Pneumonia Detection using Chest X-Ray Images

Hemanth Kumar (19BM6JP27), Karthikeya Racharla (19BM6JP32)
Subramania Bharathi (19BM6JP36)
Post Graduate Diploma in Business Analytics (2019-21)
Indian Institute of Technology
Kharagpur, WB 721 302

Abstract

This study proposes the AI framework for diagnosis of pediatric pneumonia using chest X-ray images. We constructed a convolutional neural network model from scratch to extract features from a given chest X-ray image and classify it to determine if a child is infected with pneumonia. Unlike other deep learning classification tasks with sufficient image repository, it is difficult to obtain a large amount of data for this classification task; therefore, we deployed data augmentation algorithms and used Transfer Learning approaches to improve the validation and classification accuracy of the CNN model and achieved remarkable Recall (99.23%) and F1 Score (92.53%) under different Classification setup, that are currently better and robust than State-of-the-Art results available. Our study aims to ultimately aid in expediting the diagnosis and referral of these treatable conditions, thereby facilitating earlier diagnosis, resulting in improved clinical outcomes.

1 Introduction

According to the reports given by Centers for Disease Control and Prevention, more than 1 million adults are hospitalized with pneumonia and around 50,000 die from the disease every year in the US alone. Chest X-rays are currently the best available method for diagnosing pneumonia as stated by WHO (2001), playing a crucial role in clinical care and epidemiological studies. However, detecting pneumonia in chest X-rays is a challenging task that relies on the availability of expert radiologists. In this work, we present a model that can automatically detect pneumonia from chest X-rays.

Detecting pneumonia in chest radiography can be difficult for radiologists. The appearance of pneumonia in X-ray images is often vague, can overlap with other diagnoses, and can mimic many other benign abnormalities. These discrepancies cause considerable variability among radiologists in the diagnosis of pneumonia.

1.1 Related Works

Kermany et. al (2018) have developed an Image-based deep learning classifies macular degeneration and diabetic retinopathy using retinal optical coherence tomography images, that is generalized to have potential for applications in biomedical image interpretation and medical decision making. They achieved an accuracy of 92.8%, with a Sensitivity of 93.2% and a Specificity of 90.1%

Pranav Rajpurkar et. al (2017) have implemented an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Called as CheXNet, is a

Submitted as Course Project for BM69006 Data Science Laboratory 2020.

121-layer convolutional neural network trained on ChestX-ray14, using the largest publicly available chest X-ray dataset collected by them for their use-case, which contains over 100,000 frontal-view X-ray images with 14 diseases.

2 Motivation

According to the World Health Organization (WHO), pneumonia kills about 2 million children under 5 years old every year and is consistently estimated as the single leading cause of childhood mortality (Rudan et al., 2008), killing more children than HIV/AIDS, malaria, and measles combined (Adegbola, 2012). The WHO reports that nearly all cases (95%) of new-onset childhood clinical pneumonia occur in developing countries, particularly in Southeast Asia and Africa. Bacterial and viral pathogens are the two leading causes of pneumonia (Mcluckie, 2009) but require very different forms of management.

Bacterial pneumonia requires urgent referral for immediate antibiotic treatment, while viral pneumonia is treated with supportive care. Therefore, accurate and timely diagnosis is most imperative. One key element of diagnosis is radiographic data, since chest X-rays are routinely obtained as standard of care and can help differentiate between different types of pneumonia. However, rapid radiologic interpretation of images is not always available, particularly in the low-resource settings where childhood pneumonia has the highest incidence and highest rates of mortality.

Significance and Impact of our work

- To quote an example, in Africa's 57 nations, a gap of 2.3 million doctors and nurses exists. For these populations, accurate and fast diagnosis means everything. It can guarantee timely access to treatment and save much needed time and money for those already experiencing poverty
- Potential for generalized high-impact application in biomedical imaging
- The techniques used in the projects could be extended for detecting other lung diseases where X-ray based detection techniques are currently used
- Ideal example for the same is the current ongoing research and diagnosis on detecting COVID-19 with Chest X-ray images

3 Data set

The data set we used for our classification tasks is taken from Kaggle, created by Paul Mooney at `https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia` under Creative Commons License with Attribution 4.0. The data set has three sub-divisions, namely Training, Validation and Test sets, with sub-categorization as each image category (Pneumonia/Normal). There are a total of 5,856 X-Ray images (JPEG) and 2 categories, namely Pneumonial vs. Normal images. These pediatric chest X-rays of pneumonial patients is further distinguished into viral and bacterial pneumonia to facilitate rapid referrals for children needing urgent intervention.

