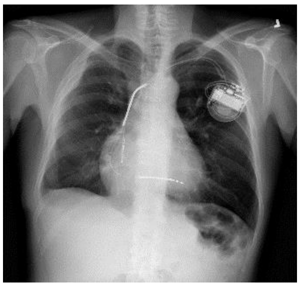


Figure 3 - Effusion and Infiltration Diagnosis

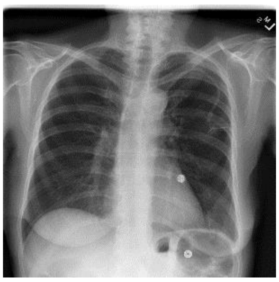


Figure 4 - Mass and Pleural Thickening Diagnosis

The goal of using deep learning methods on this dataset is to address such problems by capturing region specific details from the pixels in the overall image. The original images are of size 1028x1028, however the models described in this paper are built on smaller image dimensions with the highest dimensionality used in building Mobile Net V2 and Convolutional Neural Network models. The different dimensions, image scaling, and number of channels are presented in Appendix I.

II. Related Work

$$Weighted\ Classes = \frac{1}{Actual\ Class\ Weights}$$

Wang et.al [5] conducted a thorough analysis on the ChestX-ray8 dataset. More specifically they created this dataset by first using dependency parsing on the radiological reports associated with each x-ray in the dataset to extract the presence of a disease. The extracted diseases were used to label the x-rays. They further one-hot encoded the labels and used 3 standard loss functions including Hinge Loss, Euclidean Loss, and Cross Entropy Loss. Finally, they used Weighted Cross Entropy Loss to better address the presence of sparsity in the encoded label vectors. To more accurately detect the local structure, in the images, as it pertained to the presence of diseases, they drew bounding boxes around regions which depicted tumors and used the coordinates of these bounding boxes as additional features in their models. The different model architectures used in their research include AlexNet, GoogleNet, VGGNet-16, and ResNet-50. All models were optimized for their performance by maximizing the Area Under the Receiver Operating Curve.

III. Problems Addressed

In this paper, some of the same ideas from Wang et. al [5] are inherited. Sparsity of the labels after one-hot encoding is adjusted by using Weighted Cross Entropy Loss. However, to address the presence of class imbalance an additional step is used where the class weights for each disease are adjusted in the loss function.

a. Sparsity

To solve the issue of sparsity the following Weighted Cross Entropy Loss function is inherited from Wang et. al [5]

$$L_{W-CEL}(f(\vec{x}), \vec{y}) = \\ \beta_P \sum_{y_c=1} - \ln(f(x_c)) + \beta_N \sum_{y_c=0} - \ln(1 - f(x_c)),$$

Figure 5 - Weighted Cross Entropy Loss

In the loss function above B_p is set to $\frac{|P|+|N|}{|P|}$ and B_N is set to $\frac{|P|+|N|}{|N|}$ where |P| and |N| are the total number of '1's and '0's in the batch of image labels used in training.

3.2 Class Imbalance

In addition to adjusting sparsity as described above the loss function is also adjusted by correcting class imbalance. This is done by weighting the loss function by the inverse of the actual class weights as shown below.

These weights are further normalized on a scale [0,1]. Class weights after adjustment of class imbalance are presented in Figure 6 below.

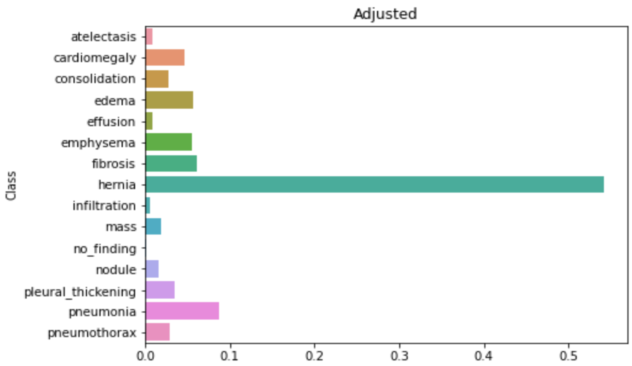


Figure 6 - Adjusted Class Weights

It can be observed that minority classes are weighted the most relative to the majority classes.

