# ISTANBUL TECHNICAL UNIVERSITY

# COVID-19 Detection Using Multi-Phase Training of ConvNets with Multi-Class Imbalanced Chest CT Scan Data

**Class: BLG 561E – Deep Learning**

**CRN: 14184**

**Lecturer: Prof. Nazım Kemal Üre**

**Students: Ahmet Talha Yiğit / 507191101**

# 1. INTRODUCTION

One of the most important problems of today’s world is the COVID-19 pandemic. Till now, more than 100.000.000 people have been infected by this virus in a period around a year. Moreover, around 2.000.000 deaths are caused by COVID-19 around the world as showed in worldometer.com. In pandemic cases, it is crucial to diagnose the infected cases as early and precise as possible to prevent the overloads in medical centres to avoid unnecessary deaths. X-ray computed tomography, also known as x-ray CT, of chest is one of the most effective ways to early diagnose the infection of COVID-19 in comparison with other diagnostic techniques such as PCR tests as argued in Esposito et al. (2020). Using chest CT scans, the diagnostic process can be more effective and faster. On the other hand, this fast diagnostic technique would also be helpful in the vaccination process to avoid vaccinating infected people. However, while detecting the COVID-19 existence using chest CT scans, even the experts may make wrong diagnostics. To avoid this situation, the process of COVID-19 detection by x-ray CT images can be seen as an image classification process, and deep learning models based on convolutional neural networks can be used to achieve the performance.

In this study, we propose a novel multi-phase training method to detect COVID-19 cases using a multi-class imbalanced chest CT scan data.

# 2. DATA

The dataset used in this study, is named as “CoronaHack -Chest X-Ray-Dataset” which is obtained from Kaggle. Originally the dataset has 5910 images with hierarchical class labels.

**Table 1.** The distribution of classes in the original dataset.

| Label | Label_1_Virus_category | Label_2_Virus_category | Image_Count |
|---|---|---|---|
| Normal | | | 1576 |
| Pnemonia | Stress-Smoking | ARDS | 2 |
| Pnemonia | Virus | | 1493 |
| Pnemonia | Virus | COVID-19 | 58 |
| Pnemonia | Virus | SARS | 4 |
| Pnemonia | bacteria | | 2772 |
| Pnemonia | bacteria | Streptococcus | 5 |

As can be seen in Table 1., the original distribution of classes is a highly imbalanced one. Streptococcus, SARS and ARDS have only a couple of observations while Normal, other Virus and other bacteria have a couple thousands of samples. In this case, we have merged Stress-Smoking and SARS with Virus, and Streptococcus with bacteria to create a cleaner dataset.

On the other hand, because there is not a clear definition of the virus category labels, it cannot be said if an empty virus category resembles another virus or it is unlabelled. Many other contributors of the dataset on Kaggle.com have assumed that this kind of observations resembles other kinds of viruses. Hence, we also treat as non-COVID-19 viruses while designing our approach.

**Table 2.** The distribution of classes after merging labels.

| Label | Image_Count |
|---|---|
| Normal | 1576 |
| COVID-19 | 58 |
| Virus | 1499 |
| bacteria | 2777 |

As shown in Table 2., the last version of the class labels is looking neater and can be used to create a multi-class classification problem. However, the class distribution is still imbalanced because of the low number of COVID-19 samples. This is the main problem in this study, and it will be addressed in the next sections to find a solution.

After having the class labels, the total dataset is divided to train, test and validations sets. To be compatible with the original train test split, the number of samples in test and validation sets are setted to 624.