Salient features under data collection mechanism:

- Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou
- All chest X-ray imaging was performed as part of patients' routine clinical care. For the analysis of chest x-ray images, all chest radio graphs were initially screened for quality control by removing all low quality or unreadable scans.
- The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system.
- In order to account for any grading errors, the evaluation set was also checked by a third expert.

The data set has original distribution as mentioned above in the table 1.

SET	Total Images	Normal	Pneumonia	Bacterial	Viral
Training	5,216	1,341	3,875	2,530	1,345
Validation	16	8	8	8	0
Testing	624	234	390	242	148

Table 1: Dataset Subdivison

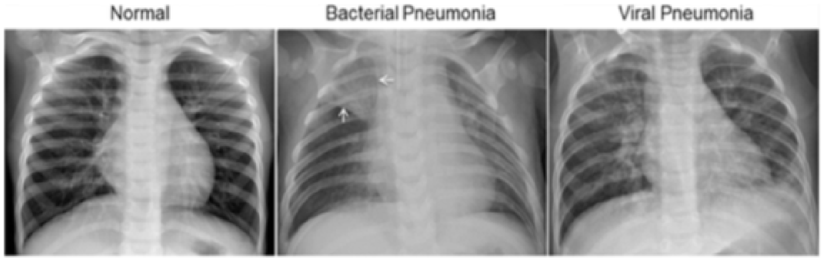

Figure 1: Sample X-Ray Images under each label

As seen in the image 1, the normal chest X-ray (left panel) depicts clear lungs without any areas of abnormal opacification in the image. Bacterial pneumonia (middle) typically exhibits a focal lobar consolidation, in this case in the right upper lobe (white arrows), whereas viral pneumonia (right) manifests with a more diffuse “interstitial” pattern in both lungs.

4 Data Pre-Processing

4.1 Data Augmentation

Data Augmentation is a strategy that enables practitioners to significantly increase the diversity of data available for training models by cropping, padding, and horizontal flipping the training data, without actually collecting new data. We have done it using Image Data Generator class from Keras.

The transformations we have used in our work includes: *Zoom, Horizontal shift, Vertical shift, Horizontal flip, Rotation and shear.*

5 Evaluation Criteria

5.1 From Confusion Matrix

Confusion Matrix evaluates the performance of a supervised classifier using a cross-tabulation of actual and predicted classes. The following evaluation metrics are be obtained from the Confusion Matrix, along with Accuracy. We used these to judge the performance of our models.

- Precision is the ratio $\frac{tp}{(tp+fp)}$, where tp is the number of true positives and fp the number of false positives. Precision intuitively describes the ability of the classifier not to label a false positive as positive.
- Recall is the ratio $\frac{tp}{(tp+fn)}$ where tp is the number of true positives and fn the number of false negatives. Recall is intuitively the ability of the classifier to identify all the positive samples. Figures 9 and 16 shows illustrative visualization of Precision and Recall for various supervised classifiers implemented.
- F1 score can be interpreted as the harmonic mean of Precision and Recall.

$$F1 = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

5.2 NLL/Cross-Entropy Loss

The most common loss function used in deep learning networks is Cross-Entropy or in other words, called as Negative Log-Likelihood Loss. It is defined as -

$$E_{\text{entropy}} = -\sum_{1}^{n}\sum_{1}^{m} y_{i,j}\ln(p_{i,j}) \tag{1}$$

where, $\mathrm{y_{i,j}}$ denotes the true value i.e. 1 if sample i belongs to class j and 0 otherwise, and $(\mathrm{p_{i,j}})$ denotes the probability predicted by the model of sample i belonging to class j.

