MULTICLASS CLASSIFICATION OF COVID-19, VIRAL AND BACTERIAL PNEUMONIA IN CHEST X-RAYS WITH CNN

DEEP LEARNING ASSIGNMENT

JURRIËN BOOGERT | STUDENTNUMBER 872949 | GROUP 2

1 PROBLEM DEFINITION

In this paper, a deep convolutional neural network (CNN) architecture is proposed to diagnose COVID-19, bacterial and viral pneumonia by classifying chest X-ray images. Included are X-rays of healthy individuals. The challenge is to achieve an accurate and effective CNN classification despite the lack of sufficient and high-quality chest X-ray images and in the light of extreme class imbalance. To try to overcome these challenges, the dataset is processed using various techniques in multiple phases. These techniques including tuning the basemodel, class weight balancing, data augmentation and transfer learning.

2 DATASET PREPROCESSING

Data is gathered from https://darwin.v7labs.com/v7-labs/covid-19-chest-x-ray-dataset. The proposed data in the assignment instructions is entitled as 'darwin dataset pull v7-labs/covid-19-chest-x-ray-dataset:all-images' and consists of 2816 cases of Bacterial Pneumonia, 127 cases of COVID-19, 1606 healty cases and 1843 cases of Viral Pneumonia with a total of 6392. I would like to point out the amounts differ from the amounts written in the assignment, but the provided link and name of the dataset correspond to the requested dataset written in the assignment. The dataset from canvas also contain a different amount of samples as written in the assignment, so we used the one provided by the link in the instructions from the document. The images are loaded into two numpy arrays. One for the images and one for the labels corresponding to the images. The dataset is split into a stratified training, validation and testset of respectively 60%, 20% and 20%. Thereafter the images are normalised. The images contain three channels, but all three are identical. To make training faster one of the channels is used, but with transfer learning using an existing model like $VGG16$ or $ResNet50$ all three channels are kept, because these models demand three channels. Deep learning architectures can't use unprocessed categorical variables like the names. Therefore these categories are one-hot-encoded. An important remark must made about the distribution of classes. Bacterial Pneumonia, COVID-19, No Pneumonia (healthy) and Viral Pneumonia are proportioned as 0.44, .02, .25 and .29 respectively. Just about 2% of the dataset are COVID-19 X-Rays. The dataset is very imbalanced.

3 BASELINE MODEL

To compare the optimized models to the baseline, the plots and performance metrics of the baseline model are shared to make informed decisions when adapting and fine-tuning the model in order for a better performance.

Looking at Fig. 1 and 2, the loss on the trainingset keeps decreasing as the loss on the validationset decreases slightly the first four epochs and than starts increasing. In addition the accuracy on the trainingset keeps increasing while on the validationset the accuracy increases up to four epochs and than starts decreasing. The model starts overfitting after four epochs.

Fig. 5 and 6 show the ROC curves for both validation and testset and here the different classes can be compared against all others. To calculate the AUC score for each class in the One-vs-Rest (OvR) approach for multi-class classification, the positive class is compared with the other classes. A higher AUC score indicates that the binary classifier is better at distinguishing the positive class from the negative classes. When comparing AUC scores between classes, a higher score indicates that the binary classifier is more accurate at classifying the positive class against the other classes. Notice an AUC score of 0.96 and 0.97 for respectively validation and testset for COVID-19, which is high in comparison to the other classes. X-rays from healthy individuals are best classified compared to the rest as can be seen in Table 1. The highest accuracy is 0.72 on the validation and 0.74 on the testset, but when the class distributions are highly skewed, accuracy may not be a dependable metric of model performance. Precision, recall, F1-score and AUC could provide metrics in case of imbalance, but it is easy to get a high AUC score when the class is very small in comparison with the rest, which is the case for the COVID-19 class. Therefore, in this case, it is also good to look at the confusion matrix and corresponding metrics per class. See Fig. 3, 4 and Table 1. Here it is seen that COVID-19 has the lowest score on both validation and testset. Simply said, the model is not useful in predicting COVID-19 and the model has difficulty differentiating between Viral Pneumonia and the rest.

