Predicting lung disease from X-ray images

Ngoc Diem Le[†]

2009466

Abstract—Image classification is one of the most important parts of digital image analysis. The objective is to use machine learning and deep learning algorithms to extract meaning from an image. In particular, medical image classification has been receiving growing attention from the medical industry because they are fast to obtain and easily accessible. In addition, the emergence of COVID-19 makes it more difficult to predict lung-related diseases such as pneumonia, so it is necessary to develop an optimal algorithm to help classify X-ray images of the lungs, thereby making the diagnosis of doctors faster and more efficient. This project aims to develop an algorithm to distinguish between COVID patients, Pneumonia patients, and healthy people by applying deep learning and transfer learning models. Despite the similar performances between four models, DenseNet still outperforms after considering all measurements with 93.45% accuracy.

Index Terms—Lung disease prediction, COVID-19, Deep Learning, Transfer Learning, Convolutional Neural Network, DenseNet.

I. Introduction

Lower respiratory infections remained the world's most deadly communicable disease, ranked as the 4th leading cause of death[1]. Diagnosis of lung-related diseases requires a high level of expertise and chest X-ray imaging is the most reliable method for lung disease prediction [1]. In 2020, the world faced an outbreak of the COVID-19 pandemic for the first time which has been causing heavy damage in many fields and taking many people's lives. Most of the people, died from COVID-19, had suffered high chest congestion (pneumonia) due to significant reduction in oxygen level which led to major heart attack [2]. Moreover, because of the arrival of COVID-19 disease, it is a very challenging task for medical experts to detect lung infections (either viral/bacterial pneumonia or COVID-19 pneumonia) from chest X-ray images [1].

With the continuous development of technology, Machine Learning plays an important role in medical imaging which has the potential in developing intelligent medical decision support systems to urge better the power of physicians. This project approached several machine learning algorithms to classify the chest X-ray images into 3 classes - Pneumonia, Covid, and Normal. The objective of this project is to compare several models in order to find the optimal model for predicting lung diseases from chest X-ray images. In particular, the four models used are Convolutional neural network (CNN), Dense convolutional neural network (DenseNet), VGG16, and VGG19.

This report is structured as follows. In Section II the dataset is introduced, the models are presented in Section III. The performance evaluation is carried out in Section IV and concluding remarks are provided in Section V.

II. Dataset

This project used the dataset collected by [3] which contain 4575 chest X-ray posteroanterior (PA) images. All of the images are divided into 3 groups which are normal (normal images), pneumonia (indicates patients with pneumonia), and covid (indicates patients with COVID-19 observed). Each group has 1525 samples which were collected from many sources such as Kaggle, NIH dataset, GitHub, Radiopaedia, The Cancer Imaging Archive, and the Italian Society of Radiology.

The figure 1 below shows the images extracted from the dataset with 5 images for each class. For convenience, these images will be labeled: 0 for Normal, 1 for Covid, and 2 for Pneumonia. Moreover, figure 2 visualizes the number of samples for each class. It can be seen that 1525 samples were used for each case.
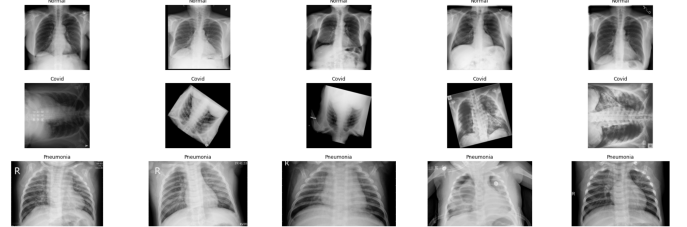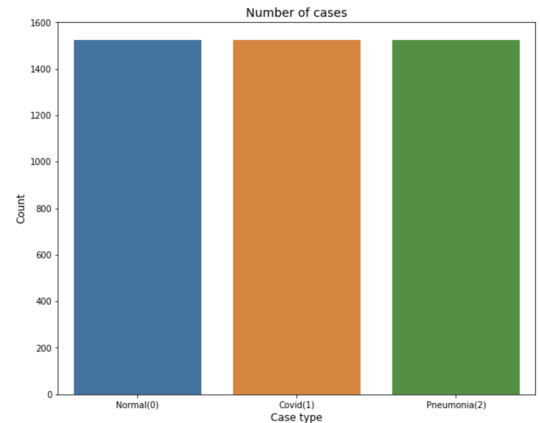