6 Methodology/Line of Work

The following steps give a broad overview of our work. Complete set of codes can be found in the Repository link: `https://github.com/KarthikeyaR/pneumonia-detection`. We have performed 2 kinds of classification - investigating the effectiveness of our CNN and transfer learning architectures in classifying these pediatric chest X-rays to detect Pneumonia vs. Normal Images and furthermore to distinguish viral and bacterial pneumonia.

For both these tasks, we have performed the steps mentioned below, in brief.

- Design Convolutional Neural Network Model and evaluation of results
- Use Transfer Learning Approaches for the classification and evaluate various model results
- Comparison of Results and selecting the best model based on F1-Score

7 Convolutional Neural Networks

CNNs are proved to have an edge over traditional Neural Networks, where they possess a visual processing scheme that is equivalent to that of humans and extremely optimized structure for handling images and 2D and 3D shapes, as well as ability to extract abstract 2D features through learning. The max-pooling layer of the convolutional neural network is effective in variant shape absorptions and comprises sparse connections in conjunction with tied weights.

When compared with fully connected (FC) networks of equivalent size, CNNs have a considerably smaller amount of parameters. Since the gradient-based algorithm is responsible for training the whole network in order to directly diminish an error criterion, highly optimized weights can be produced by CNNs. Apart from the fact that these are computationally expensive, such algorithms like CNN are data-hungry, as in they require huge training sample.

7.1 CNN Model Architecture

The figure 2 explains the architecture of the CNN used. The model has 6,638,753 trainable parameters.

7.2 Description Of Architecture

7.2.1 Separable convolution layer

This is a variation of traditional convolution.This performs a depth-wise spatial convolution followed by a point-wise convolution which mixes together the resulting output channels. Separable convolution computationally faster than the traditional convolution. Example Mobile net uses separable convolution layer.

7.2.2 Padding

Padding is done with zero values such that dimension remains the same as the image dimension after each convolution.

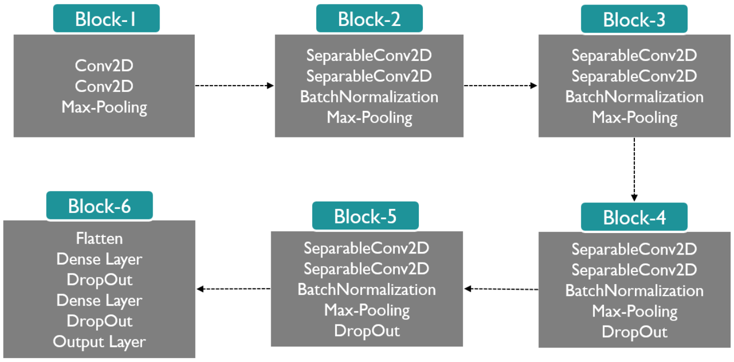

Figure 2: Brief Architecture of CNN Model

7.2.3 Batch normalisation

Standardisation of data is performed as the data passes through each of the layers. So, there will be less oscillation in cost function. And the optimal solution could be reached faster, by adjusting for the batch sample fluctuations.

7.2.4 Max pooling

This is used to reduce the amount of parameters and computation in the network.

7.2.5 Dropout

This layer is used for regularisation, so that model can be generalizable. Although using more dropout probability makes model train slow, it avoids over-fitting by making the nodes avoid dependency on specific nodes.

7.2.6 Variable learning rate

When there is no improvement in the evaluation metric, learning rate will be decreased.

7.2.7 Early stopping

When the generalization gap (i.e. the difference between training and validation error) starts to increase, instead of decreasing training is stopped.

8 Transfer Learning

One method of addressing a lack of data in a given domain is to leverage data from a similar domain, a technique known as transfer learning. Transfer learning has proven to be a highly effective technique, particularly when faced with domains with limited data. Rather than training a completely blank network, by using a feed-forward approach to fix the weights in the lower levels already optimized to recognize the structures found in images in general and retraining the weights of the upper levels with back propagation, the model can recognize the distinguishing features of a specific category of images, like fine fabric-like structures in X-Ray Images, much faster and with significantly fewer training examples and less computational power.

