Imperial College London UROP Project Report

# Exploring Best Practices for X-ray Classification

Jelaletdin Hydyrov

St John's College, University of Cambridge

October 7, 2023

## Abstract

This report explores the capabilities of the EfficientNetV2-S model [1] using PyTorch, the machine learning framework, to spot a condition called pleural effusion in chest X-rays. The images used were from the MIMIC-CXR-JPEG dataset [2] and had been labelled by CheXpert. From a subset (file p10), stratified sampling was used to pick 2800 radiographs, ensuring clarity in labels - either a solid "yes" or "no" for the presence of pleural effusion. These images were then broken down into training, validation, and testing sets.

Two variations of EfficientNetV2-S were used: one pre-trained on ImageNet and the other trained from scratch. For the training and validation phase, the respective loss curves for each version of the model plotted over 40 epochs and compared. The pre-trained version did lead to a better model over a lower number of epochs, however, it also did succumb to overfitting during the 40 epochs. The weights and biases of each model at an epoch which had the lowest validation loss were saved for testing. It was found that the pre-trained version performed better overall during testing.

## 1 Introduction

In 1895, when Wilhelm Roentgen stumbled upon a mysterious ray that could penetrate solid matter, little did he know that he was laying the foundation for a medical revolution. X-rays, named so because of their then-unknown nature, swiftly transformed the landscape of diagnostics. Fast forward to the present, and X-rays are a mainstay in medical settings globally, with millions being conducted each year to diagnose many conditions.

While the sheer volume of X-rays is a testament to their importance, it simultaneously presents a challenge: how to analyse such a large number of images effectively and correctly. In many places, especially developing regions, this challenge is compounded by a shortage of trained professionals to interpret these images.

During my time undertaking an Undergraduate Research Opportunities Programme (UROP) project provided by Imperial College London, I embarked on a journey to dabble in the possibilities of deep learning for X-ray image analysis. This was not about reinventing the wheel or claiming expertise but about earnest exploration and understanding, with all the humility of a learner at the start of a scientific journey.

The focus of this project is on the EfficientNetV2-S model and its potential in classifying chest X-rays to detect pleural effusion. Through this research, the aim is simple: to glean insights and perhaps, in some small way, to contribute to the ongoing story of X-ray imaging.

## 2 Materials & Methods

**Data Source and Preprocessing**
The MIMIC-CXR-JPEG dataset on PhysioNet contains data on many radiographs. Each study in the dataset, which has its own "study_id", is associated with a report. The dataset uses a tool, called CheXpert, to label each of these studies with their diagnoses. These labels for each study have been organised into a CSV file in the dataset. This report only considers studies that are associated with "subject_id" values that begin with 10 in order to work with a manageable amount of data on Google Cloud Storage. Moreover, only the "Pleural Effusion" column was considered and only rows with a 1.0 or 0.0 value under this column were

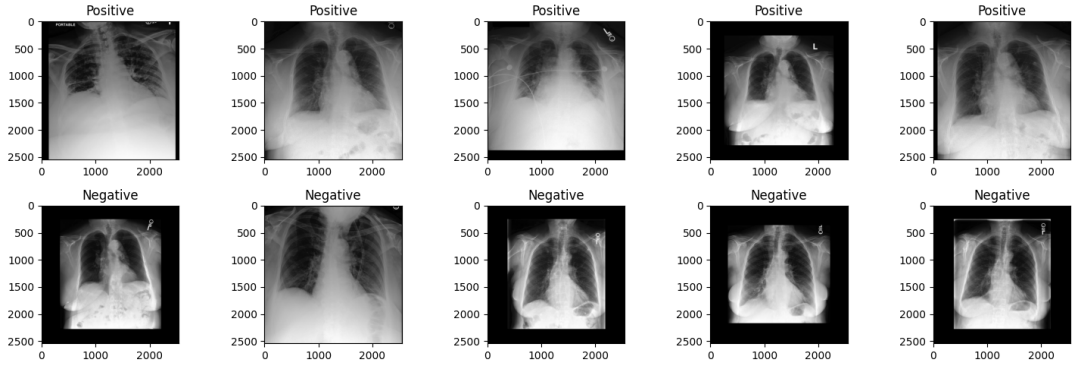


Figure 1: Sample of transformed radiographs, with their respective labels, in the AP and PA views.