Fig. 1: Example images from the dataset

Fig. 2: Frequency for each class

[†]Department of Mathematics, University of Padova, email: ngocdiem.le@studenti.unipd.it

[1]https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

A goal of supervised learning is to build a model that performs well on new data. In this project, train-test split technique is applied for evaluating the performance of the algorithms. First, the initial data is divided in a ratio of 9:1 for the training dataset and the validation dataset. Then, the training dataset is further used to divide into two datasets, the training dataset and the test dataset with the ratio of 9:1 respectively. Therefore, the training, validation, and test datasets are split in a ratio of 81:10:9 respectively. Figure 3 shows the proportion of classes in each dataset. It can be seen that in training dataset, the proportion of each class is equal, approximately 33.33%. However, in validation and test dataset, there are some imbalances. In particular, the proportion of normal cases and covid cases are the same in both datasets while the proportion of pneumonia is slightly higher than the other two in test dataset and slightly lower than the other two in validation dataset.
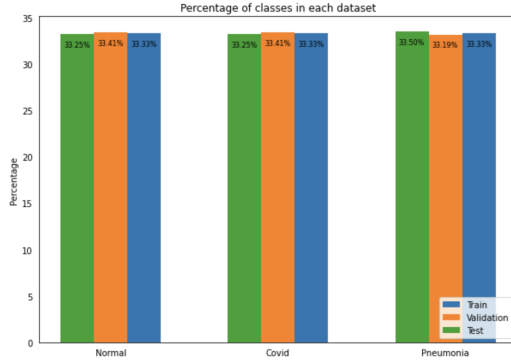

Fig. 3: Percentage of classes in each dataset

III. LEARNING FRAMEWORK

In this section, several models are conducted to evaluate their performances on X-ray images dataset. Particularly, four models were used, they are CNN, DenseNet, VGG16, and VGG19. Moreover, as mentioned above, all input images were resized to 224x224 due to the requirement of input size for model VGG16. Moreover, to help increasing generalization and adding more training data into the models, data augmentation technique is applied. In particular, zoom range is set as 0.2 and rotation range is 20. Furthermore, the optimizer used is Adam with the values varied from 0.001 to 0.0001. In order to avoid overfitting, early stopping technique is used with patience = 10 and min_delta = 0.

In CNN model, three main types of layers were used to build the architecture - Convolutional layer, Pooling layer, and Fully-connected layer. As described in figure 4, CNN model consists of 2 convolutional layers followed by a max-pooling layer and this structure is repeated three times. For each group of layers, 32 filters, 64 filters, and 128 filters were implemented with kernel size was set as 3x3 respectively. Furthermore, ReLU was used as the activation function and softmax was the activation function in the final layer which are visualized in figure 4.
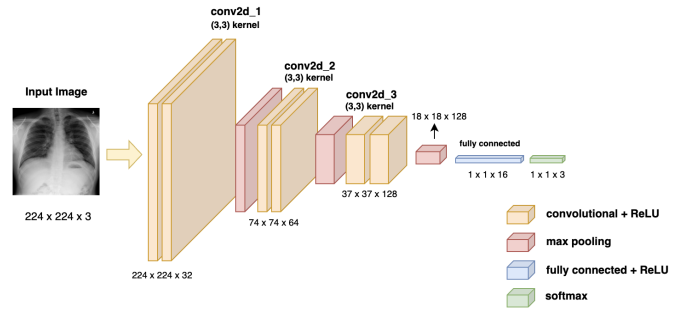