8.1 Brief Description of Architectures

8.1.1 ResNet50

The core idea of ResNet is introducing a so-called “identity shortcut connection” that skips one or more layers. The authors of ResNet argue that stacking layers shouldn’t degrade the network performance, because we could simply stack identity mappings (layer that doesn’t do anything) upon the current network, and the resulting architecture would perform the same. This indicates that the deeper model should not produce a training error higher than its shallower counterparts. They hypothesize that letting the stacked layers fit a residual mapping is easier than letting them directly fit the desired under laying mapping. Also, the residual block explicitly allows it to do precisely that.

8.1.2 DenseNet

DenseNet(Densely Connected Convolutional Networks) is one of the latest neural networks for visual object recognition. It's quite similar to ResNet but has some fundamental differences. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters.

8.1.3 InceptionV3

The Inception Module is based on a pattern recognition network which mimics the animal visual cortex. After presenting several examples of images, the network gets used to small details, middle sized features or almost whole images if they come up very often. Using the Tensorflow we adapted the InceptionV3 architecture that is pre-trained on the ImageNet data set (Szegedy et al., 2016). Retraining consisted of initializing the convolutional layers with loaded pre-trained weights and retraining the final, softmax layer to recognize our classes from scratch. Briefly, this model attempts at 'fine-tuning' the convolutional layers by unfreezing and updating the pre-trained weights on our medical images using back-propagation tended to decrease model performance due to over-fitting.

8.1.4 VGG

It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 33 kernel-sized filters one after another. This model architecture has achieved 92.7% top-5 test accuracy on ImageNet data set which contains 14 million images belonging to 1000 classes and secured top among contenders in 2014 ILSVRC Visual Recognition Challenge.

9 Results

9.1 Pneumonia vs. Normal Classification

The models are trained until saturation. The table 2 explains the model performance. Plots for Training and Validation Losses by different model tryouts are in Figures 3 to 7. Similarly, plots for Accuracy, F1-Score, Precision and Recall are in figures 8 and 9.

Models	Precision	Recall	Accuracy	F1 Score	AUC
Resnet50	92.95	95.64	92.95	94.43	92
Densenet	90.77	95.89	91.34	93.26	90
CNN	89.76	96.66	91.03	93.08	91
VGG16	87.55	99.23	90.71	93.02	88
InceptionV3	87.27	98.46	90.06	92.53	87

Table 2: Comparison of Model Results for Pneumonia Detection

9.2 Bacterial vs. Viral Classification

The models are trained until saturation. The table 3 explains the model performance. Plots for Training and Validation Losses and accuracy by different model tryouts are in Figures 10 to 14. Similarly, plots for Accuracy, F1-Score, Precision and Recall are in figures 15 and 16.

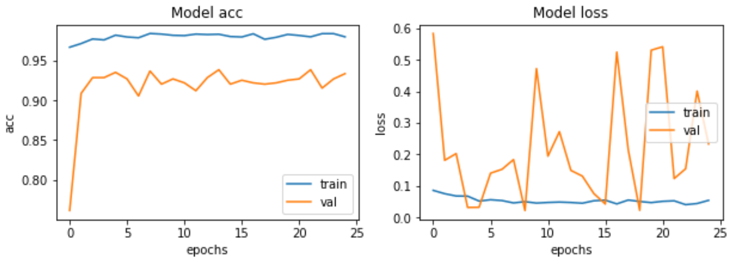

Figure 3: ResNet50 Model Accuracy and Losses (Pneumonia)

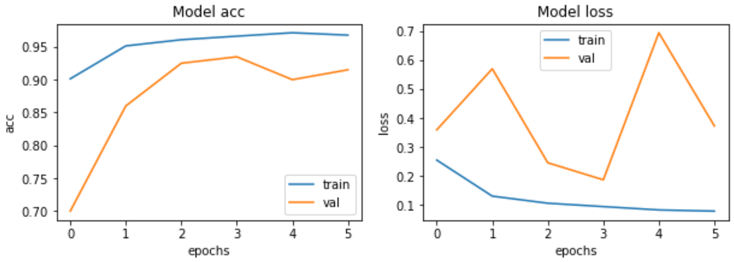

Figure 4: DenseNet Model Accuracy and Losses (Pneumonia)

