INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Investigating Feature Selection and Explainability for COVID-19 Diagnostics from Cough Sounds

Flavio Avila†, Amir H. Poorjam†, Deepak Mittal†, Charles Dognin†, Ananya Muguli‡, Rohit Kumar‡, Srikanth Raj Chetupalli‡, Sriram Ganapathy‡, Maneesh Singh†

†Verisk Analytics, Inc., Jersey City, NJ, USA
‡LEAP Lab, Indian Institute of Science, Bangalore, India

{flavio.avila, amir.poorjam, deepak.mittal, charles.dognin, maneesh.singh}@verisk.com
{nanyamuguli, rohitk, sraj, sriramg}@iisc.ac.in

Abstract

In this paper, we propose an approach to automatically classify COVID-19 and non-COVID-19 cough samples based on the combination of both feature engineering and deep learning models. In the feature engineering approach, we develop a support vector machine classifier over high dimensional (6373D) space of acoustic features. In the deep learning-based approach, on the other hand, we apply a convolutional neural network trained on the log-mel spectrograms. These two methodologically diverse models are then combined by fusing the probability scores of the models. The proposed system, which ranked 9^{th} on the 2021 Diagnosing COVID-19 using Acoustics (DiCOVA) challenge leaderboard, obtained an area under the receiver operating characteristic curve (AUC) of 0.81 on the blind test data set, which is a 10.9% absolute improvement compared to the baseline. Moreover, we analyze the explainability of the deep learning-based model when detecting COVID-19 from cough signals.

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

1. Introduction

Test, treat and track has been one of the successful strategies in containing the current ongoing novel Coronavirus disease (COVID-19) pandemic. The current gold-standard reverse transcription polymerase chain reaction (RT-PCR) and the rapid antigen tests require trained personnel to collect and test the nasal swab samples. While effective, they require the subject presence at a COVID-19 test center which limits the testing capacity. A point-of-care test for COVID-19 can act as a prescreening tool and hence increase the testing capacity. COVID-19 is found to affect the human respiratory system. According to the World Health Organization's report, 67.7% of infected people exhibit dry cough samples [1].

Although cough is a common symptom for a wide variety of medical conditions, it has been demonstrated that COVID-19 cough samples have different acoustic characteristics from cough samples caused by other respiratory ailments [2, 3, 4, 5]. Several attempts have been made to diagnose COVID-19 using cough sound recordings. In the study by Imran et al. [2], three different deep learning and classical machine learning models have been applied to a relatively small data set of 543 cough samples to independently predict the COVID-19 status of a test cough sample. The final decision is then made by a majority voting approach. Using a data set of $3,621$ cough samples, Bagad et al. [3] proposed a convolutional neural network (CNN) model to identify COVID-19 cough samples. The proposed model achieved a performance of 0.72 in terms of the area under the receiver operating characteristic curve (AUC).

In addition to the research for identifying COVID-19 in individuals, several research groups have also started collecting the sound recordings from healthy and COVID-19-positive subjects [6, 7, 8, 9]. The diagnosing COVID-19 using acoustics (DiCOVA) challenge [10] was launched to accelerate and set benchmarks for efforts towards a machine learning solution with the sound recordings collected as part of the Coswara project [6]. This paper summarizes the SpeechAUC team submission to the DiCOVA-2021 challenge. The DiCOVA challenge 2021 [10] is an effort aimed at studying the feasibility of COVID-19 diagnosis using cough, breathing and speech sounds. The challenge is organized in two tracks: the first track focuses on the cough recordings alone while the second track is based on breathing and speech sounds.

The training/development data set provided for the first track consists of 1040 cough recordings, sampled at 44.1kHz. In this data set, 75 cough sounds were recorded from COVID-19-positive subjects and the remaining were recorded from healthy subjects. Due to the limited number of positive samples, the challenge adapted a five-fold cross-validation plan. To maintain uniformity among the participants, the split of the data set into the training and validation for the five folds is also provided by the challenge organizers. Moreover, a separate test set, containing 233 cough sounds, is provided for the blind evaluation with an online leaderboard-based ranking[1].