Fig. 4: CNN architecture

A. DenseNet

For the remaining 3 models, transfer learning was implemented by using pre-trained models which provided in Keras library for use. First of all, figure 5 visualizes the architecture of DenseNet, DenseNet is considered as a convolutional neural network where each layer is connected to all other layers that are deeper in the network. In this project, DenseNet121 is used to build the model.
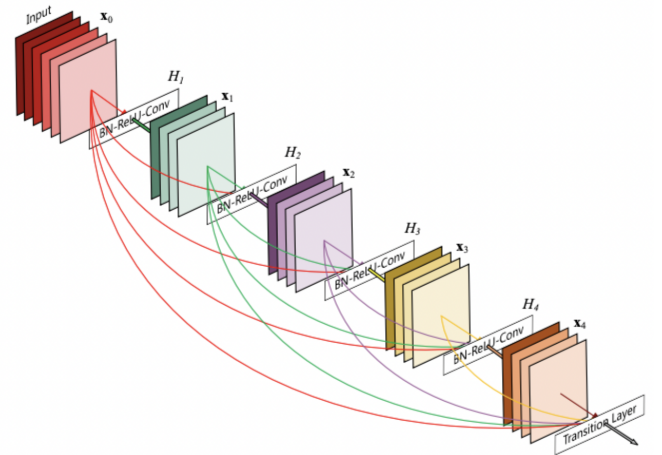

Fig. 5: DenseNet architecture

B. VGG16

The third model used is VGG16. In general, it is a CNN model proposed by Karen Simonyan and Andrew Zisserman at the University of Oxford. Figure 6 visualizes the architecture of VGG16 model. It can be seen that VGG16 is composed of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. Moreover, to explain for the number 16 in the name of the model, it is the number of layers having tunable parameters in the model which consists of 13 convolutional layers and 3 fully connected layers.

C. VGG19

VGG19 is another variant of VGG model with the number 19 indicates that the model consists of 19 layers (16 convolutional layers, 3 fully connected layer, 5 max-pooling layers,
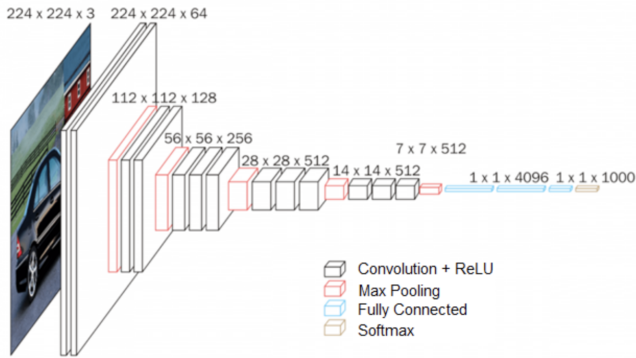
Fig. 6: VGG16 architecture

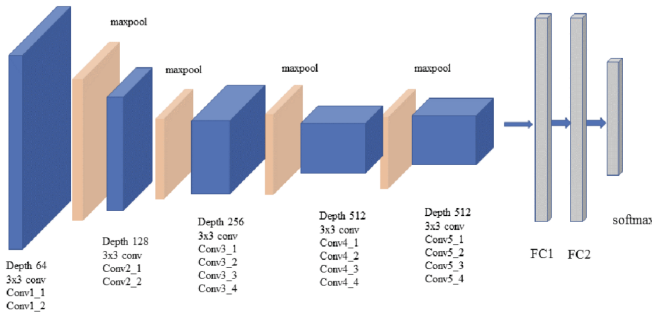and 1 softmax layer. Figure 7 visualizes the architecture of the model.

Fig. 7: VGG19 architecture

IV. RESULTS

Some machine learning techniques are used to tackle the classification task. Figure 8 visualizes the accuracy score based on test dataset of each method. In general, there is not much difference between 4 models in terms of accuracy score. Based on the figure, DenseNet gives the highest value (93.45%), followed by CNN (92.72%) and VGG19 (91.5%). Furthermore, the accuracy score of VGG16 model is lower than the 3 models above (87.86%).
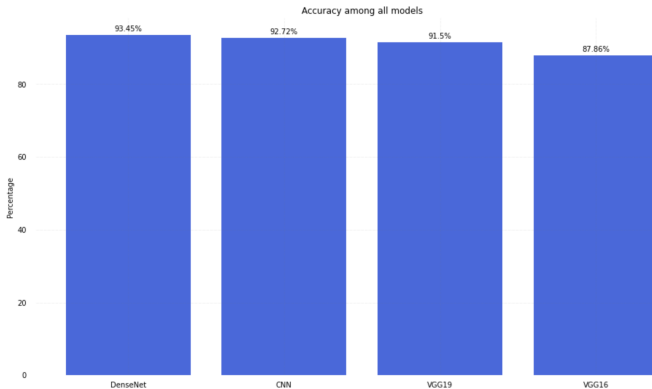

