The_Brogrammers DiCOVA 2021 Challenge System Report

Saranga Kingkor Mahanta[1], *Shubham Jain*[2], *Darsh Kaushik*[1]

[1]National Institute of Technology, Silchar
[2]independent

saranga.mahanta7@gmail.com

Abstract

A quick, efficient, and economic diagnosis procedure for Covid-19 is the utmost need of the hour. According to a recent study, asymptomatic individuals may not be entirely free of symptoms due to the virus. Such individuals may differ from healthy ones in the way they cough. The differences in the coughing sounds are subtle and not distinguishable by the human ear. However, these can be detected by Artificial Intelligence. We take a deep learning approach to analyze the acoustic dataset provided in **Track 1** of the **DiCOVA 2021 Challenge** containing cough sound recordings of both COVID-19 positive and negative individuals. We propose a ConvNet model that performs the classification between COVID-19 positive and negative with a notable AUC score of 87.07 on the blind test set provided by the same for unbiased evaluations of the models. The model takes in 15 MFCC features of the sound examples as input and produces the probability score of the classification as output.

Index Terms: COVID-19, acoustics, machine learning, respiratory diagnosis, healthcare, ConvNets

1. System Description

1.1. Methodology Overview

A duration of 7 seconds was chosen for each cough sound recording. After the required trimming and padding of the examples, 15 Mel-Frequency Cepstral Coefficients (MFCCs) [1] from each frame of the cough sound recordings were extracted using the Librosa [2] python library and used as input features into a Convolutional Neural Network (CNN) [3]. The following subsections elucidate the entire procedure.

1.2. Dataset & Pre-processing

The dataset provided for Track 1 of the DiCOVA challenge is a subset of the Project Coswara database[4] and contains a total of approximately 1.36 hours of cough sound recordings from 75 COVID-19 positive subjects and 965 non-COVID-19 subjects. The class distribution is shown in Figure 1.

A simple multi-layer neural network was trained using the provided dataset. However, due to the high class imbalance, present in the ratio of approximately 1:12 with respect to COVID-19 positive-to-negative sounds, the performance was poor. To overcome this, data augmentation was performed on the COVID-19 positive sound recordings using the Python Audiomentations[1] package. Various classes of the package namely, *Compose*, *TimeStretch*, *PitchShift*, *Shift*, *Trim*, *Gain*, *PolarityInversion* were used with different probability parameters to produce a wide range of audio recordings varied according to different aspects that define an audio instance. The data that was finally used for training after the augmentation has an improved COVID-19 positive-to-negative ratio of approximately 1:3, as shown in Figure 2.

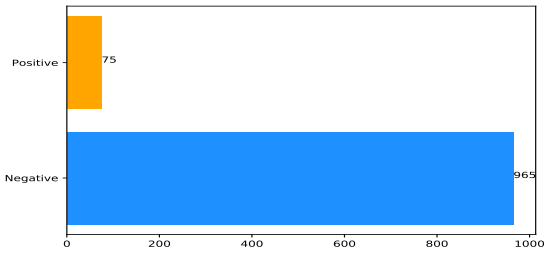

Figure 1: *Class distribution of the provided dataset*

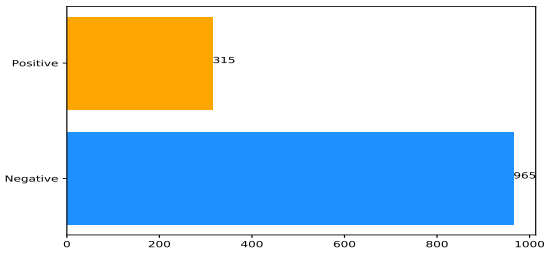

Figure 2: *Class distribution of the augmented dataset*

1.3. Feature Description

The cough sound recordings of the augmented dataset have durations ranging from approximately 0.79 seconds to 14.73 seconds, as shown in Figure 3.