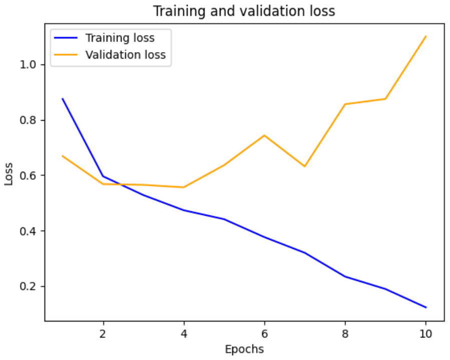

Figure 1: Loss for basemodel

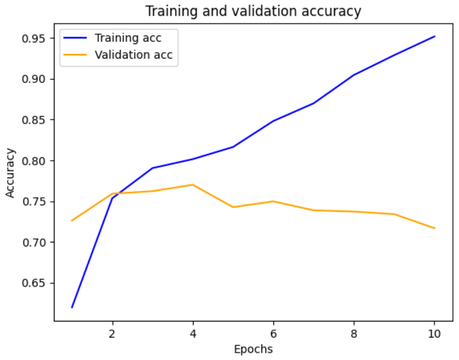

Figure 2: Accuracy for basemodel

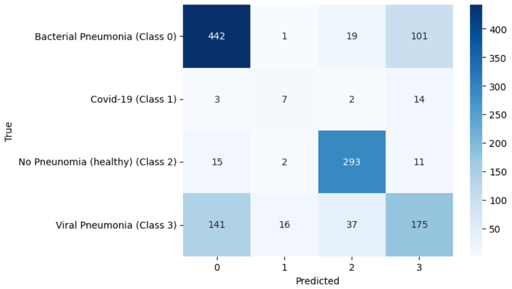

Figure 3: Confusion Matrix: validation

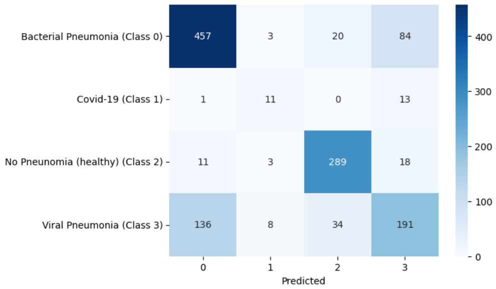

Figure 4: Confusion Matrix: test

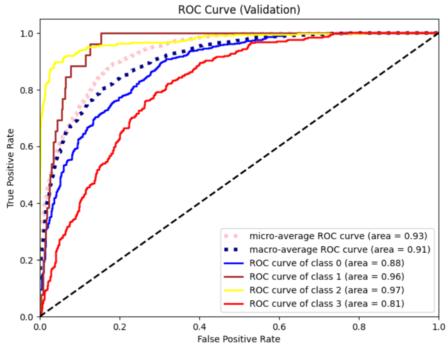

Figure 5: ROC curve: validation

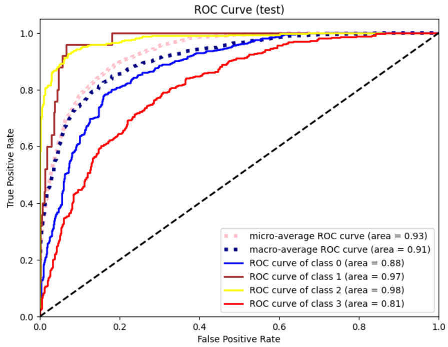

Figure 6: ROC curve: test

Table 1: classification report, basemodel

	precision		recall		f1-score	
Disease	validation	test	validation	test	validation	test
Bacterial Pneumonia (Class 0)	0.74	0.76	0.79	0.81	0.76	0.78
Covid-19 (Class 1)	0.27	0.44	0.27	0.44	0.27	0.44
No Pneumonia (healthy) (Class 2)	0.83	0.84	0.91	0.90	0.87	0.87
Viral Pneumonia (Class 3)	0.58	0.62	0.47	0.52	0.52	0.57
macro average	0.61	0.67	0.61	0.67	.61	0.66
accuracy validation/test: 0.72/0.74						