The goal of the first track is to develop a model to predict the COVID-19 status of the subject based on his/her cough sound. The participants are required to submit a score indicating the probability of COVID-19 for each recording in the blind test set and the five-fold validation sets to the online scoring server. The performance metric used for the challenge is the AUC. The receiver operating curve (ROC) is computed at a finer resolution of $1e-4$ and the AUC is computed using the trapezoidal rule. A detailed description of the challenge, the data set and evaluation criterion can be found in [10].

This paper presents promising models for distinguishing between positive and negative COVID-19 cough samples. The proposed method, a combination of feature engineering-based and CNN-based models fused at the probability score level, was able to improve the baseline AUC by 10.9%. Furthermore, we analyzed the effect of feature selection on the performance of the feature engineering-based model. We also illustrate, through a visual analysis of the CNN model, the basis on which the model makes a decision about a sample.

[1] https://competitions.codalab.org/competitions/29640

http://dx.doi.org/10.21437/Interspeech.2021-2197

2. System Description

The proposed system is an ensemble of two methodologically diverse approaches. In the first approach, the signal is summarized by a large set of features that have been prevalent in speech processing (OpenSMILE features [11]). Then, a classical machine learning technique, namely support vector machine, is used to distinguish between COVID-19 and non-COVID-19 cough samples. In the second approach, a deep learning-based model is used to identify the pattern of COVID-19 cough samples based on the log-mel spectrogram of a signal. These two approaches are described in the following sections.

2.1. Feature Engineering-Based Model

This approach, which we refer to as the *Functionals* in the rest of this paper, is based on extracting a large number of features that have been successfully applied to a wide range of speech classification problems, such as emotion recognition, Alzheimer's disease detection [12] and deception detection [13]. For feature extraction, we adopted the OpenSMILE software [11]. Before extracting the features, all audio samples were re-sampled to 16 kHz. Moreover, we applied a pre-emphasis filter to the signals to emphasize the higher frequencies by increasing the amplitude of the high frequency bands and decreasing the amplitudes of the lower bands.

The feature extraction component of the OpenSMILE tool works by calculating low-level descriptors (LLDs) across the entire signal, on a frame-by-frame basis. This large list of LLDs includes mel-frequency cepstral coefficient (MFCC), perceptual linear predictive coefficients, linear predictive coding, formants, jitter, shimmer and pitch. In order to summarize this vast quantity of time-dependent information into a fixed-length vector, a set of functionals is applied to the LLDs, producing the final features for model training. The list of functionals includes mean, moments, peak values, zero-crossings, coefficients of the discrete cosine transformation, and autoregressive coefficients [14]. For this particular model we chose the ComParE2016 set of features [14], which were used as baseline features in the Interspeech 2016 challenge. A total of 6373 features were extracted for each audio signal in the data set. We used support vector machine (SVM) with the radial basis function kernel as the classifier.

In our initial experiments, the most relevant hyper-parameters were the regularization parameter and the number of selected features. The latter parameter is an input to a random forest-based feature selection procedure. Using a random search over a wide range of values, we found that the regularization parameter had a small negative correlation with performance while the number of features had a positive correlation. More specifically, a Pearson correlation coefficient of 0.16 between the number of selected features and the AUC values is observed, indicating that more features selected tend to lead to slightly better performance on average (shown in Fig. 1). These observations motivated us to pick a small value, namely 1, for the regularization parameter and the entire feature set (6373D) was used for the training/validation.

2.2. Deep Learning-Based Model

For the deep learning-based model, we adapted a CNN model proposed in [15], which is composed of 12 convolutional layers and 2 fully connected layers. We selected this network due to its ability to transfer well on a variety of audio classification tasks [15]. Due to lack of training data, the network trained from scratch did not perform well. To tackle this problem, we initialized the weights of all layers, except the last fully connected layer, with the weights of the model pre-trained on the large database of almost 2M audio samples provided by the authors in [15]. In the training phase, we fine-tuned all the layers.