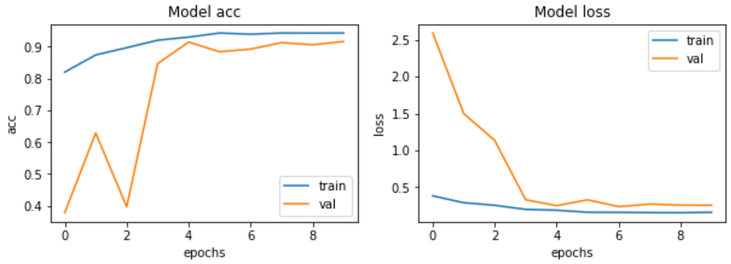

Figure 5: CNN Model Accuracy and Losses (Pneumonia)

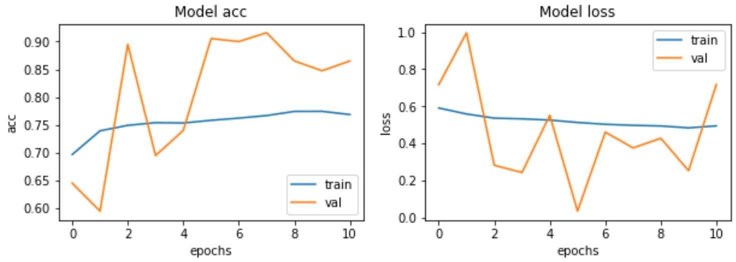

Figure 6: VGG16 Model Accuracy and Losses (Pneumonia)

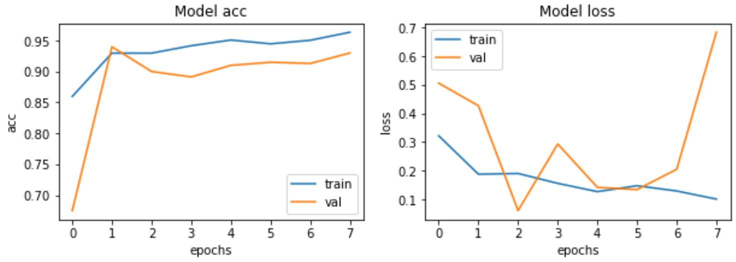

Figure 7: InceptionV3 Model Accuracy and Losses (Pneumonia)

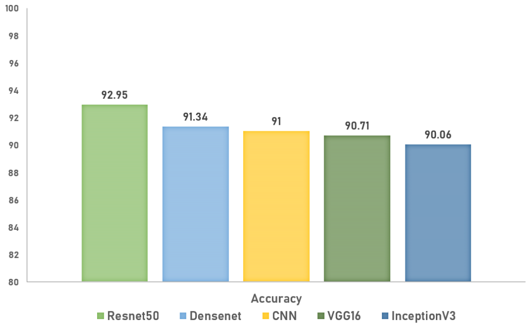

Figure 8: Accuracy of different models (Pneumonia)

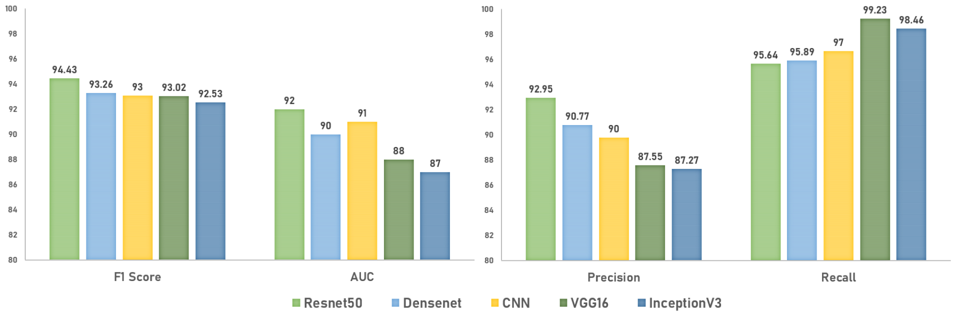

Figure 9: F1-Score, AUC, Precision and Recall for different models (Pneumonia)

Models	**Precision**	**Recall**	**Accuracy**	**F1 Score**	**AUC**
InceptionV3	87.27	98.46	90.06	92.53	0.87
Resnet50	86.98	85.81	89.74	86.39	0.89
CNN	94.74	72.97	88.21	82.44	0.85
VGG16	100	66.89	87.43	80.16	0.83
Densenet	89.47	57.43	81.28	69.96	0.77

Table 3: Comparison of Model Results for Bacterial vs. Viral Pneumonia Detection

9.3 Conclusion

Immediate treatment for someone with pneumonia is of utmost importance, so everyone with pneumonia has to be found out correctly with the model, i.e., in evaluation terms of Machine Learning, Recall has to be high for Pneumonia detection, i.e., Pneumonia vs Normal cases.