Mel-spectrograms can capture small changes in the lower frequency regions since the Mel scale contains unequal spacing in the frequency bands, unlike equally spaced frequency bands in a typical frequency spectrogram [5]. Cough sounds contain more energy in the lower frequencies [6] [6], consequently MFCCs are an apt representation for the cough sound recordings [7]. For our model, 15 MFCC coefficients were chosen per frame of each example since the lower end of the quefrency axis of the cepstrum contains the most relevant information to our particular task viz. formants, spectral envelope, etc. Moreover, choosing a higher number of cepstral coefficients would proportionally increase the complexity of the model. This may

[1]https://pypi.org/project/audiomentations/0.6.0/

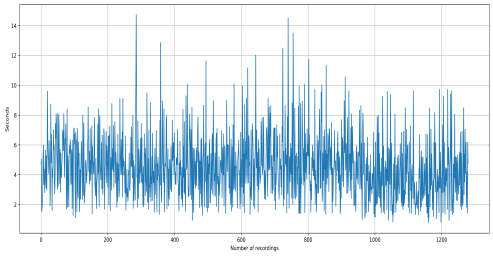

Figure 3: *Durations of all sounds of the augmented dataset*

result in a higher variance problem since the dataset is not very large. 15 MFCC coefficients were extracted from each frame of each recording. A single frame contained 2048 samples. A hop length of 512 frames was used for the framing window. The 7 second sound samples resulted in MFCC matrices having dimensions of 15x302. These matrices were then fed into our proposed CNN model.

Additionally, the input dimensions must be constant across all the training examples to be able to feed into any neural network, thus implying that the MFCC matrices of all the examples need to have a fixed dimensional size. To achieve this, each of the examples must compulsorily have a constant duration resulting in a fixed number of samples when sampled with a constant sampling rate.

Out of the 1280 recordings, it was observed that 804, 996, 1130 recordings have durations lesser than 5,6 and 7 seconds respectively. Choosing the right duration for all the recordings is crucial. Choosing a small duration will trim out important information from the sound. On the contrary, choosing the maximum duration i.e 14.73 seconds will result in a tremendous amount of sparse values in the inputs. We chose 154350 samples, which is equivalent to 7 seconds when sampled at 22050 samples/second, as the constant number of samples for all the examples because a good majority of the recordings have durations lesser than 7 seconds. Losing valuable information is a graver concern than having too many sparse values, since the dataset is relatively small. Moreover, it can be observed from Figure 3 that only a few recordings have durations above 10 seconds, since the region above 10 seconds is sparsely populated Only 150 examples had to be trimmed down while the others had to be padded with zeros to make all 1280 recordings have a constant number of samples of 154350.

1.4. Classifier Description

Out of all the models that we trained, a CNN having the architecture as shown in Figure 4 was chosen as the final model. After a commendable amount of iterations of hyperparameter tuning, the following model produced the best results.

- Convolution layer with 64 filters, kernel size of 3x3, stride of 1x1, valid padding followed by ReLU [8] activation function, accepting an input shape of 302x15x1.
- Max pooling layer with a pool size of 2x2.
- Another Convolution layer with a kernel size of 2x2, stride of 1x1, valid padding followed by ReLU activation function.
- Batch normalization layer [9]

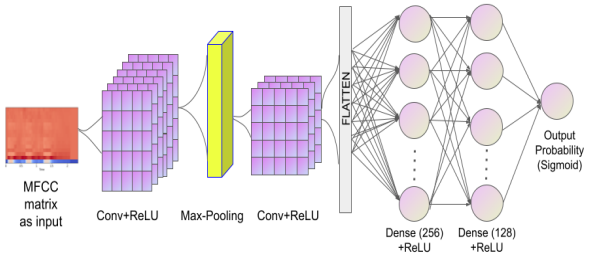

Figure 4: *Proposed CNN architecture*

- The resultant shape was then flattened for the subsequent fully connected layers.
- Fully connected layer having 256 units with kernel, bias, and activity regularizers, followed by ReLU activation function.
- A dropout layer [10] with a rate of 0.5.
- Another fully connected layer having 128 units with kernel, bias, and activity regularizers, followed by ReLU activation function.
- Another dropout layer with a rate of 0.3.
- Output layer having 1 neuron with Sigmoid activation.