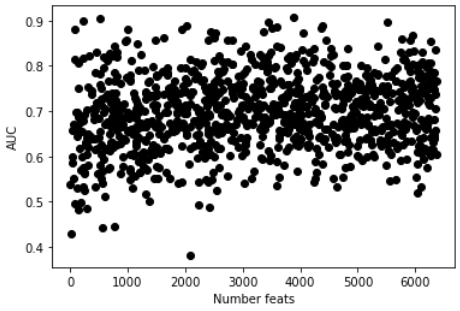


Figure 1: *Analysis of effect of number of selected features on AUC scores.*

As the input feature to this network, we used the log-mel spectrogram. Prior to feature extraction, we processed the audio samples in a similar way to the baseline system, by subtracting the signal mean and normalizing the amplitudes. Moreover, we fixed the duration of all signals to 10s. That is, we windowed the signal if it was longer than 10s and zero-padded it if it was shorter than 10s. Using the Hann window of length 32 ms, 10 ms hop and a 512-point FFT, we obtain the log-mel features of 64D. The signals were transformed into the log-mel spectrograms of size 1001×64.

Since we are working with a highly imbalanced dataset, in which the negative COVID-19 class has 15 times more samples than the positive class, we tried different loss functions aside from cross-entropy (proposed in [15]). The focal loss, dice loss and weighted cross-entropy loss are proven to work well in data imbalance settings [16]. From these loss functions, we used weighted cross-entropy with $1/15$ weight to the negative class and $14/15$ weight to the positive class. Moreover, we applied MixUp [17] and SpecAugment [18] techniques to augment the data in both classes in the same proportion to avoid introducing a bias in the minority class. For all other training parameters, we used the values proposed in [15].

2.3. Model Combination

It is well known that combining methodologically diverse models tends to improve performance. In particular, despite similar AUC performance, these two methods have relatively small correlation between output probabilities, ranging from 0.31 to 0.45 depending on the validation fold. This observation motivated us to perform averaging of the normalized probability scores of the two models.

3. Results

3.1. DiCOVA Challenge Results

The two models described in Sec. 2 were first trained and validated using all five folds provided by the challenge organizers. The primary metric for model parameter selection was the average validation AUC calculated over those folds. In Table 1, we can compare the average validation AUC and the corresponding AUC on the blind test set of the proposed and the baseline systems. The baseline model is a multi-layer perceptron, consisting of a single hidden layer with 25 hidden units, hyperbolic

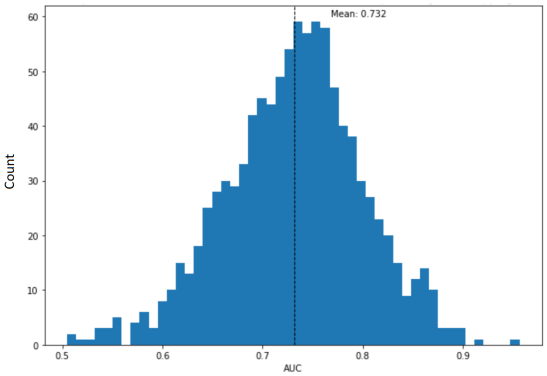


Figure 2: *Distribution of AUC values calculated by a Monte Carlo analysis across several training/validation trials.*

Table 1: *Performance comparison between the baseline and the proposed systems in terms of AUC.*

System	Average AUC	
	Blind Test Set	Validation Set
Baseline	0.6985	0.685
Functionals	0.8041	0.743
CNN	0.7464	0.781
Score Fusion	0.8075	0.780

tangent activation functions, and the l_2 regularization penalty [10]. The model is trained with 39 dimensional MFCC together with the *delta* and *double-delta* coefficients calculated from the frames of the training audio signals.

Since the model based on functional features is faster to train, once the hyper-parameters were selected (as explained in Sec. 2.1), we also validated the model using both a leave-one-out (LOO) and a Monte Carlo-based cross-validation approach. Here, we select random five folds, with four of them used for training and the remaining one for validation. The resulting distribution of scores provides us with some intuition about the range of expected AUCs on the blind test set, as can be seen in Fig. 2. The LOO aggregate AUC was about 0.72. This analysis indicated that the model had strong average performance with significant number of experiments providing AUC above 0.80. The model achieved a test AUC of 0.804 (shown in Table 1) which improved the baseline performance by 10.9%.

