ChestVision

Jonathan Chen (xc228)
Omkar Garad (omg22)
Jae Kim (jk2765)
Genki Miyamoto (gm565)
Aryan Patil (ap2365)

Abstract

This research investigates the efficacy of computer vision techniques in identifying pleural effusion, cardiomegaly, and pneumonia in chest radiographs using the CheXpert dataset. Given the crucial role of timely diagnosis in emergency scenarios, we undertake preprocessing on a dataset consisting of 224,316 radiographs from 65,240 patients, labeled for 14 chest conditions. The study employs three architectures (DenseNet-121, EfficientNet, and Vision Transformer) to train models, considering both consolidated and individual approaches. In evaluating the consolidated model, EfficientNet demonstrates a superior AUROC mean (0.837) compared to Vision Transformer (ViT) (0.779) and DenseNet (0.796). This exceptional performance leads to the selection of EfficientNet for developing individual disease models, showcasing its versatility and robustness across diverse diagnostic contexts. The significance of this work lies in addressing a critical gap in leveraging computer vision for enhanced emergency diagnostics. The study not only contributes valuable insights into optimal model configurations for medical imaging analysis but also underscores the potential of EfficientNet in advancing disease detection from chest radiographs.

Background

The chest radiograph is an imaging test that looks at the organs near the chest area, in particular heart, lungs, bronchi, aorta, pulmonary arteries, etc. These radiographs are also effective in revealing fluid in or around the lungs, and are crucial for the diagnosis of many diseases and conditions such as collapsed lung, pneumonia, broken ribs, emphysema, cancer, and etc. Among these conditions, pleural effusion, cardiomegaly, and pneumonia are amongst the most common presented in the emergency room. Pleural effusion is an abnormal accumulation of fluid in the pleural space between the lungs and the chest wall. Cardiomegaly refers to an abnormally enlarged heart caused by various different reasons. Pneumonia is inflammation and fluid in the lungs caused by a bacterial, viral, or fungal infection, making it difficult to breathe and sometimes causing a fever or cough with abnormal or bloody mucus. Pleural effusion is treated through drainage via therapeutic thoracentesis or through a test tube (tube thoracostomy), cardiomegaly is treated either through surgery or implantation of devices, such as pacemaker or implantable cardioverter defibrillator. Pneumonia can be treated using antibiotics, but may require oxygen therapy, IV fluids, or drainage if the condition has developed. Early diagnosis of these diseases when patients present themselves to the emergency room leads to better health outcomes, saving more lives and reducing medical costs. The use of computer vision on chest radiograph imaging can be used to screen for pleural effusion, cardiomegaly, and pneumonia in order to facilitate earlier treatment through hastened diagnosis.

The aim of the project is to assess how accurate pleural effusion, cardiomegaly, and pneumonia can be detected from chest radiographs as well as which models best detect effusion, cardiomegaly, and pneumonia from chest radiographs. The approach to address these questions consist of the following steps: finding a large dataset of chest radiographs that are labeled with the particular condition, preprocessing the images through resizing, normalizing, cropping, noise reduction, and other methods, training consolidated models of DenseNet-121, EfficientNet and Vision Transformer architectures, performing a comparative analysis on the consolidated models to determine a model used for binary classification of the three diseases, create and train binary classification models for the three diseases and compare to the consolidated model. The subsequent comparative analysis evaluates the models' performances, aiming to identify optimal configurations for accurate detection. This focused exploration contributes to the advancement of computer vision applications in medical imaging, specifically in enhancing emergency diagnostic capabilities.

Dataset

The dataset that is being used for this project is the CheXpert dataset, a large dataset of chest radiographs collected by the Stanford ML Group. The dataset consists of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest radiographic observations: Enlarged Cardiomediastinal, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices as well as a distinct label “No Findings” if there were no conditions detected in the image. For each image each condition was labeled either negative (“no evidence of that condition"), uncertain (“possibility that the condition is present), or positive (“evidence that the condition is present”). In addition to the labels, the dataset contains the path to the image, demographic information, particularly sex and age, and image information. Image information consisted of whether the image was taken frontal or lateral and whether the radiograph is anterior posterior or posterior anterior.