Once some one is diagnosed with Pneumonia, it's important to identify whether it is a viral pneumonia case or bacterial pneumonia, as the patient has to be treated accordingly because mistreatment as a viral case could be fatal. Moreover, the treatments for both are different. Hence, Precision is very important for classifying types of Pneumonia into Bacterial or Viral.

From our results, we clearly see that our transfer learning models have performed better than our conventional CNN model. For Normal vs Pneumonia case - all our transfer learning implementations have performed remarkably well (ResNet50 being more stable in terms of Model loss and Accuracy) and has produced Recall better than the benchmark score (93.2%); **VGG16** (Recall: 99.23%) being the champion model. For viral and bacterial case - Our models have produced better F1 scores and Precision than the benchmarks of F1 (62.36%) and Precision(48.11%); **InceptionV3** (F1-Score 92.53%) being the champion model.

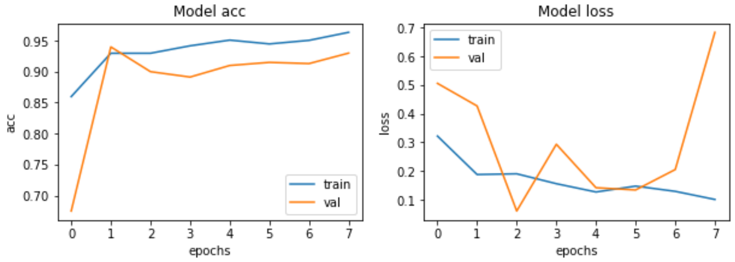

Figure 10: InceptionV3 Model Accuracy and Losses (Bacterial vs. Viral)

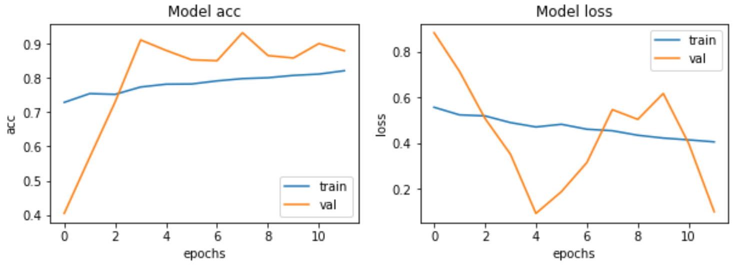

Figure 11: ResNet50 Model Accuracy and Losses (Bacterial vs. Viral)

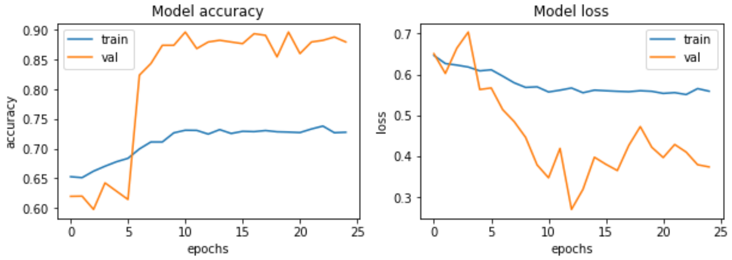

Figure 12: CNN Model Accuracy and Losses (Bacterial vs. Viral)

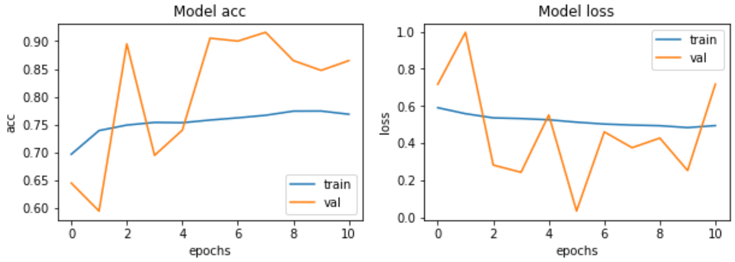

Figure 13: VGG16 Model Accuracy and Losses (Bacterial vs. Viral)