For the CNN model, we trained two models, each for every train-val fold by using different random initializations of the last fully connected layer. We then took the best performing model of each train-val fold, resulting in five models in the end. The models were trained for $10,000$ iterations. During the test phase, we took the weighted average of the test scores from all five models. To assign weights, we applied the softmax function to the AUC values of validation fold. As presented in Table 1, this model achieved an AUC of 0.746 in the blind test set which improved the baseline AUC by 4.8%.

Finally, by combining the models at score-level, the performance of the CNN and Functionals models was improved by 6.11% and 0.34%, respectively. Although the improvement is marginal, we expect that combining two methodologically different models might improve the robustness of the system. However, more research is needed to support this hypothesis.

Table 2: *Average AUC comparison for the Functionals system of different feature set sizes.*

Model	Average AUC			
	K=1000 S=210	K=1500 S=352	K=2000 S=523	K=2500 S=737
SVM (Linear Kernel)	0.718	0.747	0.717	0.727
SVM (RBF Kernel)	0.735	0.754	0.743	**0.768**

3.2. Feature Selection in the Functionals System

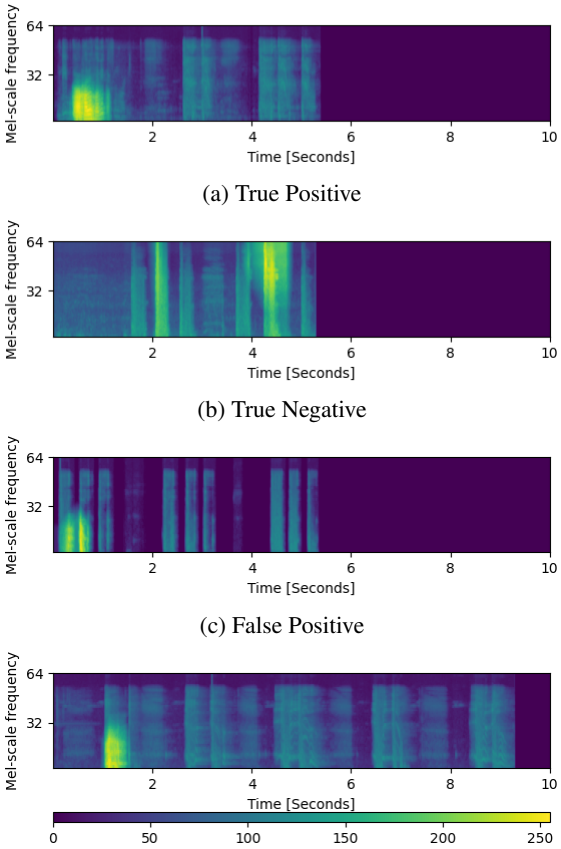
In this section, we further explore the SVM model for feature importance analysis. For the sake of interpretability, it is important to find the features that are most influential to the target class prediction score by the model. This will facilitate inspecting whether the acoustic features are influenced by the medical condition or by other confounding factors. To this aim, we train an SVM model with a linear kernel on the training data for each fold. We then rank the features in decreasing order of the absolute value of the corresponding SVM coefficients. Five orderings of the features, one from each fold, are obtained in this manner. In each ordered list, we select the first K features and take an intersection across the 5 lists to get a set of S important features. For $K = 1000$, we found $S = 140$ in our experiments, and for $K = 2500$, S was equal to 540. We train SVM models on the reduced feature set for the five folds. The average validation AUC for the five folds is found to be 0.90 ± 0.02 for $K = 1000$ and 0.96 ± 0.02 for $K = 2500$. We have also looked at the categories of the features obtained for $K = 1000$. The features are the LLDs derived from the set {MFCCs, auditory spectrogram, FFT magnitude, local jitter and shimmer of the raw samples, F0, zero crossing rate, etc.}.

Since the experiments are performed on the total development data, the validation set for a particular fold will be in the training set for a different fold. This can induce bias in the feature sets chosen. To avoid this, we repeat the experiment on the train and validation sets of fold-1. We treat the fold-1 validation set as the blind test set and generate five new sub-folds of train and validation sets from fold-1 training samples. We perform the feature selection process described above using the sub-folds. A single SVM is then trained on the DiCOVA fold-1 train set and evaluated on the corresponding validation set (blind test). The performance for different values of K and S, for SVM with linear and RBF kernels, is given in Table 2. The SVM model with RBF kernel trained using these features obtained an AUC of 0.768, and the SVM with RBF kernel is better than the linear-SVM. Results shown in Table 2 confirm the bias when the feature selection is performed jointly across the train-validation folds in a cross-validation style evaluation setup.