These labels were determined using a labeler which extracts observations from the impression section of radiology reports that summarizes key findings in the radiographic study. A large list of phrases were manually curated by board-certified radiologists to match various ways observations can be mentioned in these reports. After the observations are extracted, the reports was split and tokenized using NLTK into sentences. Then, these sentences are parsed and a universal dependency graph was computed using CoreNLP for each sentence. Each mention of observations is classified to arrive at a final label. In the dataset, “1” represents a positive observation, “-1” represents an uncertainty label, and “0” represents negative observation. A blank cell represents no mention of the disease entirely.

The original CheXpert containing 224,316 chest radiographs of 65,240 patients is too large to be stored on personal computers, so a downsampled subset of the CheXpert dataset containing 190,527 chest radiograph of 64,540 patients is used instead. This dataset was preprocessed through resizing the images, cropping in order to cut off unwanted letters on certain images, appling contrast limited adaptive histogram equalization, applying gamma correction, normalizing the images, and converting PIL images into tensors.

The dataset preprocessing involved the following 6 steps:

1.) Elimination of Undesired Labels & Images:
All the labels unrelated to the 3 target diseases were removed from the dataset ('Sex', 'Age', 'Frontal/Lateral', 'AP/PA', 'No Finding', 'Enlarged Cardiomediastinum', 'Lung Opacity', 'Lung Lesion', 'Edema', 'Consolidation', 'Atelectasis', 'Pneumothorax', 'Pleural Other', 'Fracture', 'Support Devices'). Also, all the lateral x-ray images were removed since not all the patients had lateral radiographs.

2.) Resizing & Cropping:
To ensure all the images in the dataset have a consistent size, each image was resized to the size (320, 320). Also, each image was cropped at the center with the cropping size of 250 to focus more on the chest area, which is relevant to the task. Some of the images in the dataset included undesired letters in the rims, so these letters were also removed by the cropping.

3.) Adaptive Histogram Equalization (CLAHE):
Adaptive Histogram Equalization (CLAHE) was also applied to enhance the contrast of an image, and make it easier for a model to detect features.

4.) Gamma Correction:
Gamma correction was also applied to adjust and even the brightness of each image.

5.) Normalization:
Each image was also normalized to help the model converge faster during training.

6.) Conversion from PIL Images into Tensors:
Each image was converted from Python Imaging Library (PIL) format into tensors, which are the standard input format for deep learning models in frameworks like PyTorch.