Fig. 8: Accuracy for each model

A. Convolutional Neural Network (CNN)

For the first model, figure 9 shows the accuracy between the train and validation dataset. After 20 epochs, the accuracy obtained from training dataset is approximately the same with validation dataset, 0.932 and 0.931 respectively. Moreover, it can be seen that the line represents the accuracy scores obtained by validation dataset is fluctuated. In general, not only this model but also for the remaining 3 models, it is observed that the line shows validation accuracy is usually above the line of training accuracy, which can be explained because of the imbalance between classes in validation dataset.
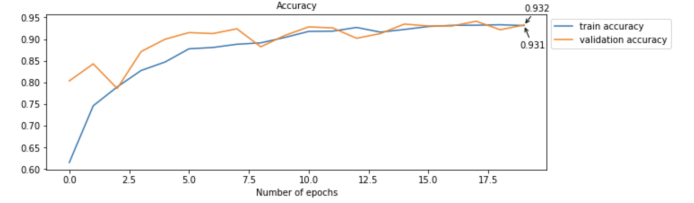

Fig. 9: Accuracy in CNN model

Figure 10 visualizes the Receiver Operating Characteristics (ROC) curve and the Area Under the Receiver Operating Characteristics (AUC) based on test dataset. This figure tells how much the model is capable of distinguishing between classes. It can be easily seen that the red line indicates the COVID class is very close to the upper left corner of the graph and moreover, the area of this case is pretty high (0.97) which shows that the model can distinguish very well the patient who has COVID, followed by the normal case (AUC = 0.94) and pneumonia case (AUC = 0.92).
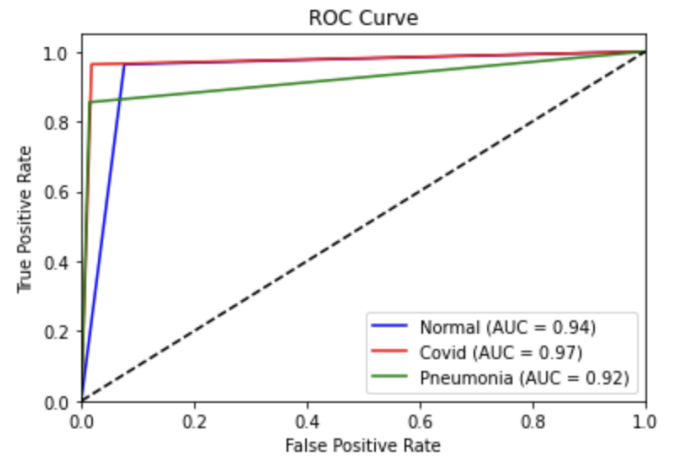

Fig. 10: ROC curve and AUC in CNN model

Considering about precision and recall, figure 11 shows the value computed by CNN model on test dataset. In terms of precision, pneumonia accounted for the highest value (0.97), in addition, Covid and Normal have the highest recall value (0.96). Which measure should be chosen to find a good model? In theory, precision is the proportion of predicted positives is

truly positives and recall is the proportion of actual positives is correctly classified. In terms of detecting lung disease in human patients, "positive" means that the patient has lung disease and "negative" means that the patient is healthy. In this case, probably people want to make sure that the classifier has high recall, so that as many lung diseases as possible are correctly detected. Therefore, this model is considered as a good model in terms of predicting patient who has COVID19 and also for healthy people due to the high values of recall for this case, 0.96.

	precision	recall	f1-score	support
Normal	0.86	0.96	0.91	137
Covid	0.96	0.96	0.96	137
Pneumonia	0.97	0.86	0.91	138