4 Infrastructure - Tools

In this research Python is used as the programming language along with NumPy, Pandas, Scikit-Learn and Matplotlib for data manipulation and TensorFlow for feature engineering and modeling.

4.1 Data Pre-processing and Feature Engineering

Each image used for training, validation, and testing was decoded based on specific height and width used to build each model. Mobile Net V2 and CNN were built using image height and width of 299x299, whereas both Vision Transformer models, B16 and B8 were built using image height and width of 244x244. Each image tensor is further augmented by randomly cropping it, adjusting its brightness, flipping the image horizontally, and adjusting its contrast. Parameters for adjustment of brightness and contrast are given in Appendix II. Each image tensor is further scaled based on the model used. For Mobile Net V2 and CNN images are scaled between 0 and 1, whereas for both Vision Transformers, images are scaled between -1 and 1. In addition, labels for each image are one-hot encoded.

5. Machine Learning Framework/Approach

The sub-sample of 6000 images used is split as 4000, 1000, and 1000 images for training, validation, and testing respectively. Each model is trained and validated individually and finally their performance is measured on the test set.

5.1 Target and Baseline Models

Using the sub-sample of the entire dataset, as described in the previous sections, 4 different models are built, Mobile Net V2, CNN, Vision Transformer B8, and Vision Transformer B16. Both Vision Transformer models employ transfer learning and are imported from TensorFlow Hub [6]. Mobile Net V2 is used as the baseline model, and all other models are compared in their performance to it.

5.2 Evaluation Criteria

The objective of this research is to show that Computer Aided Diagnosis systems can be effective is a clinical setting to supplement existing workflows of physicians. Having a system which can accurately diagnose a disease in a given patient is useful, however incorrectly diagnosing a patient that does not have a presence of a disease can have severe negative consequences ranging to unneeded treatments, additional medical exams and tests, and development of further potential complications resulting for taking unneeded medication. Keeping this matter under consideration, it is just as important to control the presence of False Positive as it is to identify True Positives. In order to account for this phenomenon, each model is optimized for Precision, Recall, and Overall Accuracy as shown below.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{FN + TP}$$

$$Accuracy = \frac{TP + TN}{Total\ Samples\ Used}$$

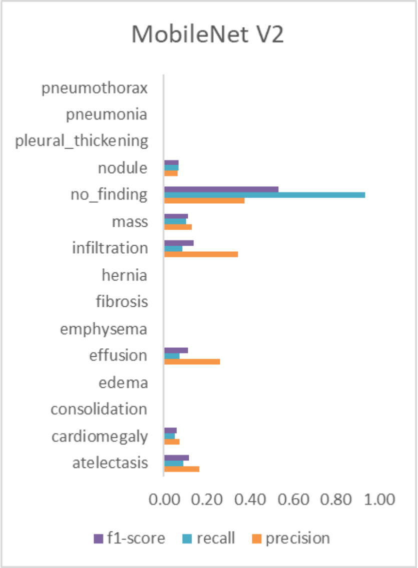
Where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives.

5.3 Model Performance

Model Performance for each model, based on Precision, Recall, and F1-Score, on the test set, along with training and validation Loss and Accuracy, on the training and validation sets, is presented in the sections below. For Precision, Recall, and F1-Score, thresholds for predictions are optimized for each disease classification.

I – Mobile Net – V2

Mobile Net V2 is used as the baseline model to solve this problem. The model is built using RELU activation for the convolutional and dense layers with a dropout rate of 0.15. In the output layer of the model Sigmoid activation is used. Figure 7 shows Precision, Recall and F1-Score for Mobile Net V2 and Figure 8 shows the training and validation loss and accuracy curves.