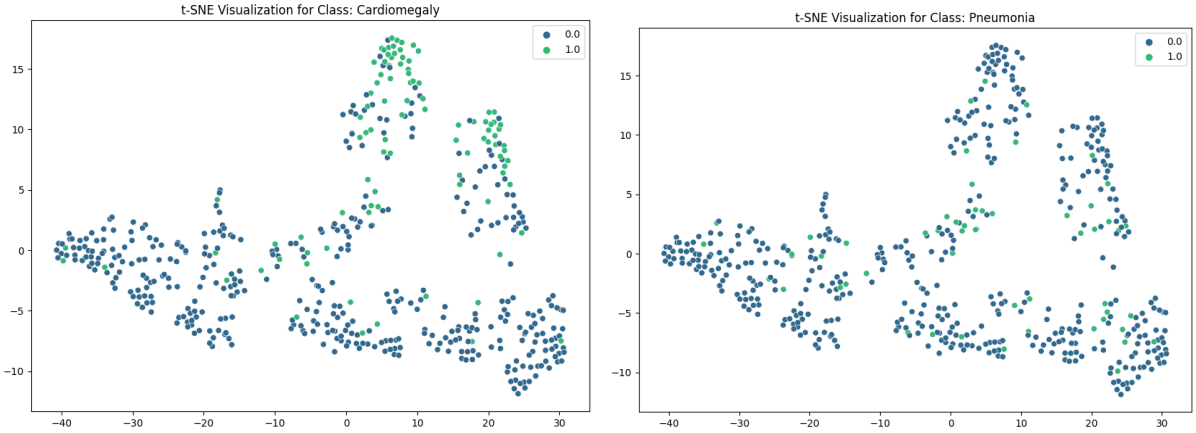
Analysis

The first model that was trained was the DenseNet-121 Architecture. This was the architecture that the authors of the original CheXpert paper had the best results with when the paper was published in 2019. To handle uncertainty labels (-1), the approach of changing all uncertainty labels to 1 was used for this project due to demonstrated success by the creators of the dataset. There are a multitude of other approaches to handling the uncertainty labels, but the purpose of this project is not to determine the best approach.

After the dataset is modified to reflect this approach, preprocessing was conducted using the steps outlined in the dataset section. The training of the model uses the Adam optimizer with specific learning rates and other parameters. The training function performs one training epoch iteration over the training data, computes the binary cross-entropy (BCE) loss, performs backpropagation, and updates the model's parameters. The model is then trained using federated learning, which allows a model to be trained across multiple decentralized devices or servers holding local data samples without exchanging them. The training dataset was divided up into five so that federated learning can be performed using five decentralized devices.

Similarly, EfficientNet and Vision Transformer were trained. EfficientNet was chosen for its superior performance in balancing model efficiency and accuracy, crucial for handling a large medical imaging dataset. Vision Transformer (ViT) was selected to explore attention mechanisms, beneficial for capturing intricate patterns in chest X-ray images and potentially improving diagnostic precision. In order to compare DenseNet-121, EfficientNet and Vision Transformer, 1 epoch per round, 1 round, and 5 clients (federated learning) were used. Then AUROC, t-SNE, and confusion matrix were found to analyze the accuracy of the model. Through comparative analysis of the three consolidated models, the best architecture was used to train three binary classification models for the three diseases. These binary classification models are compared to the consolidated model to analyze the difference in performance between binary classifications models and consolidated models.

Results



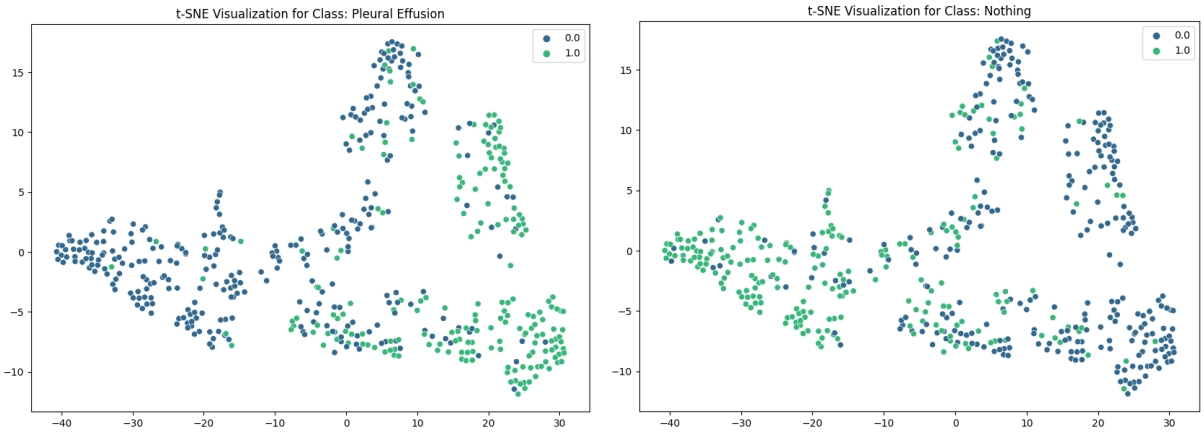


Figure 1. T-SNE graphs for consolidated models

Figure 1 shows the t-SNE graphs for the different classes used in the consolidated models for DenseNet-121, Vision Transformer, and EfficientNet. These graphs visually represent high-dimensional data in a lower-dimensional space. From the clustering on these graphs, it can be hypothesized which classes are more easily classified than others.