**Table 3.** Train, validation and test splits and their class distributions.

| Label | Train_Image_Count | Validation_Image_Count | Test_Image_Count |
|---|---|---|---|
| Normal | 2191 | 293 | 293 |
| COVID-19 | 46 | 6 | 6 |
| Virus | 1242 | 167 | 167 |
| bacteria | 1183 | 158 | 158 |
| TOTAL | 4662 | 624 | 624 |

As Table 3. shows, the development set is splitted into three stratified sets as train, validation and test sets. After this step, the main aim of the study is to get the best performance on the determined test data, by using the training and validation sets. Here one of the most important class is COVID-19, the performance on it matters much. For example, if the model classifies a COVID-19 case as another class, it prevents to take necessary actions to slow down the pandemic and may cause to vaccinate someone with COVID-19 infection. On the other hand, if the it classifies another case as COVID-19, it may cause not to vaccinate a suitable person which is not preferred. In short, the model should be as precise as possible about the COVID-19 class prediction.

After creating the final labels, and splits, the dataset is resized to 128x128 pixels and normalized with the mean and standard deviation obtained from ImageNet dataset. As a result, the dataset cleaning process for the base model is finished and it is ready to create different models.

In the next sections of the study, a base model which is a basic convolutional classifier train directly with the imbalanced data, an alternative convolutional classifier trained with oversampled data and multi-stage and two-stage trained classifiers will be analysed. Their corresponding data related processes will be mentioned under their own sections.

# 3. METHODOLOGY

As the first step of the study, a base model without any loss modification and data augmentation will be built to benchmark the alternative modified models. The model architecture is designed in this step, and it is used for all of the alternative methods. The rest of the hyperparameters such as learning rate, and decay rate are tuned for each alternative model individually.

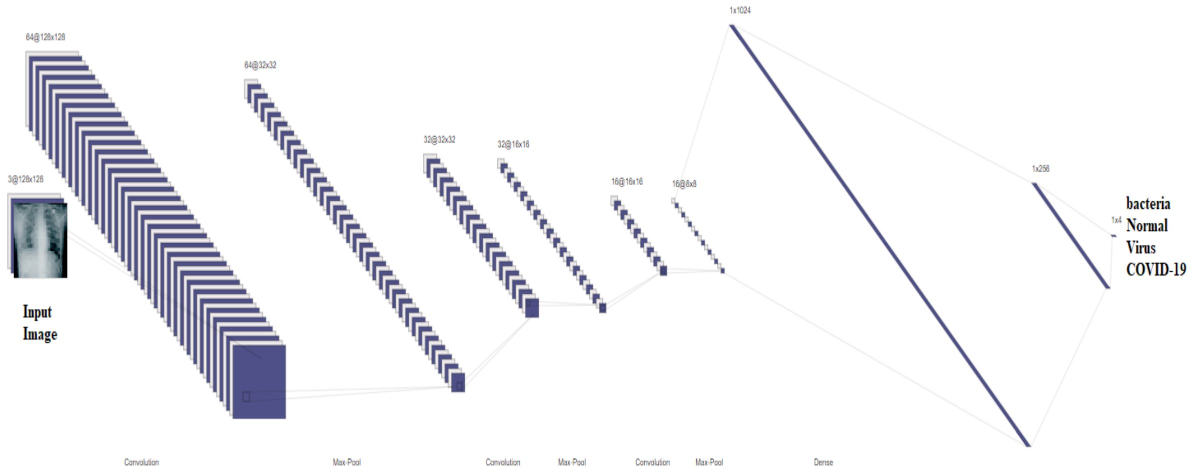


**Figure 1.** The model architecture which will be trained with many different techniques. (Original PNG file can be found in the project folder.)

As shown in Figure 1., the model has a basic ConvNet architecture. For regularization of the model, after each convolutional layer, batch normalization is applied; moreover, after the first dense layer, a dropout with 0.4 is applied. After having the architecture of the model, the nest step is to train it with many alternative methods and compare the results.

The methods which will be compared are as follows:

- The first alternative methods is training the model directly using the data as itself to create a baseline for other methods.
- The second method is training the model with the oversampled data by replicating the COVID-19 samples.
- The third method is two-stage training which firstly trains the model with under sampled data to give equal importance for feature extraction parts, and then fine-tunes the classifier part with all data to make the model fit the original data distribution. This method is proposed in Lee et al. (2016).
- The final method multi-stage training is basically the innovation that we have thought about the two-stage training method. We had the intuition that we can make the model learn the final distribution gradually by step by step transfer learning application. The method is explained in detail under it own section.