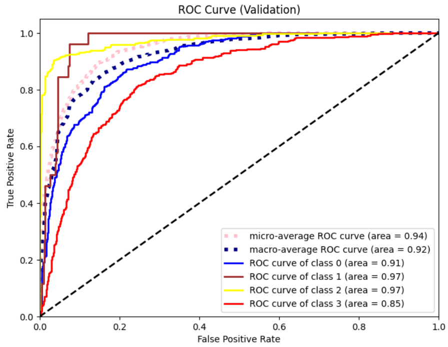

Figure 7: ROC curve: validation, 4 epochs

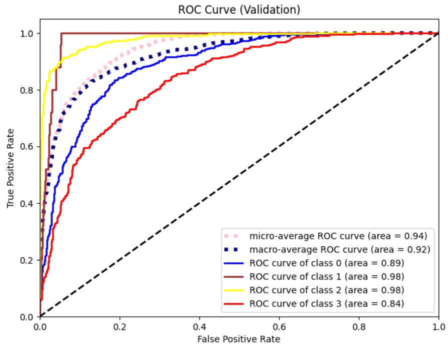

Figure 8: ROC curve: test, 4 epochs

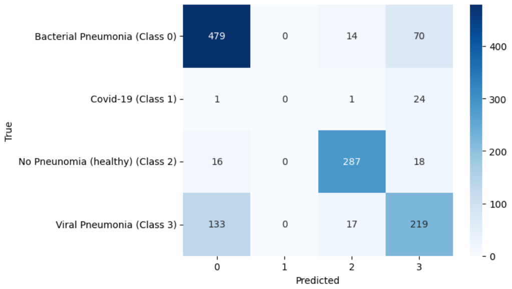

Figure 9: Confusion Matrix: validation, 4 epochs

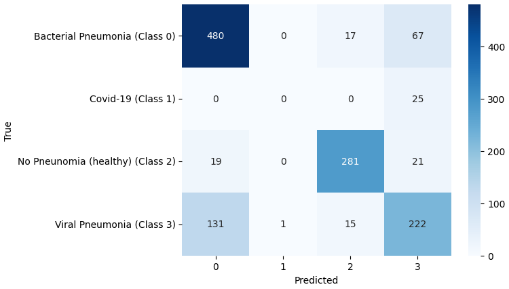

Figure 10: Confusion Matrix: test, 4 epochs

4 PERFORMANCE IMPROVEMENTS

As mentioned in the previous section the model starts overfitting after four epochs. We can prevent this with early stopping after four epochs. The results are shown in Fig. 7,8,9,10 and Table 2.

From looking at the AUC scores it seems there is an improvement, especially for class 3 against all others. But note from Table 2 and Fig. 9,10 that the scores are ill-defined for the minority class COVID-19 on all three metrics. COVID-19 has no score on the validationset, since it never predicts the class as being positive and just one time predicts Viral Pneumonia wrongly as COVID-19 on the testset, and since the full dataset is very large in comparison to this minority class the scores still result in 0.00. Simply said, this model is useless in predicting COVID-19. The performance of all other classes improve though.

Another approach to prevent overfitting is reducing the model's capacity, i.e. reducing learnable parameters as inspired by Reshi et al. (2021). Reshi et al. also suggest using max pooling to minimize overfitting, because others may not identify the sharp features easily as compared to max pooling, and a batch normalization layer. This is a technique that normalizes the input layer's scaling and activation, improving the learning process between hidden units by speeding it up and prevents the model to overfit slightly. Because we reduced the model's capacity drastically, adding dropout layers prevented the model from overfitting, but diminishing performance, so we decided to leave them out. In my view a simpler model with less parameters is preferable over a complex model with dropoutlayers, because it is less computational demanding and thus less energy consuming. As Reshi et al. also suggest we ended up using 6 convolutional layers and just two fully connected layers. It was a process of trial and error changing parameters and evaluating performance. As optimizer we settled with ADAM, since RSMprop doesn't improve accuracy. With Adagrad training and test accuracy reach a plateau, doesn't overfit on training data but oscillate around 0.73. SGD seems to get stuck in a local minimum since the validation accuracy stays the same with a learning rate of 0.001. When increasing the learningrate, it

Table 2: classification report, basemodel, 4 epochs

	precision		recall		f1-score	
Disease	validation	test	validation	test	validation	test
Bacterial Pneumonia (Class 0)	0.76	0.76	0.85	0.85	0.80	0.80
Covid-19 (Class 1)	0.00	0.00	0.00	0.00	0.00	0.00
No Pneumonia (healthy) (Class 2)	0.90	0.90	0.89	0.88	0.90	0.89
Viral Pneumonia (Class 3)	0.66	0.66	0.59	0.60	0.63	0.63
macro average	0.58	0.58	0.58	0.58	.58	0.58
accuracy validation/test: 0.77						