Fig. 11: Classification report of CNN model

B. DenseNet

In DenseNet model, the accuracy score in train dataset after 20 epochs is slightly higher than in the validation dataset, 0.937 and 0.934. Moreover, based on the performance in test dataset, DenseNet model and CNN model give nearly the same behavior with AUC = 0.98 for COVID case, followed by normal case with AUC = 0.95, and pneumonia case with AUC = 0.93. Figures 12 and 13 visualize the results of accuracy and ROC curve respectively. Figure 14 shows the classification report for DenseNet model, this model has significantly high values compared to the previous model, especially for the recall of normal case (0.98) and precision of Pneumonia case (0.98). However, the recall of Pneumonia case in both model are very small (0.86) compared to other values.
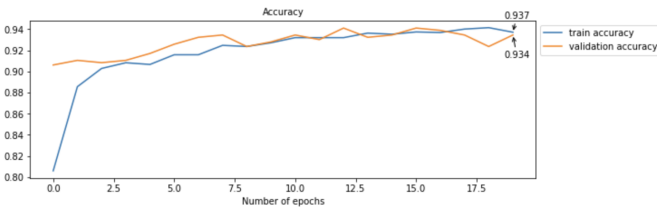

Fig. 12: Accuracy in DenseNet model

C. VGG16

Figure 15 shows the accuracy score in train dataset and validation dataset. After 20 epochs, the line represents the accuracy for validation dataset is completely above the blue line for train dataset with 0.908 and 0.874 respectively. Furthermore, based on figure 16, the AUC values for 3 classes are very low compared to other models, around 0.91. About the classification report, this model shows a very low value of the recall for Covid case (0.82) and the normal case accounted for the highest value of recall (0.98). Furthermore, Pneumonia still has the highest precision value (0.97) which shown in figure 17.
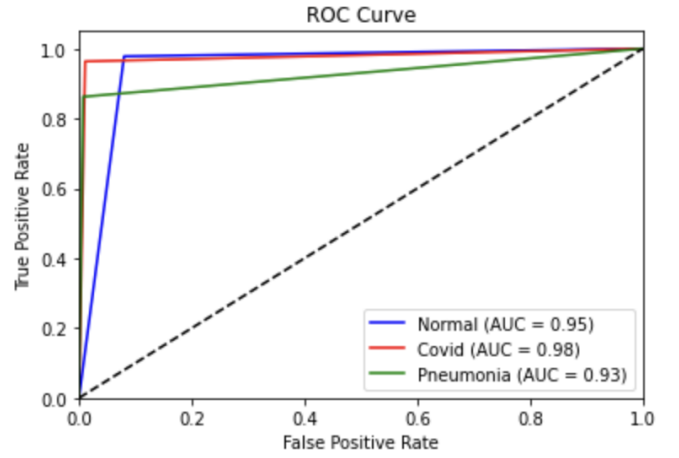

Fig. 13: ROC curve and AUC in DenseNet model

	precision	recall	f1-score	support
Normal	0.86	0.98	0.91	137
Covid	0.98	0.96	0.97	137
Pneumonia	0.98	0.86	0.92	138

Fig. 14: Classification report of DenseNet model
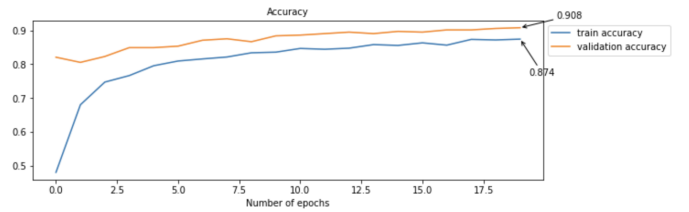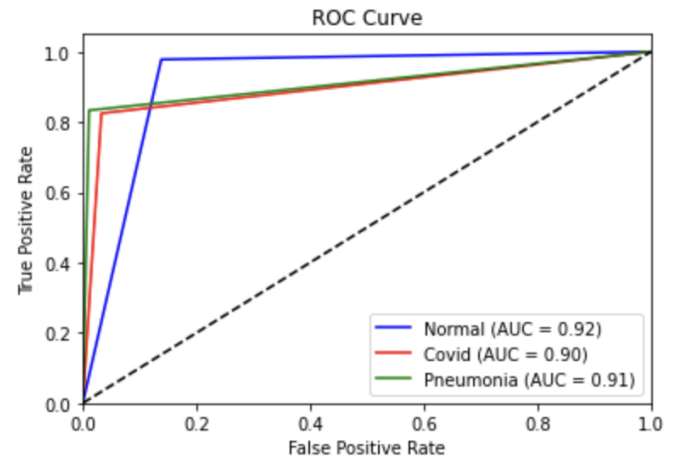