The model was trained and evaluated using the stratified k-fold cross-validation technique [11] with 5 folds, similar to the number of folds provided in the challenge. The Adam optimizer [12] was used with an initial learning rate of 0.0001 while training on the examples which were further divided into mini-batches of size 32, to implement mini-batch gradient descent with respect to the binary cross-entropy loss function. The training took place over 200 epochs per fold.

1.5. Results

Figure 5 depicts the ROC [13] curves obtained from the model evaluation of each fold. The validation accuracy averaged over all five folds resulted in 94.61% with a standard deviation of 2.62%. In a similar manner, a mean ROC-AUC score of 97.36 was achieved over the folds. Furthermore, for each fold, we further evaluated confusion matrices for a decision threshold giving 80% sensitivity. Averaging these confusion matrices over the folds for each model, the following approximate confusion matrix was obtained as shown in Figure 6.

Lastly, the proposed model achieved a **Test AUC score of 87.07** on the blind test set, hence claiming the top position of the leaderboard before the competition deadline. This can be claimed as a truly unbiased evaluation.

2. Acknowledgement

We would like to thank the organizers of the DiCOVA 2021 Challenge namely, Sriram Ganapathy, Prasanta Kumar Ghosh, Neeraj Kumar Sharma, Srikanth Raj Chetupalli, Prashant Krishnan, Rohit Kumar, and Ananya Muguli for conducting this competition whose outcomes may have the potential to make a

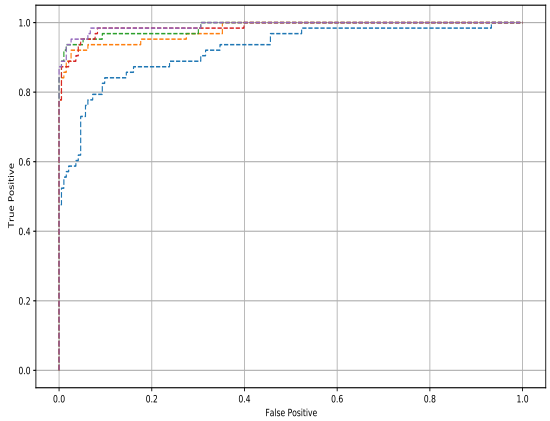

Figure 5: *ROC curves depicting model performance on each fold*

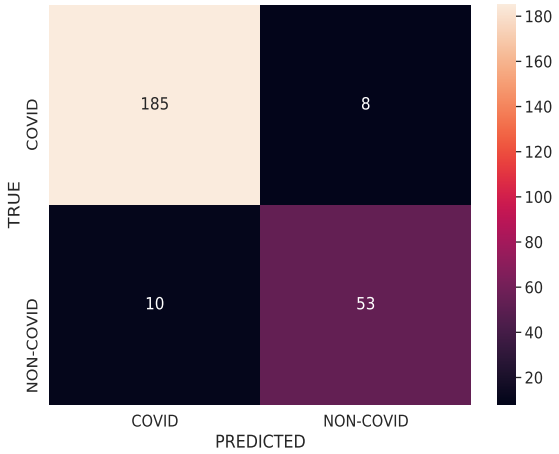

Figure 6: *Averaged model decisions computed at 80% sensitivity*

huge impact on a global level in dealing with the ongoing pandemic. We are also grateful for the timely replies which cleared any kind of queries that arose.

3. References

[1] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *International conference on computer graphics, simulation and modeling*, 2012, pp. 135–138.

[2] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[3] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE international symposium on circuits and systems*. IEEE, 2010, pp. 253–256.

[4] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara – a database of breathing, cough, and voice sounds for COVID-19 diagnosis," in *Proc. INTERSPEECH, ISCA*, 2020.

[5] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[6] J. Knocikova, J. Korpas, M. Vrabec, and M. Javorka, "Wavelet analysis of voluntary cough sound in patients with respiratory diseases," *J Physiol Pharmacol*, vol. 59, no. Suppl 6, pp. 331–40, 2008.

[7] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[8] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," *arXiv preprint arXiv:1611.01491*, 2016.

[9] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" *arXiv preprint arXiv:1805.11604*, 2018.

[10] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8609–8613.

[11] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 1, pp. 1–12, 2000.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.