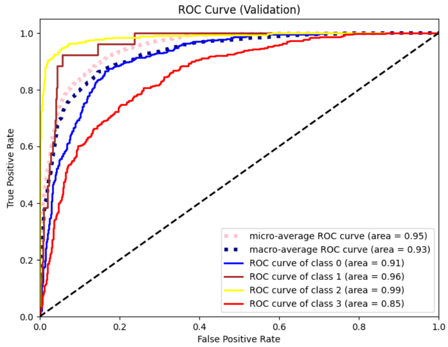

Figure 11: ROC curve: validation, tuned

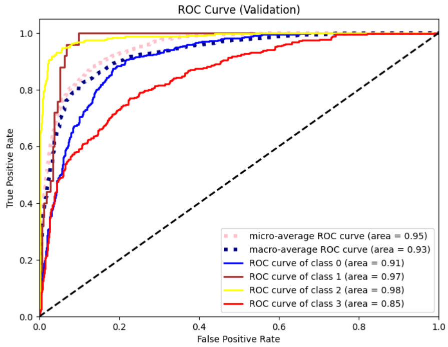

Figure 12: ROC curve: test, tuned

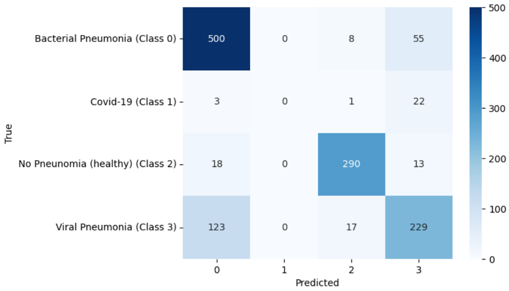

Figure 13: Confusion Matrix: validation, tuned

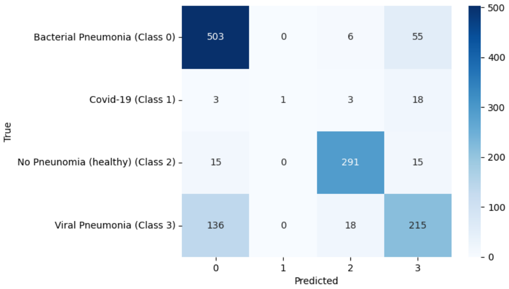

Figure 14: Confusion Matrix: test, tuned

moves out of the local minimum, but accuracy doesn't improve much. Adding L1 regularization with 0.01 to the two dense layers before the last dense layers restricts the model from overfitting, but accuracy is diminished as well. With lowering the L1 to 0.0001 overfitting stil happens, but later than without regularization. Changing L1 for L2 with 0.001 increases validation accuracy slightly by 1% but training set still overfits. We didn't use regularization in the final model. As seen from Fig. 12,13,14 and Table 3 overall accuracy – of the tuned model – as well as precision, recall and F1-score for all classes except COVID-19 after 5 epochs are higher than the basemodel. Note precision for COVID-19 is perfect on the testset, since True Positive is one and False Positive is none. On the contrary recall is still very low since all except one case is misattributed as another class. As a result F1-score is also low. Clearly the imbalance need to be addressed. We tried using class weights when training the model, so the model adds proportional weights to the different classes when minimising the loss. This can help overcome the bias towards majority classes and make the model perform better on the minority classes. As can be seen from Figure 15,16 the model slowly overfits. The validation loss increases, but the accuracy reaches a plateau. we stopped the training at eight epochs and computed the metrics as can be seen in Fig. 17,18,19,20 and Table 4. An increase in performance on the minority class, but class 3 degrades a bit using class weights, but just a few percent. The macroaverages are better compared to the reduced model. Only not for precision on the testset. We tried improving further on this by data augmentation. We only rotated, shifted, sheared and zoomed slightly on the fly when training the model. We didn't flip the images since X-rays are always frontally taken and never upside down. The resulting loss and accuracy curves are interesting, because they overlap quite well, this means the model generalizes well to unseen data (see Fig. 21,22). Downside of this way of data augmentation is that training is very slow and it takes a lot of epochs, since it sees 'new' data every epoch, since different augmented examples are created every training epoch. Validation loss and accuracy reaches a plateau around 41 epochs, so we used the model after 41.

