SAPIENZA
UNIVERSITÀ DI ROMA

FUNDAMENTALS OF DATA SCIENCE
FINAL PROJECT

COVID-19 Binary Classification

Professor:
Fabio Galasso

Authors:
Lazzari - 1917922
Violano - 2148833
Perrone - 2128080
Biagioli - 2096774

Winter Semester 2023

1 Introduction

In 2020, the global COVID-19 pandemic began, significantly impacting our daily lives and subsequently our university careers. Now, armed with the knowledge gained in this course, we have chosen to tackle, in our own way, one of the major challenges of that period: *Binary Classification of COVID-19*.

2 Data

Once the project objective was defined, we searched for a dataset containing images of both healthy and diseased patients. This search, conducted on the Kaggle platform, led us to choose the following dataset available at this link. The dataset includes $21,165$ chest X-ray images categorized into the following classes: $3,616$ images of COVID-19 patients, $1,345$ images of patients with lung opacity, $6,012$ images of patients with viral pneumonia, and $10,192$ images of healthy patients.

3 Method

For this problem, we initially chose to implement a *Fully Connected Neural Network*. However, after reading some papers related to the topic, like Paper 1 and Paper 2, we concluded that for this type of problem, it would be more appropriate to implement a *Convolutional Neural Network*.

4 Preprocessing

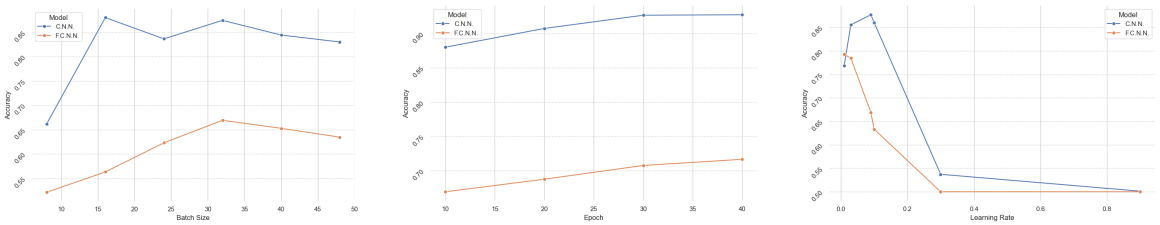
Given the nature of our problem and our data, we chose to implement the following preliminary steps: (1) removed unnecessary classes from the dataset, (2) balanced the two classes of diseased and healthy patients, (3) normalized and converted all images to grayscale, (4) converted the images from three channels to a single channel, and (5) applied resizing to the images, transitioning from a size of 299×299 to 150×150 in order to reduce the number of parameters for the Fully Connected Neural Network.

5 Parameters and Metrics

By reading the previously mentioned papers and conducting some tests, we have decided to set the following parameters for both neural networks: 2 convolutional layers and 2 pooling layers for CNN, 3 linear layers and 2 dropout layers for FCNN, Relu as the activation function. As for the optimization algorithm, we chose SGD over Adam for its speed.

In order to evaluate the performance of our solutions, we used the following metrics: *Accuracy*, *Recall / Sensitivity*, *Precision*, *F1 Score*, and *Confusion Matrix*.

We trained our neural networks through k-fold cross-validation by varying hyperparameters in order to obtain optimal results. In a medical context, such as binary classification of COVID-19, it is essential to minimize the overall error because false negatives could contribute to the spread of the pandemic, while false positives would strain the healthcare system. Therefore, we decided to maximize accuracy, and as observable in Figures 1.a, 1.b, 1.c, the second Neural



(a) *Batch Size variation* **(b)** *Epoch variation* **(c)** *Learning Rate variation*

Figure 1: *Tuning parameters*

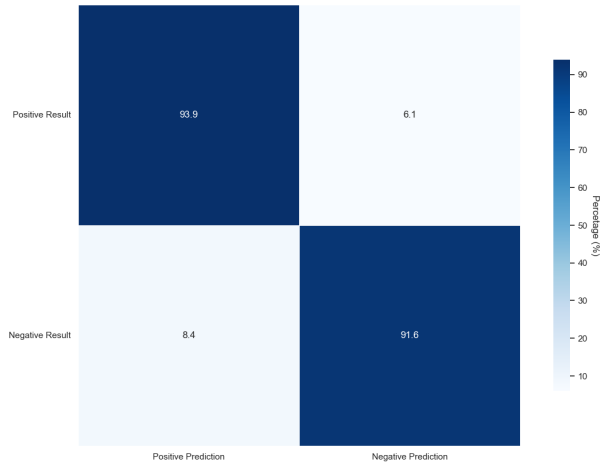


Figure 2: *CNN Confusion Matrix*

Network yields better results. For this network, we identified the following optimal parameters: (1) Batch Size = 16, (2) Epochs = 30, (3) Learning Rate = 0.09.

6 Results

With our model, we achieved an accuracy of 0.93, which is generally a good result. However, considering the classification goal, we aimed to reach at least a value of 0.95 because, from the literature, we learned that this is usually required as a minimum threshold in these contexts. Nevertheless, we are quite satisfied with the obtained results for the other metrics: Precision = 0.93, Recall = 0.92, and F1-Score = 0.93. These values reflect what we observed in the Confusion Matrix, shown in Figure 2.

7 Conclusions

After various attempts, we managed to achieve a relatively high accuracy value. However, our model could be further improved by utilizing more computational power and additional datasets.

8 Source Code

https://github.com/FrancescoLazzari/FDS-Final_Project_2023