AUROC	DenseNet-121	Vision Transformer	EfficientNet
AUROC mean	0.7965	0.7797	0.8375
Cardiomegaly	0.8287	0.8225	0.8450
Pneumonia	0.6687	0.6635	0.7611
Pleural Effusion	0.8649	0.8267	0.8917
None	0.8238	0.8060	0.8521

Table 1: AUROC Comparison: DenseNet vs ViT vs EfficientNet.

Table 1 shows the mean AUROC as well as individual class AUROC results of the 3 trained models. It clearly shows that the consolidated disease detection model using EfficientNet outperforms DenseNet as well as Vision Transformer models.

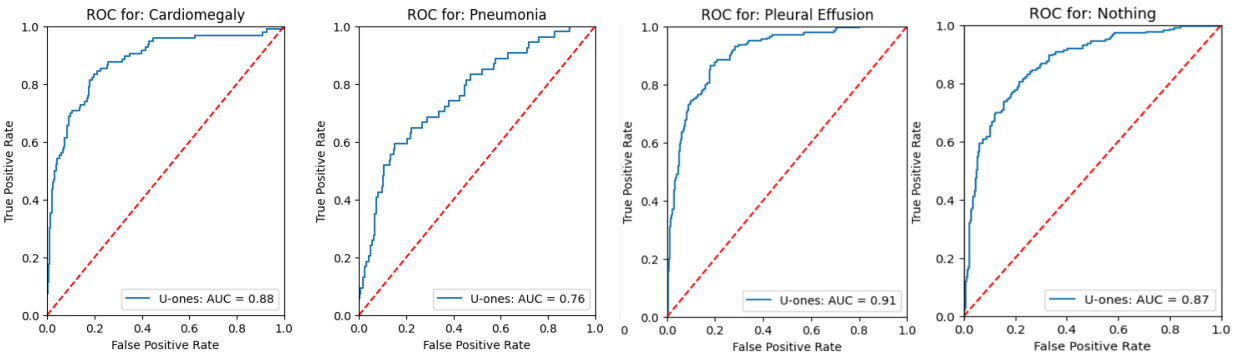


Figure 2: AUROC graphs for EfficientNet

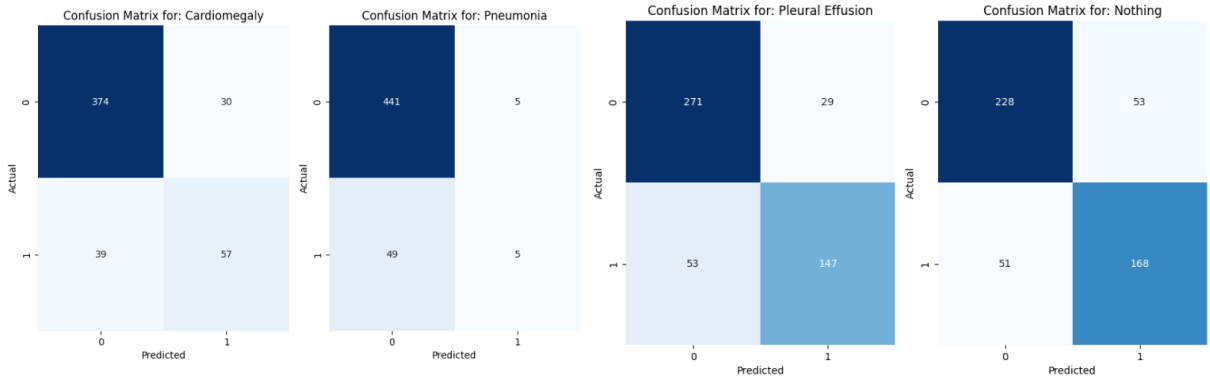


Figure 3: Confusion matrices for EfficientNet

The Vision Transformer model is outperformed by both the DenseNet-121 model and the EfficientNet model. The AUROC graphs and the confusion matrices for EfficientNet, the best performing model, is displayed in figures 2 and 3.