After training all of the alternative models, they will be compared to each other, and analyzed by their advantages and disadvantages.

# 4. MODEL BUILDING

## 4.1. Base Model

The base model is the one which is trained with the imbalanced (original) version of the dataset. The model is trained with 0.003 learning rate, cross entropy loss, Adam optimizer, and exponential learning rate schedular with 0.96 decay rate for 50 epochs.

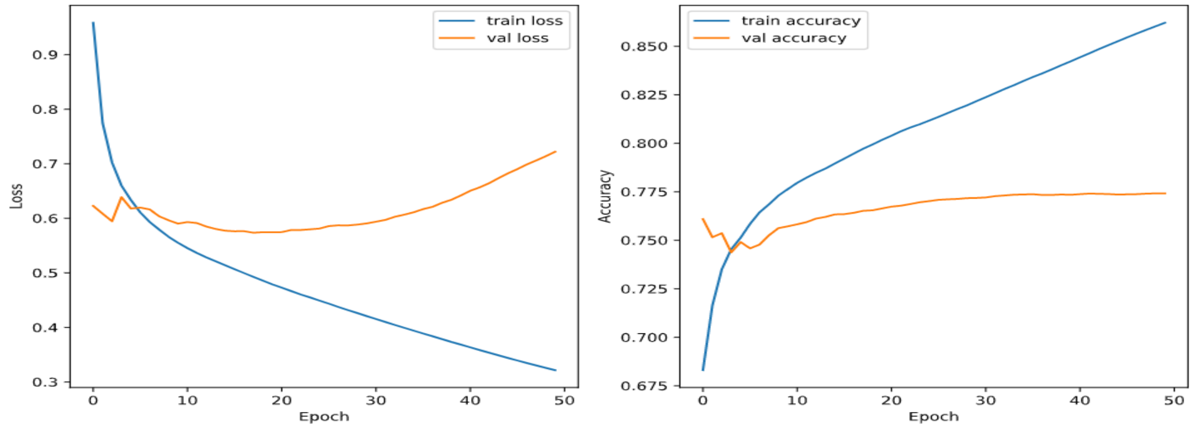


**Figure 2.** The training and validation losses and accuracies during the training of base model.

As it is seen on Figure 2., the model has already started to overfit after 17$^{th}$ epoch, but it also increased it validation accuracy till 40$^{th}$ epoch and reached an accuracy plateau. Hence, it is enough for the model to trained for 50 epochs.

After training the model, the confusion matrix is constructed on test set to see the accuracy and the other performance metrics of the model.

**Table 4.** Confusion matrix of base model on test set.

| Predicted / True | Normal | Virus | bacteria | COVID-19 |
|---|---|---|---|---|
| **Normal** | 160 | 3 | 4 | 0 |
| **Virus** | 4 | 93 | 59 | 2 |
| **bacteria** | 7 | 36 | 250 | 0 |
| **COVID-19** | 1 | 1 | 0 | 4 |

As shown in Table 4., the accuracy of the base model is pretty high with 82%, however it is also seen that the performance on COVID-19, minority class, is not so satisfying. As the goal of this study is to achieve the best performance on COVID-19 cases, the alternative models will mainly focus on increasing the performance on it instead of the overall accuracy. The precision of COVID-19 cases in base model is 0.66, which means 66% of COVID-19 predictions are actual COVID-19 cases. On the other hand, recall is also calculated as 0.66, which means 66% of actual COVID-19 cases are predicted as COVID-19. Finally, the F score is calculated as 0.66 as well. For the other alternatives, the criteria will be the F score because it shows the performance on the minority class.

## 4.2. Training Model with Oversampled Data

The first alternative is training the same model with oversampled data. By oversampling, both all information in the dataset is retained, and the performance on minority class is expected to increase.