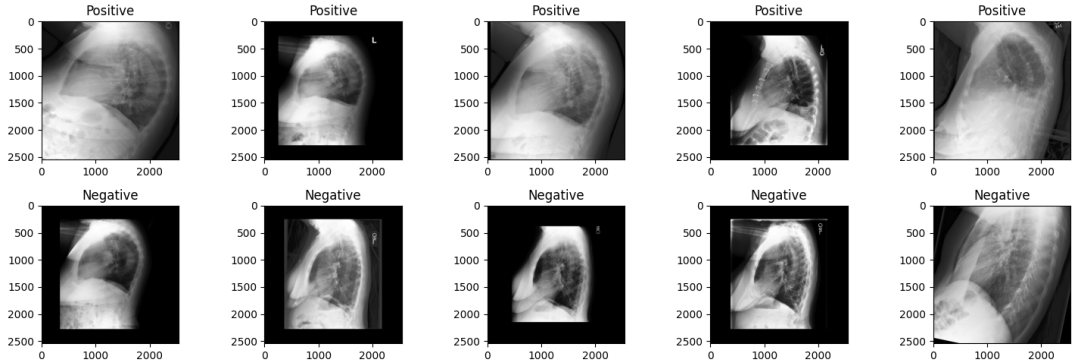


Figure 2: Sample of transformed radiographs, with their respective labels, in the lateral and LL views.

kept in order to remove ambiguity on whether a radiograph had signs of pleural effusion or not. Then, the metadata CSV file was used to find the id values of each radiograph associated with these studies in the Google Cloud bucket created to store the radiographs. Given that the dataset contains multiple different view types, two dataframes were produced to encapsulate the different views. Overall, there are four view types present for the radiographs we have filtered out for the purpose of this report. This includes the left lateral (LL), lateral, anterior to posterior (AP) and posterior to anterior (PA) views. The LL and lateral views were taken to be equivalent and so was the AP and PA views. The end result were two Pandas dataframes with "dicom_id" values on the first column and a 1 or 0 value in the "Pleural Effusion" column next to it to represent the presence and absence of pleural effusion, respectively, for the radiograph with the "dicom_id" value associated with it. One dataframe was for the lateral and LL views and the second was for the AP and PA views.

To transform the radiographs to get it ready for training the model, using the metadata CSV file, every radiograph had been centre cropped using the median width and height of all the radiographs filtered out, irrespective of view type. This is because the models will be trained using a combination of views. As the radiographs in the dataset are grayscale in nature, they were also transformed into 3-channel RGB. The effect of all these transformations can be seen in Figure 1 and Figure 2.

Then, using stratified sampling, 2800 radiographs were chosen from the two dataframes, where 1400 radiographs were picked from the AP and PA dataframe and 1400 from the lateral and LL dataframe. Stratified sampling was used to preserve the proportion of positively and negatively diagnosed radiographs from the two dataframes. The effect of this can be seen in Figure 3. From these 2800 radiographs, a 70-15-15 split was used for the Train-Valid-Test split. Two other testing datasets were also produced using stratified sampling to see the performance of the models on just the lateral and LL views as well as the AP and PA views. Therefore, there was one training set of 1960 radiographs with combined views, one validation set of 420 radiographs with combined views and

three testing sets of 420 radiographs each.

**Model and Training**
The EfficientNetV2_s model was chosen for its relatively less computationally intensive nature when comparing models that achieve similar accuracies on ImageNet. Two variations of this model underwent training: one pre-trained on ImageNet, and the other trained from scratch. Each model's evolution was tracked across 40 epochs, where the training loss and validation loss curves were then produced.

To preserve the version of each model that would be likely to have the best performance in the testing phase, after each epoch, the validation loss was kept track of. The version at the epoch where the validation loss was lowest was saved and kept for the testing phase. During the testing phase, the accuracy, precision and sensitivity of the models were calculated for each testing set. For this report, Google Colab's A100 GPU was used to speed up the whole process. In terms of specifics, cross-entropy was used for the loss, stochastic gradient descent with a learning rate of 0.001 and moment of 0.9 was used for the optimizer and for the dataloaders, a batch size of 3 was used for training and 1 for validation and testing.

# 3 Results

From Figure 5, we can observe that the pre-trained model achieved its lowest validation loss first when compared to Figure 4. However, in later epochs it did experience overfitting, which can be observed by the rising validation loss curve.

Table 1 also shows us that the best version of the pre-trained model performed better than the best version of the model trained from scratch. Moreover, it seems that both models performed best on the lateral and LL testing set.