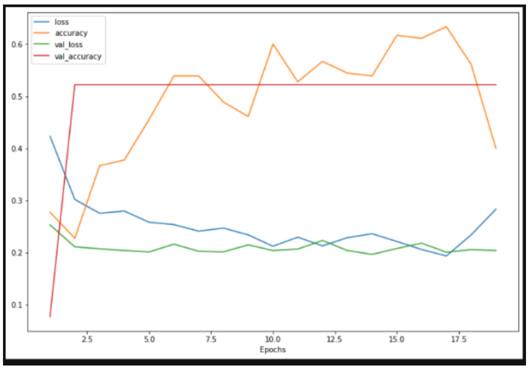


Figure 8 – Mobile Net Loss and Accuracy Curves

Figure 7 - Mobile Net V2 - Precision, Recall, and F1-Score

It can be observed that this model does not perform well on minority classes present in the dataset and tends to classify majority of the images as "no_finding" i.e. presence of no disease.

II – CNN

The CNN model is also built by using RELU activation in the convolutional and dense layers along with Sigmoid activation in the output layer. In addition, a dropout rate of 0.25 is used prior to the output layer. Figure

9 shows Precision, Recall and F1-Score for the CNN model and Figure 10 shows training and validation loss and accuracy curves.

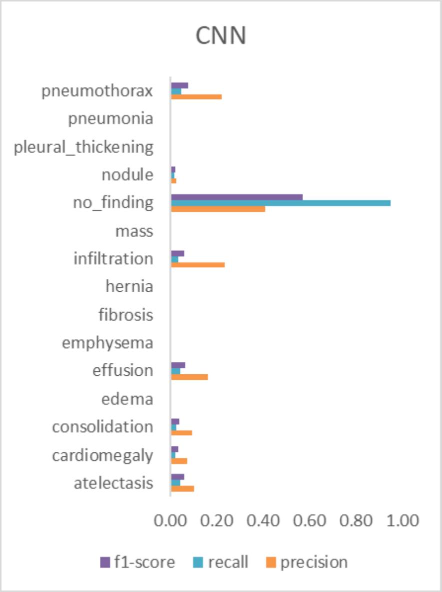


Figure 9 - CNN - Precision, Recall, and F1-Score

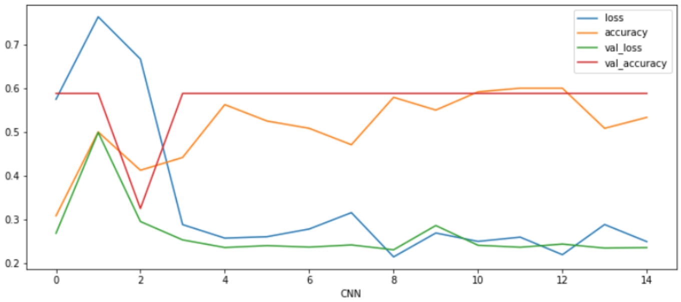


Figure 10 - CNN - Precision, Recall, and F1-Score

The CNN model gave inferior performance compared to the Mobile Net V2 model. Even though it was able to capture more disease classifications, majority of the images were given a majority label of "no_finding".

III – Vision Transformer B16

The Vision Transformer B16 Model is imported from TensorFlow Hub. This model has its parameters fine-tuned on the ImageNet-1k dataset. In addition, the model is given an output layer with Sigmoid activation. Figure 11 shows Precision, Recall and F1-Score for the CNN model and Figure 12 shows the training and validation loss and accuracy curves.