Fig. 15: Accuracy in VGG16 model

Fig. 16: ROC curve and AUC in VGG16 model

D. VGG19

For the last model, VGG19, the accuracy records on validation dataset in the 200th$^{\text{th}}$$_{\text{i0 th}}$ epoch is 0.928 and there are fluctuations in the process due to the visualization in figure

	precision	recall	f1-score	support
Normal	0.78	0.98	0.87	137
Covid	0.93	0.82	0.87	137
Pneumonia	0.97	0.83	0.90	138

Fig. 17: Classification report of VGG16 model

18. Moreover, in terms of classifying patients with disease and without disease, the ability to classify COVID patients is the highest with AUC = 0.95, followed by normal cases and pneumonia with the same values (AUC = 0.93), which is visualized in figure 19. The figure 20 shows all the measures belong to 3 classes. This model has the highest precision in Covid case (0.97) followed by pneumonia (0.96). Considering about recall, normal case has the highest value among the other 2 groups, which is 0.96.
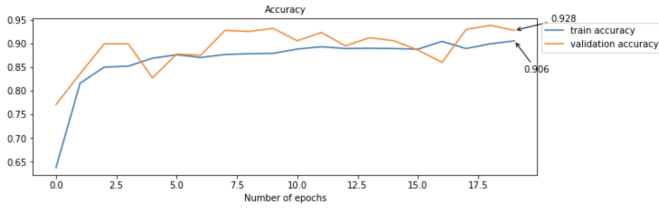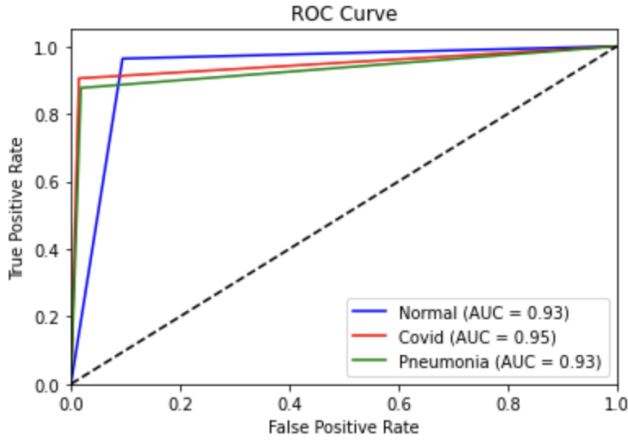

Fig. 18: Accuracy in VGG19 model

Fig. 19: ROC curve and AUC in VGG19 model

	precision	recall	f1-score	support
Normal	0.84	0.96	0.89	137
Covid	0.97	0.91	0.94	137
Pneumonia	0.96	0.88	0.92	138

Fig. 20: Classification report of VGG19 model

E. Performance Comparision

In order to have an overview as well as easily choose the best model for classification task, the performance of models is shown for comparison.

Figure 21 visualizes the number of parameters in 4 models. VGG19 is the model has the highest parameter value, and VGG16 model has the $20nd^{ndi0th}$ place with about 71.07% compared to the first model, followed by DenseNet (40.87%) and CNN got the last position with 1.98%.
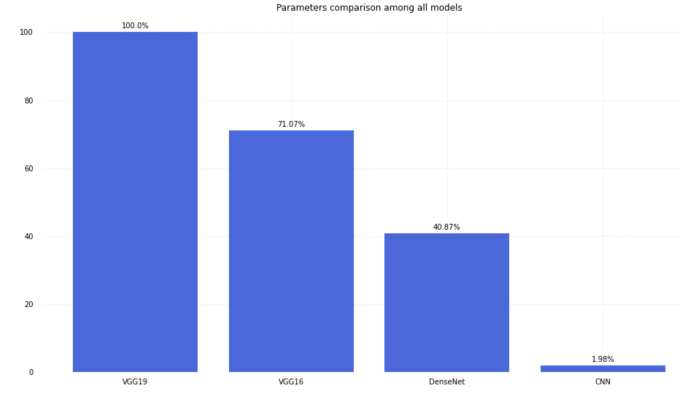