# 4 Discussion

It was surprising to see that both models performed best on the lateral and LL testing set as it seems reasonable to think that the AP and PA views of a radiograph would have more information related to whether there is pleural effusion, as seen when comparing Figure 1 and

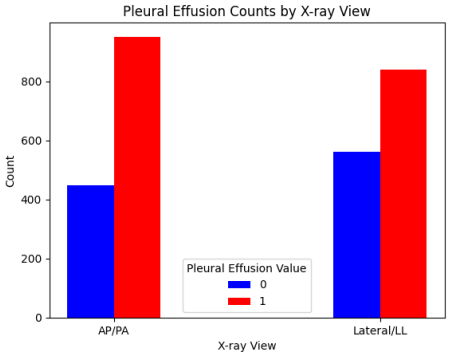


Figure 3: Bar chart showing the frequency of radiographs from each view and of each class used for the training-validation-testing split.

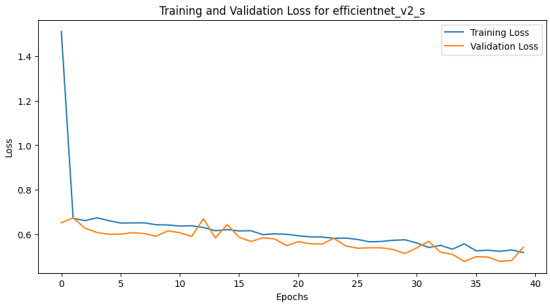


Figure 4: Training and validation loss curves for EfficientNetV2-S (not pre-trained).

Figure 2. However, perhaps this is explained by the different proportion of radiographs with and without signs of pleural effusion in the different datasets, as seen on Figure 3. This could be determined in future experiments, if perhaps, we keep the proportion of the presence and absence of pleural effusion for both general view types the same. It would also be of interest to see if the effectiveness of training could be optimised by having specific proportions of positive and negative diagnoses.

Furthermore, it would be interesting to see

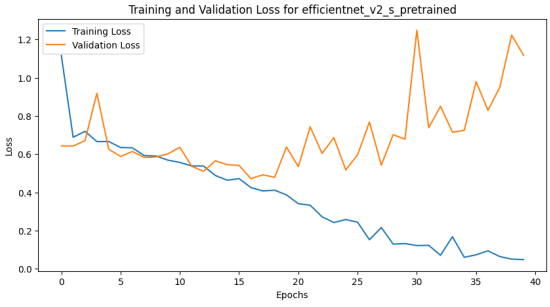


Figure 5: Training and validation loss curves for EfficientNetV2-S (pre-trained).

Table 1: Model Performance Metrics

| Model Name | View Type | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| EfficientNetV2-S | Combined | 76.67% | 89.93% | 53.29% |
| | AP and PA | 75.48% | 91.58% | 41.48% |
| | Lateral and LL | 78.57% | 87.30% | 65.48% |
| Pre-trained EfficientNetV2-S | Combined | 79.05% | 86.19% | 66.45% |
| | AP and PA | 79.29% | 86.32% | 64.44% |
| | Lateral and LL | 83.81% | 86.51% | 79.76% |

the overall performance of both models if it had been trained with more images and over many more epochs. However, due to the limitations of using Google Colab, this would be difficult to do, perhaps accessing alternative options such as data science workstations would be better approaches for deep learning tasks of this nature in general.

## Conclusions

Long-term implications from my findings would suggest using pre-trained models as a starting point for training deep learning models for X-ray classification. Also, ensuring more training data and epochs for training and validation would suggest an improvement in results from the testing phase. However, this would likely need to be done using tools other than Google Colab.

## Acknowledgements

My deepest appreciation goes to Xiaodan Xing and Dr Guang Yang of Imperial College London and the UROP initiative for this invaluable opportunity. A heartfelt thank you also goes out to St John's College, University of Cambridge, for their financial backing. This research journey, enriched by the guidance and encouragement of many, will forever be a cherished academic milestone.

## References

[1] Tan, M. and Le, Q.V. (2021) EFFICIENTNETV2: Smaller models and faster training, arXiv.org. Available at: https://doi.org/10.48550/arXiv.2104.00298 (Accessed: 06 October 2023).

[2] Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., and Horng, S. (2019) 'MIMIC-CXR-JPG - chest radiographs with structured labels' (version 2.0.0), PhysioNet. Available at: https://doi.org/10.13026/8360-t248 (Accessed: 06 October 2023).

Johnson AE, Pollard TJ, Berkowitz S, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. 2019 Jan 21. Available at: https://arxiv.org/abs/1901.07042. (Accessed: 06 October 2023).

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. Available at: https://physionet.org/content/mimic-cxr-jpg/2.0.0/ (Accessed: 06 October 2023)