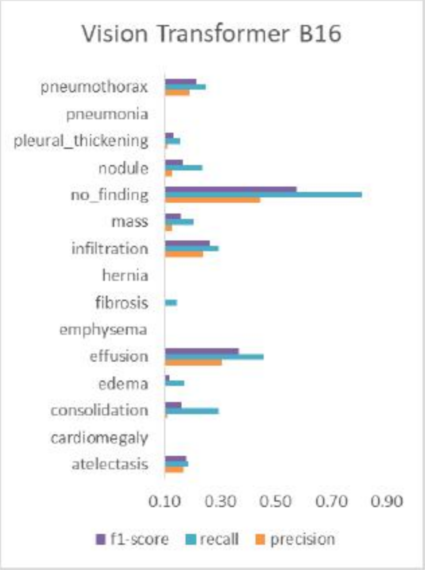


Figure 11 – B16 - Precision, Recall, and F1-Score

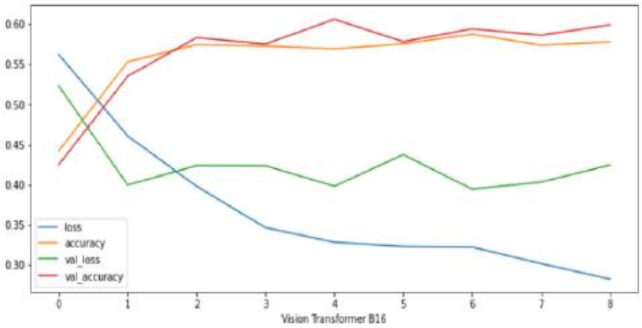


Figure 12 - B16 Loss and Accuracy Curves

The Vision Transformer B16 performed better than both the Mobile Net V2 and the CNN model and was able to capture more disease classes. However, it still suffered from the case of not being able to classify minority classes and over classified images as “no_finding”.

IV – Vision Transformer B8

The Vision Transformer B8 model is also imported from TensorFlow Hub, and its parameters are fine-tuned on the ImageNet-1K dataset as well. In addition, for the output layer Sigmoid activation is used. Figure 13 shows Precision, Recall and F1-Score for the CNN model and Figure 14 shows the training and validation loss and accuracy curves.

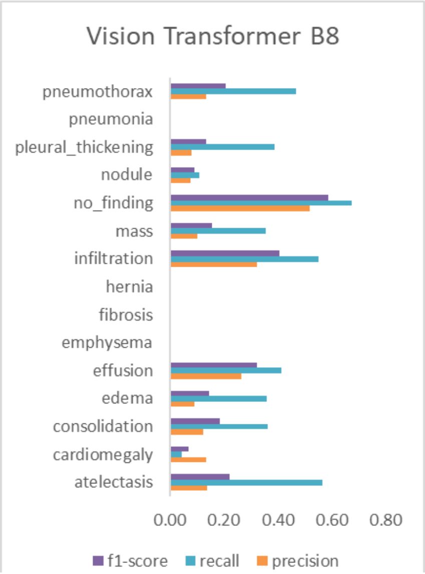


Figure 13 – B8 - Precision, Recall, and F1-Score

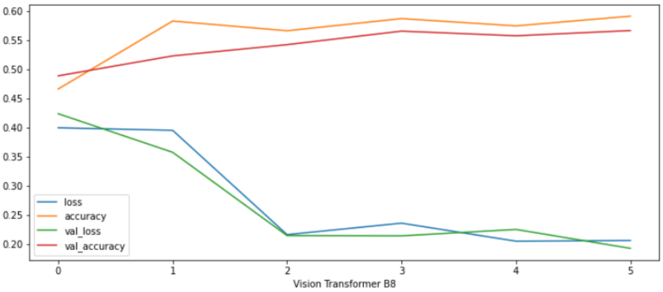


Figure 14 - B8 Loss and Accuracy Curves

This model performed significantly better than the other models and was able to capture significantly more disease classifications. It also did not overly classify images as “no_finding” relative to the other disease classifications, however this model also lacked in classifying the minority classes such as Hernia, Fibrosis, and Pneumonia.

6 Comparative Analysis

Amongst all the model architectures used, the Vision Transformer B8 model performed best across all evaluation metrics including Precision, Recall, and F1-Score. Figure 15 and Table 2 show a weighted average of Precision, Recall, and F1-Score for each model across all disease classes.