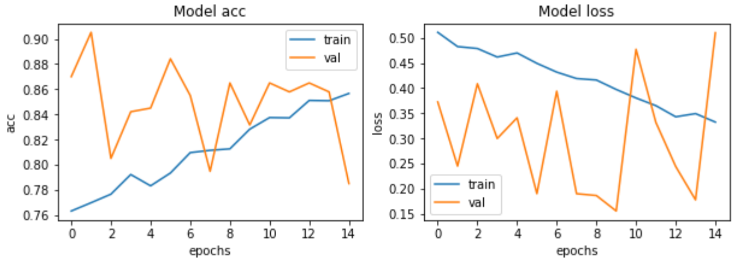

Figure 14: DenseNet Model Accuracy and Losses (Bacterial vs. Viral)

This also means that our model is more generalizable and robust than the existing architectures. We also showcased the power of the transfer learning system to make highly effective classifications, even with a very limited training data set.

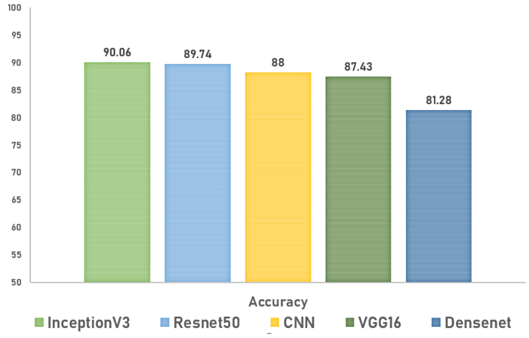

Figure 15: Accuracy of different models for Bacterial vs. Viral Classification

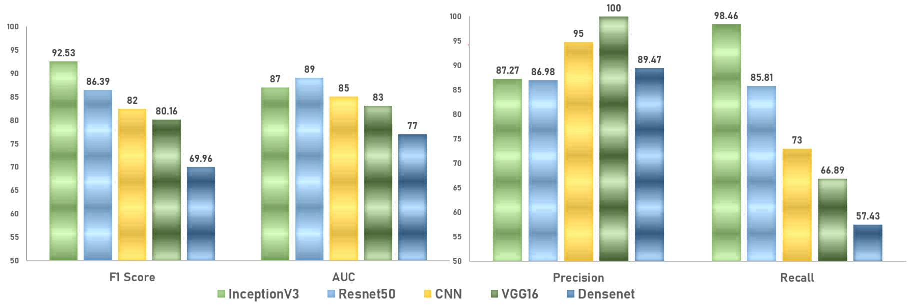

Figure 16: F1-Score, AUC, Precision and Recall for different models (Bacterial vs. Viral)

10 Future Scope

Our study will go on a long way in improving the health of at-risk children in energy-poor environments. As our analysis was limited by depth of data, increased access to data and training of the model with radiological data from patients and non-patients in different geography can create significant improvements in the model. Hence, future studies could entail use of images from varied manufacturers, so that the system will be universally useful and acceptable.

In principle, the techniques we have described here could potentially be extended in a wide range of medical images across multiple disciplines such as Ophthalmology, CT Scans etc. There is also the concept of **Occlusion testing** to identify areas of greatest importance used by model while assigning a diagnosis. The greatest benefit of an occlusion test is that it reveals insights into the decisions of neural networks, which are infamously known as 'black boxes' with no transparency.

Acknowledgments

We would like to thank Prof. Sujoy Bhattacharya for allowing us to pursue such an interesting topic for our course project, guiding us throughout and evaluating the project results.

References

Kermany et. al, Michael Goldbaum, W. C. M. A. L. K. Z. Identifying medical diagnoses and treatable diseases by image-based deep learning. In *Cell 172, 1122–1131*, 2018.

Pranav Rajpurkar et. al, Jeremy Irvin, K. Z. B. Y. A. Y. N. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. 2017.