Fig. 21: Parameters comparison among all models

Figure 22 shows the time consumed by 4 models. After 20 epochs, CNN has the least time, 35.3s, followed by VGG16 (36.53s) and DenseNet (36.71s), and VGG19 consumed more time with 37.56s.
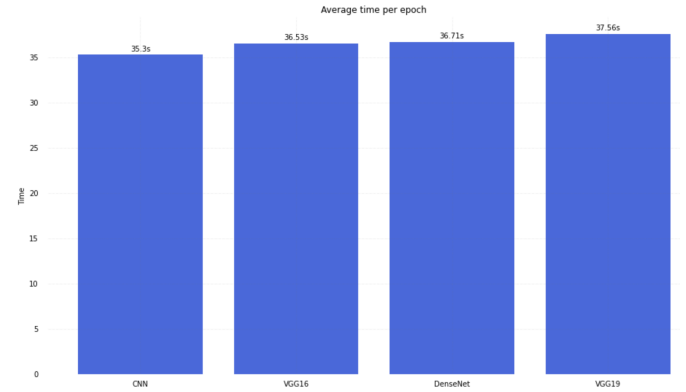

Fig. 22: Average time per epoch

Finally, compared between accuracy and time per epoch, it can be seen in figure 23 that CNN is the model uses the least amount of time but at the same time gives a very high accuracy (92.7%). However, in terms of accuracy, DenseNet gives the highest value (93.4%) and also has a small amount of time per epoch (36.71s). The other two models, VGG16 and VGG19 have slightly higher execution time and slightly lower accuracy.

V. CONCLUDING REMARKS

In conclusions, the project implemented machine learning techniques to predict lung disease from X-ray images. The dataset was collected by Amanullah Asraf, Zabirul Islam [3] with 4575 chest X-ray images which equally divided into 3 groups - Normal, Pneumonia, and Covid. Four models
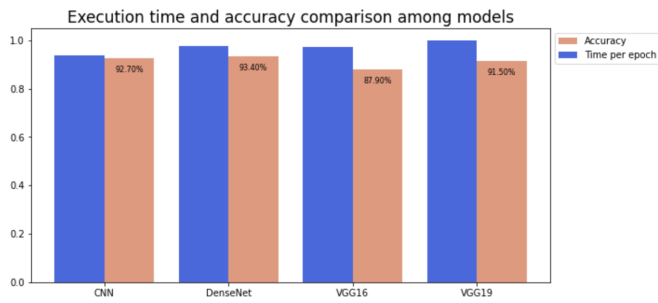
Fig. 23: Execution time and accuracy comparison among models

were applied to tackle classification task which are Convolutional neural network (CNN), Dense convolutional neural network (DenseNet), VGG16, and VGG19. In general, CNN and DenseNet have better results compared to VGG16 and VGG19. In particular, the accuracy score varies from 87.86% to 93.45% with DenseNet having the highest accuracy value. Furthermore, when other measures are considered, DenseNet consistently outperforms due to its high precision values in classifying pneumonia and Covid cases (0.98) and especially this model has high recall value in distinguishing healthy people (0.98) and Covid patients (0.96), these are the values we are most interested in when detecting the disease.

REFERENCES

[1] S. Goyal and R. Singh, "Detection and classification of lung diseases for pneumonia and covid-19 using machine and deep learning techniques," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–21, 2021.

[2] X. Chen, S. Laurent, O. A. Onur, N. N. Kleineberg, G. R. Fink, F. Schweitzer, and C. Warnke, "A systematic review of neurological symptoms and complications of covid-19," *Journal of neurology*, vol. 268, no. 2, pp. 392–402, 2021.

[3] A. Asraf and Z. Islam, "Covid19, pneumonia and normal chest x-ray pa dataset. mendeley data v1 (2021)."