Weighted Average	**Precision**	**Recall**	**F1-Score**
MobileNet-V2	0.21	0.28	0.19
CNN	0.20	0.29	0.20
Vision Transformer B16	0.25	0.41	0.31
Vision Transformer B8	0.27	0.47	0.33

Table 1 - Model Comparison

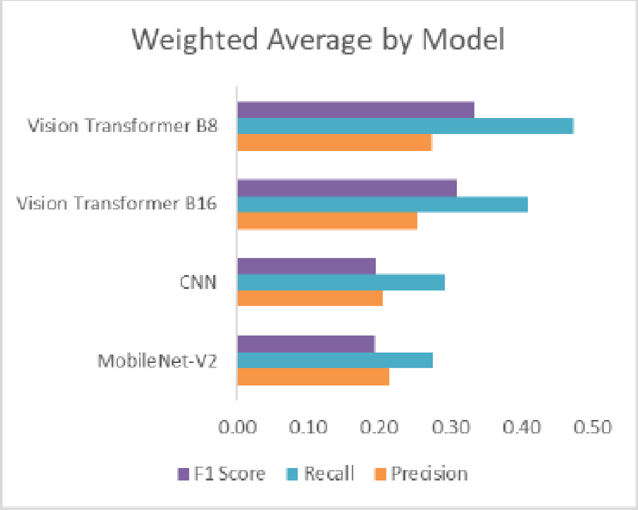
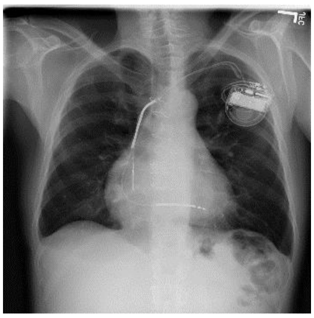


Figure 15 - Model Comparison

Although the vision transformer models outperformed Mobile Net V2 and CNN, their performance is still not optimal for a computer aided diagnosis system. However, it can be observed that transfer learning can be a useful technique in a healthcare setting. An example x-ray using Vision Transformer B8 with actual and predicted labels is shown below.



Predicted label(s):
Atelectasis, Effusion, Infiltration

Actual label(s):
Infiltration

Figure 16 - Model Comparison

Detailed results by each disease class, from all models, are presented as tables in Appendix II.

7 Conclusion and Future Considerations

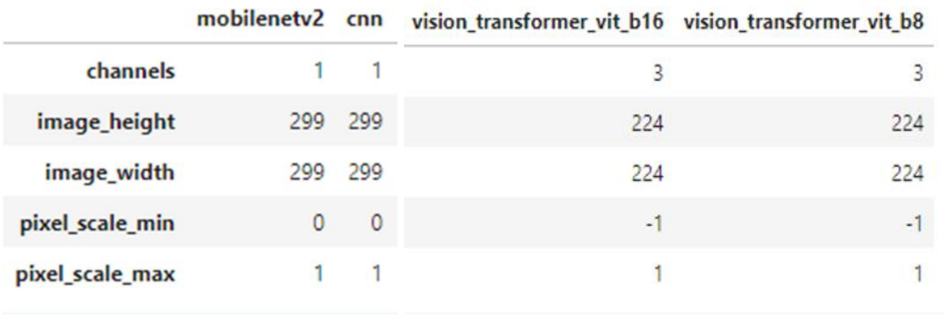
Some of the key takeaways from this research are that image classification is a relatively complex task, and it is even more complex when applied in a healthcare setting, particularly in the case of x-ray images due to the local structure of different diseases. It is also noted that inaccurately diagnosing patients for diseases which they do not have can have unwanted consequences, and therefore during model building it is important to carefully identify which evaluation metrics need to be optimized during training and validation. Choice of a loss function can also have a significant impact of the final outcome as well as the treatment of encoding multi-label outputs. In addition, thresholds used to classify outputs can vary across classes, and need to be optimized to get better predictions.