Table 3: classification report, tuned

	precision		recall		f1-score	
Disease	validation	test	validation	test	validation	test
Bacterial Pneumonia (Class 0)	0.78	0.77	0.89	0.89	0.83	0.82
Covid-19 (Class 1)	0.00	1.00	0.00	0.04	0.00	0.08
No Pneumonia (healthy) (Class 2)	0.92	0.92	0.90	0.91	0.91	0.91
Viral Pneumonia (Class 3)	0.72	0.71	0.62	0.58	0.67	0.64
macro average	0.60	0.85	0.60	0.61	.60	0.61
accuracy validation/test: 0.80/0.79						

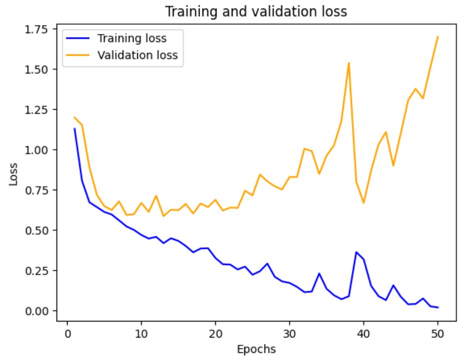

Figure 15: Loss for tuned and weighted model

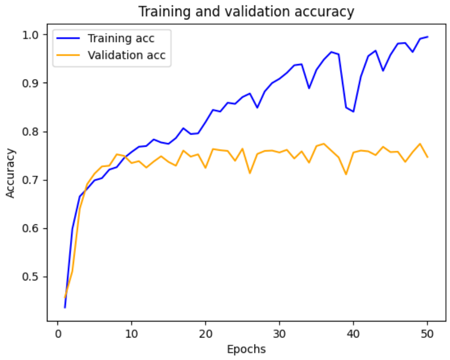

Figure 16: Accuracy for tuned and weighted model

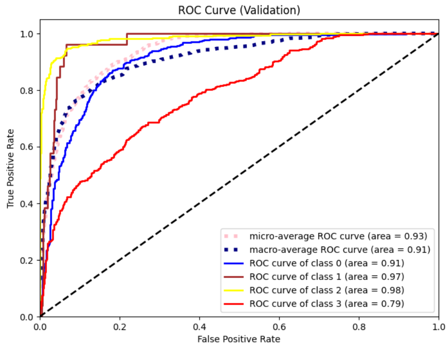

Figure 17: ROC curve: validation, tuned and weighted

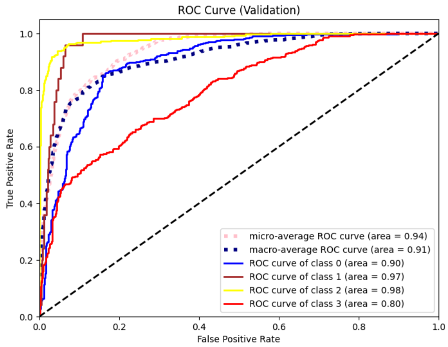

Figure 18: ROC curve: test, tuned and weighted

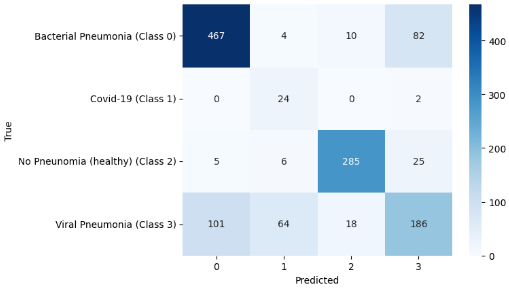

Figure 19: Confusion Matrix: validation, tuned and weighted model, 8 epochs

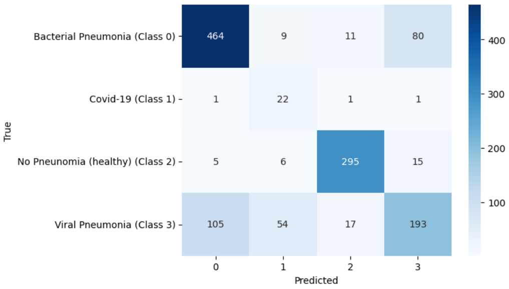

Figure 20: Confusion Matrix: test, tuned and weighted model, 8 epochs

Table 4: classification report, tuned and weighted

	precision		recall		f1-score	
Disease	validation	test	validation	test	validation	test
Bacterial Pneumonia (Class 0)	0.82	0.81	0.83	0.82	0.82	0.81
Covid-19 (Class 1)	0.24	0.24	0.92	0.88	0.39	0.38
No Pneumonia (healthy) (Class 2)	0.91	0.91	0.89	0.92	0.90	0.91
Viral Pneumonia (Class 3)	0.63	0.67	0.50	0.52	0.56	0.59
macro average	0.65	0.66	0.79	0.79	.67	0.67
accuracy validation/test: 0.75/0.76						