**Table 5.** Distribution of oversampled training set.

| Label | Train_Image_Count |
|---|---|
| Normal | 2191 |
| COVID-19 | 1196 |
| Virus | 1242 |
| bacteria | 1183 |
| TOTAL | 5812 |

As shown in Table 5., the imbalance of the data is well decreased with the increased number of COVID-19 samples, which are created by replicating the original training samples. The model is trained with the same hyperparameters, optimizer and loss function to compare.

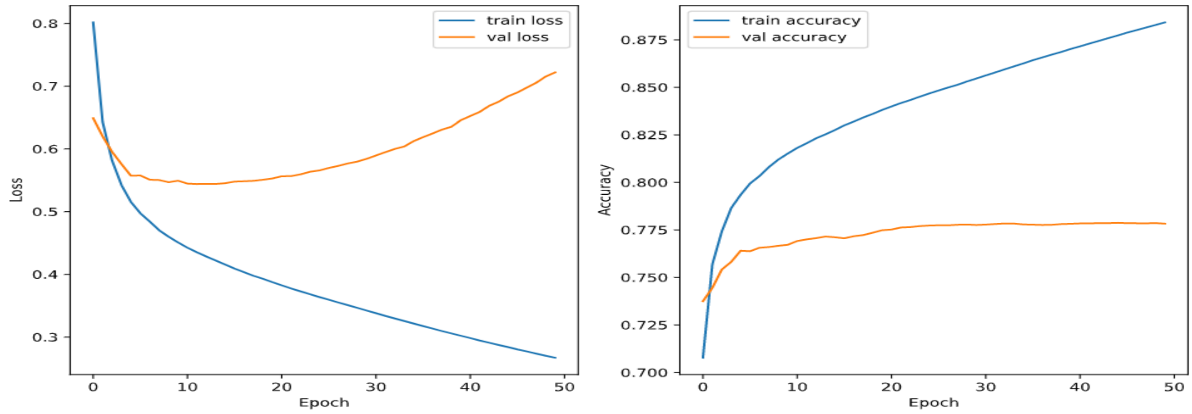


**Figure 3.** The training and validation losses and accuracies during the training with oversampled data.

Figure 3. shows that the training process is really similar with the base model. However, the model converged faster and overfitting has started in earlier than base model.

**Table 6.** Confusion matrix of base model on test set.

| **Predicted / True** | **Normal** | **Virus** | **bacteria** | **COVID-19** |
|---|---|---|---|---|
| **Normal** | 160 | 4 | 2 | 0 |
| **Virus** | 3 | 102 | 51 | 2 |
| **bacteria** | 1 | 44 | 248 | 0 |
| **COVID-19** | 0 | 1 | 0 | 5 |

As shown in Table 6., the performance on COVID-19 cases is increased by using oversampled data. The accuracy of this model is 83% and the F score around 0.77 is calculated.

## 4.3. Two-Phase Training with Undersampling

After experimenting with oversampling, an innovative technique of training deep neural networks with imbalanced data proposed by Lee et al. (2016) is applied. This technique is basically based on transfer learning. In the first phase of training, the data is undersampled to avoid imbalance and make the convolutional parts to learn more about the minority class. And in the second phase, the trained network is transferred, and the dense layers of it are fine-tuned with the original data to help the model to grasp the overall distribution.

**Table 7.** Confusion matrix of two-phase trained model on test set.

| Predicted / True | Normal | Virus | bacteria | COVID-19 |
|---|---|---|---|---|
| **Normal** | 157 | 7 | 3 | 0 |
| **Virus** | 4 | 96 | 56 | 2 |
| **bacteria** | 4 | 47 | 242 | 0 |
| **COVID-19** | 0 | 0 | 0 | 6 |

As shown Table 7., this approach has improved the f score of minority class to 0.86. It is seen that this approach has a potential to improve with novel innovations.