Although the models used in this paper are built and tested on a sub-sample of the entire dataset, given more computational resources the methods described in this paper can be scaled to the entire dataset. Due to the data hungry nature of deep learning methods, model training and validation on more data can potentially improve the outcome for each model. In addition to large scale training, other loss functions such as Euclidean Loss and Hinge Loss can be used along with varying batch sizes during training. Finally, many different model architectures can be explored in solving this problem including large scale pre-trained image models such as the many variants of Inception and ResNet models.

Through this paper it is shown that using open-source tools the area of computer aided diagnosis can be explored and applications which incorporate models for diagnosing patients can be built. It is also shown that transfer learning can prove to be a useful technique as, on a sub-sample of the dataset, it is able to generalize better relative to training models from the beginning. Building such intelligent and automated systems can prove to be extremely beneficial since it can save time required to diagnose a patient and also provide an additional source of validation for the diagnosing physicians.

Appendix I

Image dimensionality, Scaling, and Channels per model.



	mobilenetv2	cnn	vision_transformer_vit_b16	vision_transformer_vit_b8
channels	1	1	3	3
image_height	299	299	224	224
image_width	299	299	224	224
pixel_scale_min	0	0	-1	-1
pixel_scale_max	1	1	1	1

Appendix II

Data Pre-processing Parameters

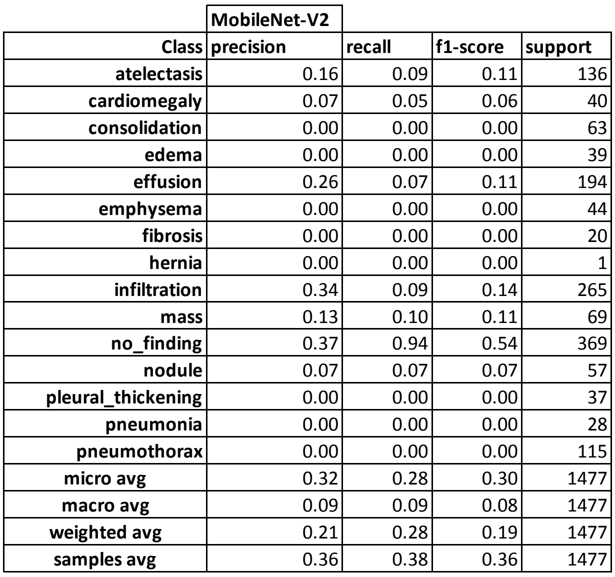
Brightness – Max Delta	63/255
Contrast Lower	0.2
Contrast Upper	1.8