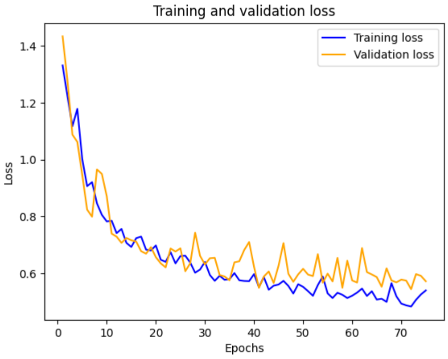

Figure 21: Loss for tuned, weighted and augmented model

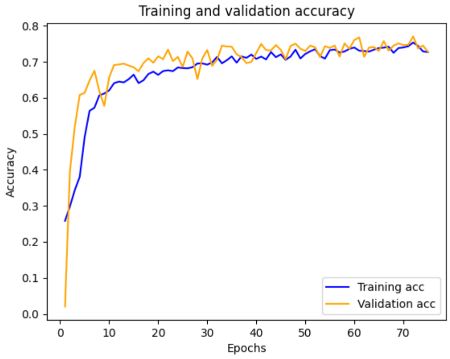

Figure 22: Accuracy for tuned, weighted and augmented model

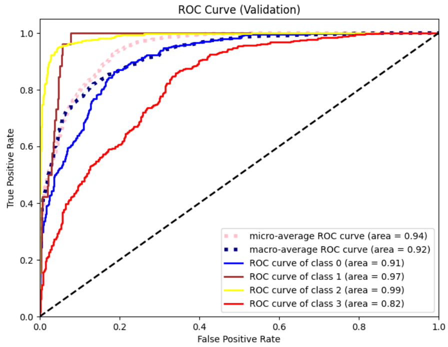

Figure 23: ROC curve: validation, tuned, weighted, augmented

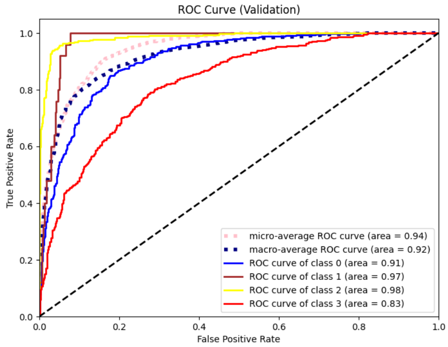

Figure 24: ROC curve: test, tuned, weighted, augmented

Fig. 25,26 look very much alike which confirms the idea that the model generalizes well to unseen data. Their performance is almost identical. Table 5 show the metrics again. The results are comparable with the previous model, but it still seems very hard to differentiate between Viral Pneumonia and the rest when looking at the ROC curves and AUC scores from Fig. 23,24, i.e. the models often classify Viral Pneumonia as another class, which especially results in a low recall score. In contrast, precision is very low for COVID-19.

When using class weights individual performance metrics average out and the scores for COVID-19 improve, but the performance on classifying Viral Pneumonia degrades. There seems to be a trade-off.

4.1 *Transfer Learning Model*

As a last resort we implemented pre-trained models $VGG16$ and $ResNet50$. Both combined with data augmentation and weighted classes. Fig. 27,28 show the jittery training curves and show slight signs of overfitting. The best accuracy and loss for validation corresponds to epoch 18 and the metrics are presented by Fig. 29,30,31,32 and Table 6. We didn't include the fully connected layers and added them ourselves to be learned with training. We didn't want the weights of the transfer model to be updated, therefore we set the trainable property on each of the layers in the models to false. The model generalizes well again to

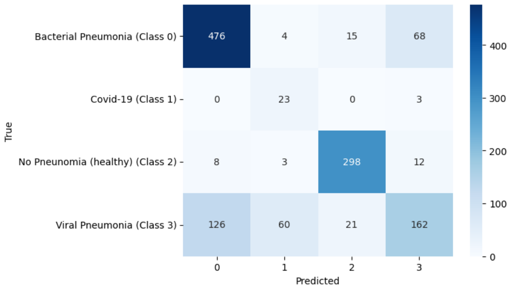

Figure 25: Confusion Matrix: validation, tuned, weighted and augmented model, 41 epochs

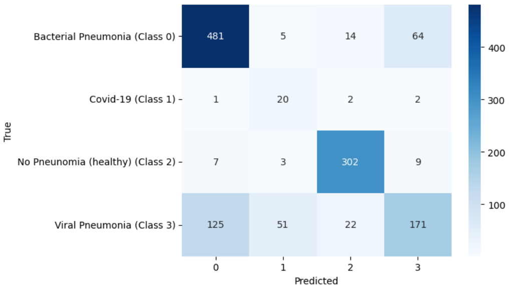

Figure 26: Confusion Matrix: test, tuned, weighted and augmented model, 41 epochs