AUROC	Cardiomegaly	Pneumonia	Pleural Effusion
AUROC Mean	0.8916	0.7701	0.9158
Disease	0.8914	0.7699	0.9159
Not Disease	0.8917	0.7702	0.9157

Figure 4. AUROC of EfficientNet binary classifications on individual diseases

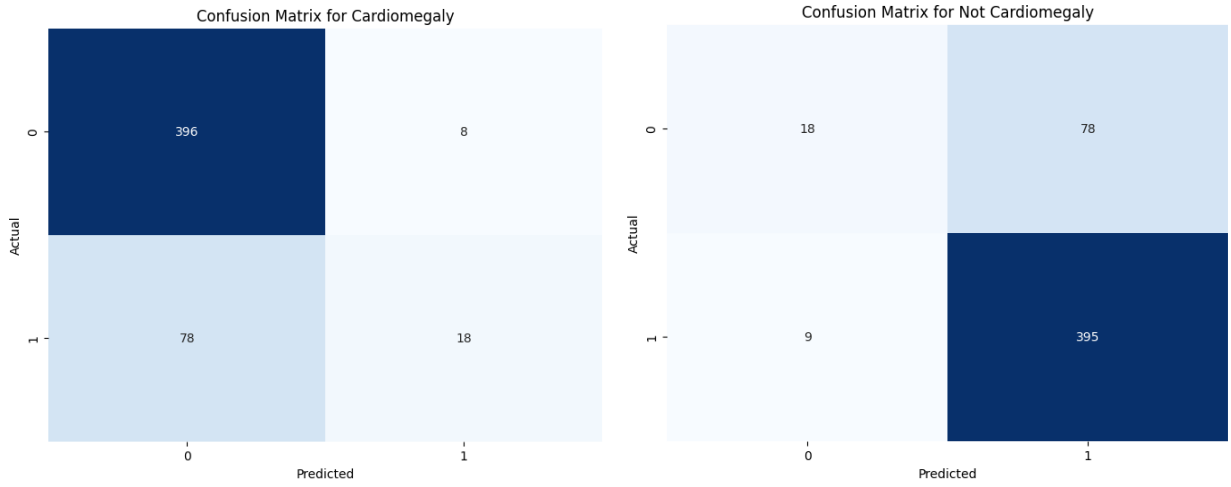


Figure 5. Confusion Matrices for EfficientNet Cardiomegaly binary classification

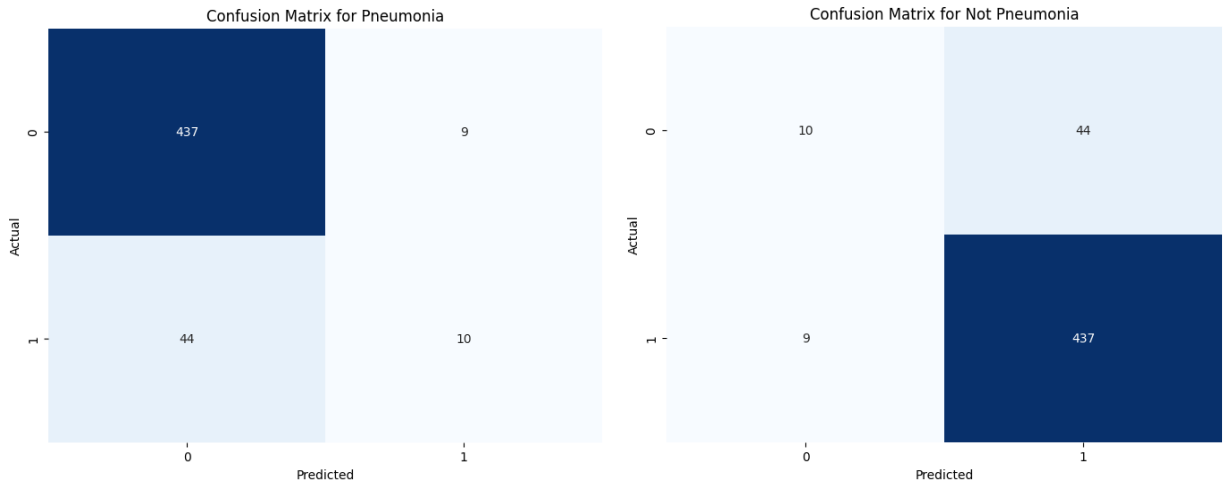


Figure 6. Confusion Matrices for EfficientNet Pneumonia binary classification

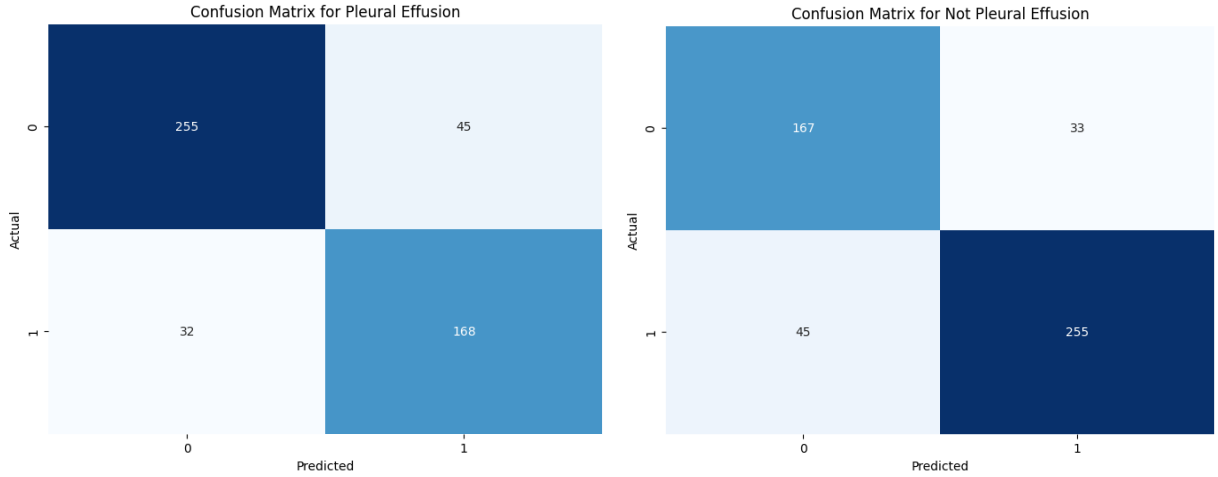


Figure 7. Confusion Matrices for EfficientNet Pleural Effusion classification

The three figures above are the confusion matrices of each of the binary classification EfficientNet models used to classify whether that disease is detected in the image or not.

Interpretation of Results

The t-SNE graphs' purpose is to represent high-dimensional data in a lower-dimensional space. For these graphs, it is important to note clusters. For Cardiomegaly, the clustering occurs at the top of the graph. For Pleural Effusion, the clustering occurs at the bottom right of the graph. For nothing, the clustering occurs at the left of the graph. The Pneumonia graph is the only graph that doesn't have an obvious clustering. Due to this, it is likely that Pneumonia is the hardest to classify – this is demonstrated in figure 1, as the AUROC for Pneumonia was the worst out of all the classes for all three consolidated models: DenseNet-121, EfficientNet, and Vision Transformer.

When comparing the performance in classification between the three consolidated models, EfficientNet outperforms DenseNet-121 and Vision Transformer. AUROC is a measure of the trade-off between

sensitivity and specificity at various thresholds. Since the Vision Transformer model has lower AUROCs for all four classes, it has worse discrimination and predictive capabilities. There are two reasons for this. First, the training of vision transformers is computationally expensive, and since there was a lack of computational resources for this project, not enough training rounds were used for the vision transformer to outperform the other models. Also convolutional neural networks are more suited for capturing spatial hierarchies in images through their use of local receptive fields and pooling operations, since spatial relationships are crucial for classification of diseases, the models with CNN architecture performed better. DenseNet-121 is worse-off than EfficientNet because EfficientNet uniformly scales the depth, width, and resolution of the network, allowing it to achieve better performance with fewer parameters. In addition, EfficientNet is newer than Dense-Net121.