Additional hyper-parameters

Batch Size	15
Shuffle Buffer	20*batch_size
Validation Batch Size	10

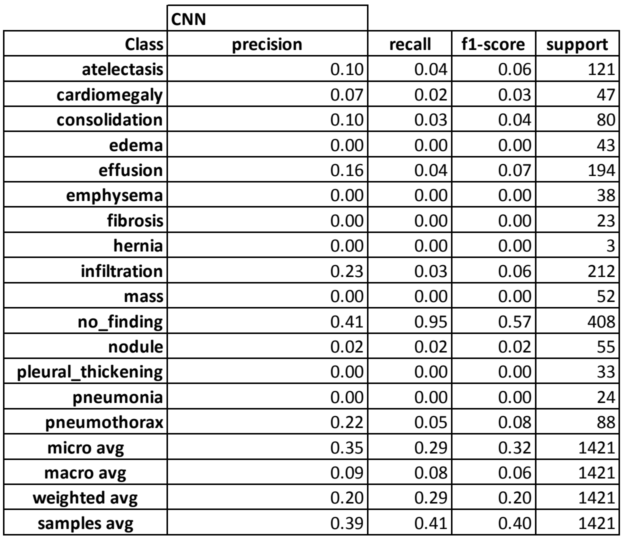
Detailed Model Results

Mobile Net V2



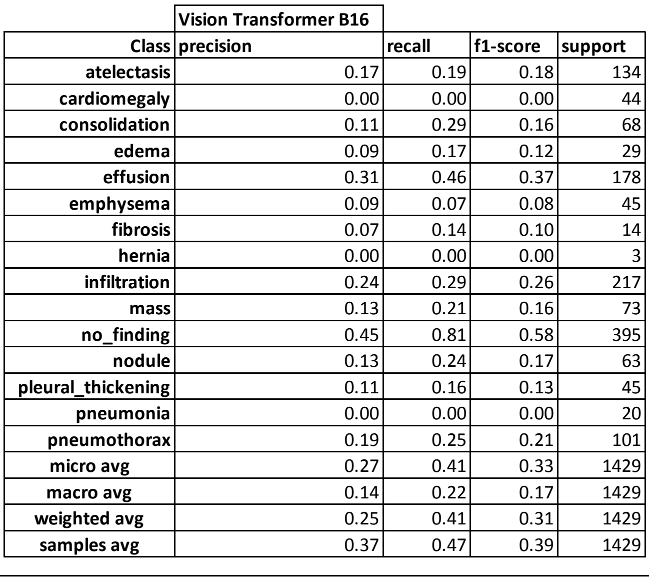
	MobileNet-V2			
Class	precision	recall	f1-score	support
atelectasis	0.16	0.09	0.11	136
cardiomegaly	0.07	0.05	0.06	40
consolidation	0.00	0.00	0.00	63
edema	0.00	0.00	0.00	39
effusion	0.26	0.07	0.11	194
emphysema	0.00	0.00	0.00	44
fibrosis	0.00	0.00	0.00	20
hernia	0.00	0.00	0.00	1
infiltration	0.34	0.09	0.14	265
mass	0.13	0.10	0.11	69
no_finding	0.37	0.94	0.54	369
nodule	0.07	0.07	0.07	57
pleural_thickening	0.00	0.00	0.00	37
pneumonia	0.00	0.00	0.00	28
pneumothorax	0.00	0.00	0.00	115
micro avg	0.32	0.28	0.30	1477
macro avg	0.09	0.09	0.08	1477
weighted avg	0.21	0.28	0.19	1477
samples avg	0.36	0.38	0.36	1477

CNN



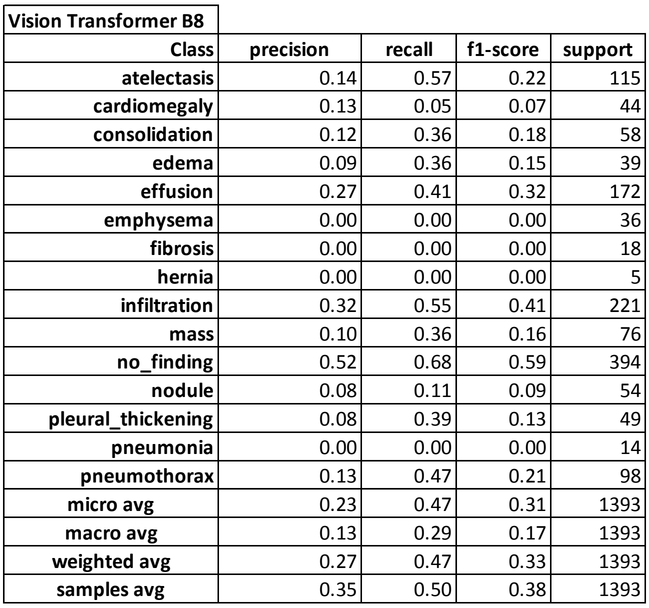
	CNN			
Class	precision	recall	f1-score	support
atelectasis	0.10	0.04	0.06	121
cardiomegaly	0.07	0.02	0.03	47
consolidation	0.10	0.03	0.04	80
edema	0.00	0.00	0.00	43
effusion	0.16	0.04	0.07	194
emphysema	0.00	0.00	0.00	38
fibrosis	0.00	0.00	0.00	23
hernia	0.00	0.00	0.00	3
infiltration	0.23	0.03	0.06	212
mass	0.00	0.00	0.00	52
no_finding	0.41	0.95	0.57	408
nodule	0.02	0.02	0.02	55
pleural_thickening	0.00	0.00	0.00	33
pneumonia	0.00	0.00	0.00	24
pneumothorax	0.22	0.05	0.08	88
micro avg	0.35	0.29	0.32	1421
macro avg	0.09	0.08	0.06	1421
weighted avg	0.20	0.29	0.20	1421
samples avg	0.39	0.41	0.40	1421