Table 5: classification report, tuned, weighted and augmented

	precision		recall		f1-score	
Disease	validation	test	validation	test	validation	test
Bacterial Pneumonia (Class 0)	0.78	0.78	0.85	0.85	0.81	0.82
Covid-19 (Class 1)	0.26	0.25	0.88	0.80	0.40	0.38
No Pneunomia (healthy) (Class 2)	0.89	0.89	0.93	0.94	0.91	0.91
Viral Pneumonia (Class 3)	0.66	0.70	0.44	0.46	0.53	0.56
macro average	0.65	0.65	0.77	0.76	.66	0.67
accuracy validation/test: 0.75/0.76						

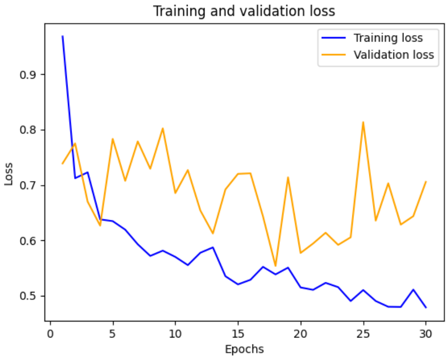

Figure 27: Loss for weighted, augmented and transferred model: $VGG16$

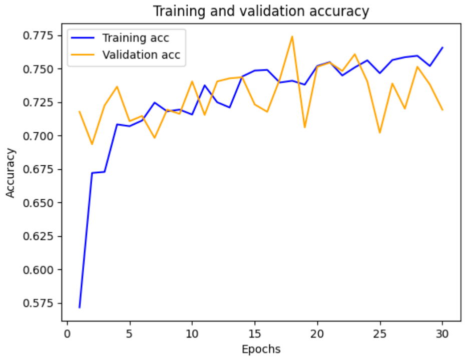

Figure 28: Accuracy for weighted, augmented and transferred model: $VGG16$

unseen data on the testset. Especially the performance on class 3 is better than the previous model. When looking at F1-score on the testset all scores are better, except for class 2, which score slightly lower with 0.89 instead of 0.91, so not much degrading. Recall is slightly worse for COVID-19 and Bacterial Pneumonia, while precision is scoring higher, but not for the healthy class. ROC curves are especially steep on the testset, which is important, because it is ideal to maximize the TPR and minimize the FPR.

For $ResNet50$ we tried lowering the learningrate with a factor 100 to 0.0001, because performance didn't improve with incremental epochs, but loss and accuracy didn't improve, so we concluded this model is not the right choice for our particular problem. After 20 epochs the validation accuracy didn't come higher than 0.54.

5 DISCUSSION

Different methods and techniques are used to increase the overall performance of the model in classification of four different classes. No method or technique can solve the problem of differentiating between all classes with equal performance in the light of an imbalanced dataset where the minority class (COVID-19) has few samples. It is common practice to select a model based on lowest loss and highest accuracy on the validation set. Therefore the basemodel to compare against is the one with results shown in Fig. 7,8,9,10 and Table 1. Reducing the capacity of the model resulted in comparable results, but faster training time. Using class weights resulted in increased performance in classifying COVID-19 correctly, but decreased performance in