## 4.4. Multi-Phase Training with Undersampling

In this study, we propose an innovation over two-phase training approach introduced in Lee et al. (2016). The proposed multi-phase training approach is applicable for multi-class classification problems. It is basically based on the idea of not directly giving the original distribution to the pretrained model but giving it gradually. To do so, 3 versions of randomly undersampled data has prepared.

**Table 8.** Data used in training phases.

| Label | Original (Phase-4) | Phase-1 | Phase-2 | Phase-3 |
|---|---|---|---|---|
| Normal | 2191 | 46 | 1183 | 1242 |
| COVID-19 | 46 | 46 | 46 | 46 |
| Virus | 1242 | 46 | 1183 | 1242 |
| bacteria | 1183 | 46 | 1183 | 1183 |
| TOTAL | 4662 | 184 | 3595 | 3713 |

Table 8. shows the undersampled versions of the original dataset which are used to train the network to both learn the original distribution and the minority classes better. The training method can be explained as follows:

- In the first phase, all layers of the model are trained with the distribution shown in Phase-1 column of Table 8.

- In the second phase, the first convolutional block, consisting a convolutional layer and a batch normalization layer, parameters are frozen. Then, the rest of model parameters are trained with the Phase-2 column of Table 8.
- In the third phase, the parameters of first two convolutional blocks are frozen. Then, the rest of model parameters are trained with the Phase-3 column of Table 8.
- In the final phase, all of the convolutional blocks are frozen, and the dense layers are trained with the original data shown in Table 8.

This method has been applied step by step in this study; however, it can be designed as a single training procedure using a flexible training loop.

**Table 9.** Confusion matrix of multi-phase trained model on test set.

| Predicted / True | Normal | Virus | bacteria | COVID-19 |
|---|---|---|---|---|
| **Normal** | 159 | 3 | 5 | 0 |
| **Virus** | 8 | 88 | 61 | 1 |
| **bacteria** | 5 | 34 | 254 | 0 |
| **COVID-19** | 0 | 0 | 0 | 6 |

As can be seen in Table 9., our approach of multi-phase training has achieved the best performance minority class by increasing the F score to 0.92. It can be said that the gradual training has a better impact on accuracy of both majority and minority classes.

# 5. Final Comparison

After training models with all the different approaches, the comparison of their performance has been made.

**Table 10.** Comparison of models based on accuracy and F-score on test set.

| Model | Accuracy | F-Score |
|---|---|---|
| **Base Model** | 0.817 | 0.66 |
| **Oversampling** | **0.829** | 0.77 |
| **Two-Phase Training** | 0.807 | 0.86 |
| **Multi-Phase Training** | 0.814 | **0.92** |

As Table 10. shows, the best accuracy is achieved by oversampling method. The proposed multi-phase training method has improved both accuracy and f-score of multi-phase training approach. Although the accuracy of multi-phase training approach is even lower than base model which has been trained directly with the imbalanced original data, it shows a much better performance on COVID-19 cases.

# 6. Conclusion

As the result of this report, it can be said that using multi-phase training approach with random undersampling may be a beneficial approach for training deep learning model with imbalanced data. In this study, it showed a superior performance compared to the two-phase approach and random oversampling methods. On the other hand, a similar approach can be combined with random oversampling instead of random undersampling. The performances of all models have a possibility to be increased with better designed model architectures, hyper parameters and optimizers.

# 7. References


CoronaHack -Chest X-Ray-Dataset, https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset, last accessed 01.02.2021

Esposito, Antonio & Palmisano, Anna & Scotti, Giulia & Morelli, Marco & Vignale, Davide & Cobelli, Francesco & Tonon, Giovanni & Tacchetti, Carlo. (2020). Why is chest CT important for early diagnosis of COVID-19? Prevalence matters. 10.1101/

Lee, H., Park, M., & Kim, J. (2016). Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. 2016 IEEE International Conference on Image Processing (ICIP). doi:10.1109/icip.2016.7533053

Worldometer, https://www.worldometers.info/coronavirus/, last accessed 01.02.2021