Vision Transformer B16



	Vision Transformer B16			
Class	precision	recall	f1-score	support
atelectasis	0.17	0.19	0.18	134
cardiomegaly	0.00	0.00	0.00	44
consolidation	0.11	0.29	0.16	68
edema	0.09	0.17	0.12	29
effusion	0.31	0.46	0.37	178
emphysema	0.09	0.07	0.08	45
fibrosis	0.07	0.14	0.10	14
hernia	0.00	0.00	0.00	3
infiltration	0.24	0.29	0.26	217
mass	0.13	0.21	0.16	73
no_finding	0.45	0.81	0.58	395
nodule	0.13	0.24	0.17	63
pleural_thickening	0.11	0.16	0.13	45
pneumonia	0.00	0.00	0.00	20
pneumothorax	0.19	0.25	0.21	101
micro avg	0.27	0.41	0.33	1429
macro avg	0.14	0.22	0.17	1429
weighted avg	0.25	0.41	0.31	1429
samples avg	0.37	0.47	0.39	1429

Vision Transformer B8



Vision Transformer B8				
Class	**precision**	**recall**	**f1-score**	**support**
atelectasis	0.14	0.57	0.22	115
cardiomegaly	0.13	0.05	0.07	44
consolidation	0.12	0.36	0.18	58
edema	0.09	0.36	0.15	39
effusion	0.27	0.41	0.32	172
emphysema	0.00	0.00	0.00	36
fibrosis	0.00	0.00	0.00	18
hernia	0.00	0.00	0.00	5
infiltration	0.32	0.55	0.41	221
mass	0.10	0.36	0.16	76
no_finding	0.52	0.68	0.59	394
nodule	0.08	0.11	0.09	54
pleural_thickening	0.08	0.39	0.13	49
pneumonia	0.00	0.00	0.00	14
pneumothorax	0.13	0.47	0.21	98
micro avg	0.23	0.47	0.31	1393
macro avg	0.13	0.29	0.17	1393
weighted avg	0.27	0.47	0.33	1393
samples avg	0.35	0.50	0.38	1393

References

[1]. viso.ai. 2021. Top 10 Applications of Deep Learning and Computer Vision in Healthcare - viso.ai. [online] Available at: <https://viso.ai/applications/computer-vision-in-healthcare/> [Accessed 6 December 2021].

[2]. DataToBiz. 2021. 11 Really Helpful Use Cases of Computer Vision in Medicine. [online] Available at: <https://www.datatobiz.com/blog/computer-vision-in-medicine-use-cases/> [Accessed 6 December 2021].

[3]. StartUs Insights. 2021. 5 Top Computer Vision Startups Impacting the Healthcare Industry. [online] Available at: <https://www.startus-insights.com/innovators-guide/5-top-computer-vision-startups-impacting-the-healthcare-industry/> [Accessed 6 December 2021].

[4]. Nihcc.app.box.com. 2021. Box. [online] Available at: <https://nihcc.app.box.com/v/ChestXray-NIHCC> [Accessed 6 December 2021].

[5]. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).

[6]. Tfhub.dev.2021. TensorFlow Hub [online] Available at:<https://tfhub.dev/sayakpaul/collections/vision_transformer/1> [Accessed 6 December 2021].