Since the most effective consolidated model was the EfficientNet, three binary classification models using EfficientNet were created for all three of the diseases to assess whether binary classification will lead to better results. The AUROC mean for the Cardiomegaly binary classification model was 0.8916 while the EfficientNet's consolidated model's AUROC for Cardiomegaly was 0.8450. The AUROC mean for Pneumonia binary classification model was 0.7701 while the EfficientNet's consolidated model's AUROC for Pneumonia was 0.7611. The AUROC mean for the Pleural Effusion binary classification model was 0.9158 while the EfficientNet's consolidated model's AUROC for Pleural Effusion was 0.8917. The AUROC values for the binary classification models using EfficientNet were higher than the consolidated model for every single class. The reason for this is the binary classification models only have to classify whether a disease is present while the consolidated model must classify for all three of the diseases. In addition, the dataset is imbalanced because there is far more data where the disease is not present compared to where the disease is present, the binary classification models are better at handling this due to a focused objective and simpler optimization.

Conclusion

The milestones established in the original proposal were to clean and filter the dataset, create binary classification models for Cardiomegaly, Pleural Effusion, and Pneumonia using a model of any architecture, creating a consolidated model to compare to the binary classification models, and perform comparative analysis. Although the consolidated model was trained first, the group has met all of these milestones, with an additional milestone added on being met as well – comparative analysis between three different models to discover the best model for the consolidated model. The two main challenges that were faced was inability to store large amounts of data on personal computers, which resulted in the usage of a downsampled dataset and lack of computational resources which resulted in usage of less rounds and epochs. If the larger dataset was used, and there were more computational resources available, the performance of all the models would be improved. However, even without the best performance, comparative analysis can still be conducted since all models were trained in the same environment and parameters.

In conclusion, the project emphasizes the significant impact of computer vision in emergency diagnostics through chest radiograph analysis. The application of DenseNet-121, EfficientNet, and Vision Transformer demonstrates the importance of selecting an accurate model when using computer vision to detect diseases in radiographs. The results of this project highlight the efficiency and accuracy of EfficientNet, as well as the fact that binary classification models were more accurate than consolidated models in general. Next steps for this project would be to assess whether it is more efficient and accurate to run binary classification models one by one on images or to run one consolidated model. In addition, newer architectures have been innovated which could beat the accuracy of EfficientNet – these must be studied as well.

Detailed List of Contributions By Each Team Member

Jonathan: background, analysis, results, interpretation of results, conclusion of final paper, DenseNet-121 training and validation.

Genki: Dataset, Preprocessing, ViT implementation, training & validation,

Aryan: ViT, individual diseases training and validation

Omkar: EfficientNet training and validation, individual diseases training and validation

Jae: EfficientNet training and validation, abstract & conclusion of final paper

References

https://stanfordmlgroup.github.io/competitions/chexpert/
https://github.com/rajpurkarlab/cheXpert-test-set-labels/tree/main
https://arxiv.org/abs/1901.07031
https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/#:~:text=The%20area%20under%20the%20receiver,use%20to%20evaluate%20classification%20models.
https://keras.io/api/applications/densenet/
https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a
https://arxiv.org/abs/2010.11929
https://paperswithcode.com/method/efficientnet#:~:text=EfficientNet%20is%20a%20convolutional%20neural,resolution%20using%20a%20compound%20coefficient.
https://www.ncbi.nlm.nih.gov/books/NBK448189/#:~:text=Pleural%20effusion%20is%20the%20accumulation,%2C%20malignancy%2C%20or%20inflammatory%20conditions.
https://www.medparkhospital.com/en-US/disease-and-treatment/what-is-cardiomegaly#:~:text=difficulty%20pumping%20blood.-,Cardiomegaly%20or%20an%20enlarged%20heart%20is%20when%20your%20heart%20is,normal%20activities%20of%20daily%20living.
https://www.nhlbi.nih.gov/health/pneumonia#:~:text=Pneumonia%20is%20an%20infection%20that,or%20fungi%20may%20cause%20pneumonia.