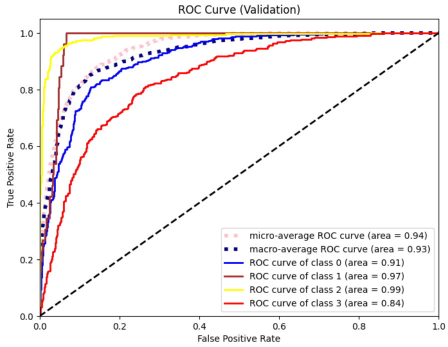

Figure 29: ROC curve: validation, tuned, weighted, augmented, transferred

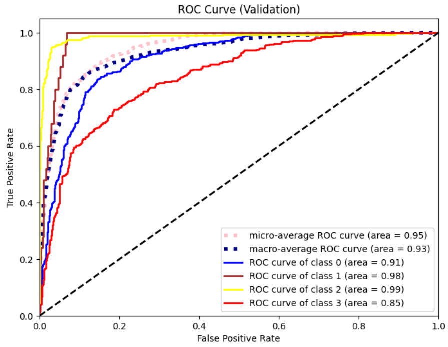

Figure 30: ROC curve: test, tuned, weighted, augmented, transferred

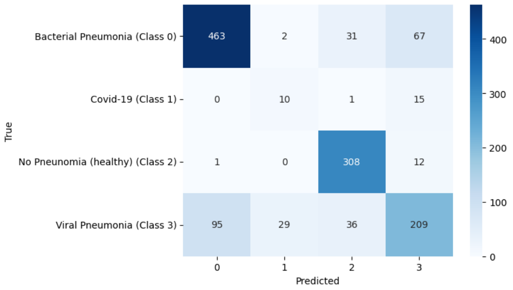

Figure 31: Confusion Matrix: validation, weighted, augmented and transfered model, 18 epochs: $VGG16$

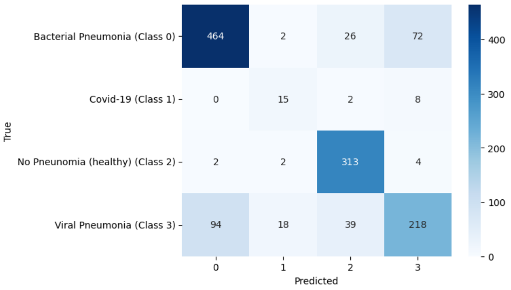

Figure 32: Confusion Matrix: test, weighted, augmented and transfered model, 18 epochs: $VGG16$

Table 6: classification report, weighted, augmented and transferred: $VGG16$

	precision		recall		f1-score	
Disease	validation	test	validation	test	validation	test
Bacterial Pneumonia (Class 0)	0.83	0.83	0.82	0.82	0.83	0.83
Covid-19 (Class 1)	0.24	0.41	0.38	0.60	0.40	0.48
No Pneunomia (healthy) (Class 2)	0.82	0.82	0.96	0.98	0.88	0.89
Viral Pneumonia (Class 3)	0.69	0.72	0.57	0.59	0.62	0.65
macro average	0.65	0.69	0.68	0.75	.66	0.71
accuracy validation/test: 0.77/0.79						

classifying Viral Pneumonia. Using data augmentation made the training much slower, but made generalizing to unseen data (on the testset) more predictable. Only the pre-trained model $VGG16$ resulted in a overall accuracy worth of evaluating. Interestingly the performance on unseen data was the same or better than on the validation set, but the scores were very jittery on the validation set. The eventual model using the pre-trained model, weighted classes and data augmentation resulted in a better performing model for on all metrics when compared to the absolute basemodel after ten epochs, but not when compared to the basemodel selected after four epochs, but here the performance was better for COVID-19.

It seems like it is impossible to get a model performing well on all metrics for all classes when there is not enough and equal amounts of data for all classes to train on and when the X-ray images are of low resolution. I would argue that the detailed structure of different pneumonia's get lost when using images of just $156X156$ pixels. I would also argue pre-trained models don't add much to the performance, when they are not trained on specific pneumonia X-ray images and and as a result fail for this classification task (Jiang et al., 2021). Other papers (Naseem et al., 2022; Umar Ibrahim et al., 2022; Verma et al., 2022) suggest it is possible to get a distinctive performance, but they use more images, balanced datasets, higher resolution images and fewer classes.

REFERENCES

Jiang, Z.-P., Liu, Y.-Y., Shao, Z.-E., & Huang, K.-W. (2021). An improved vgg16 model for pneumonia image classification. *Applied Sciences*, *11*(23), 11185.

Naseem, M. T., Hussain, T., Lee, C.-S., & Khan, M. A. (2022). Classification and detection of covid-19 and other chest-related diseases using transfer learning. *Sensors*, *22*(20), 7977.

Reshi, A. A., Rustam, F., Mehmood, A., Alhossan, A., Alrabiah, Z., Ahmad, A., Alsuwailem, H., & Choi, G. S. (2021). An efficient cnn model for covid-19 disease detection based on x-ray image classification. *Complexity*, *2021*, 1–12.

Umar Ibrahim, A., Ozsoz, M., Serte, S., Al-Turjman, F., & Habeeb Kolapo, S. (2022). Convolutional neural network for diagnosis of viral pneumonia and covid-19 alike diseases. *Expert Systems*, *39*(10), e12705.

Verma, D. K., Saxena, G., Paraye, A., Rajan, A., Rawat, A., & Verma, R. K. (2022). Classifying covid-19 and viral pneumonia lung infections through deep convolutional neural network model using chest x-ray images. *Journal of Medical Physics*, *47*(1), 57.