4. Explainability Analysis

The CNN model is a black box, associational machine that has a tendency to learn the simplest path—which might not necessarily be the correct one—to associate the input features with the target classes. Thus, it is important to understand on what basis the model is making a decision about a cough sample.

In this section, we address the explainability of the CNN model using the score-class activation maps (CAM) technique. The Score-CAM [19] is a visualization method which provides a post-hoc visual explanation for a trained CNN model by highlighting the regions in the input space that are most influential



(a) True Positive

(b) True Negative

(c) False Positive

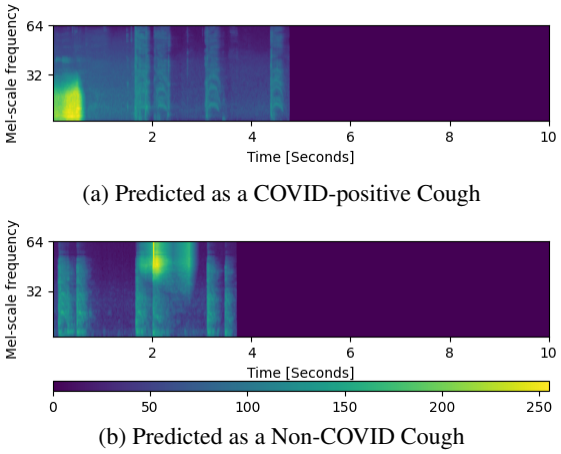
(d) False Negative

Figure 3: *Saliency maps plotted on top of the log-mel spectrograms of four different samples. The ground truth label of samples (a) and (d) is COVID positive, while samples (b) and (c) belong to the non-COVID class. The intensity of the saliency maps, shown on a scale of 0 to 255, indicates the importance of the regions of the spectrogram to the target class.*

to the prediction. In this method, activation maps are applied to different regions of the input spectrogram and the changes in the target score are observed. Then, the activation maps, weighted by the target class prediction scores, are linearly combined to form a saliency map which highlights the regions in the spectrogram where the network focuses.

Figure 3 shows the spectrograms of four different cough samples, selected from the validation set of the fold-1. The spectrograms are overlaid with the saliency maps generated by applying the Score-CAM technique to the CNN model trained on the training subset of the corresponding fold. The intensity of the saliency maps, represented on a scale of 0 to 255, shows which regions in the spectrograms are important for the target class. The first thing we notice in the plots is that the model focuses on the regions in the spectrogram which contain cough samples. This is an important observation since it implies that the decisions are mainly based on the cough samples and not other factors such as noise, breathing, or zero-padding.

A cough sound is typically composed of three phases, namely the opening burst of air expelled from the lungs, the noise-like flow of air and the closure of the vocal folds [20]. Most differences in various cough samples are typically reflected in the second phase of a cough [21]. For example, Smith et al. [22] reported a variety of differences in the duration and the pattern of the second phase of cough samples with and without mucus. Another study by Chatrzarrin et al. [21] showed that the major differences between the dry and wet cough samples are in the second phase of the cough samples. They showed that higher frequency bands of dry cough samples have much less energy than those of wet cough samples. Fig. 3 shows that the model focuses more on the second phase of the cough samples. Moreover, given that COVID-19 cough samples are categorized as a dry cough [1], we can observe that when a cough sample is predicted as COVID-positive, the model has focused on the lower frequency bands (Fig. 3-a), while the higher frequency regions are more influential when the model detects a non-COVID sample (Fig. 3-b). These observations are consistent with and supports previous studies.



(a) Predicted as a COVID-positive Cough

(b) Predicted as a Non-COVID Cough

Figure 4: *Saliency maps plotted on top of the log-mel spectrograms of 2 different samples selected from the blind test set. (a) yielded the highest score for the COVID-positive class, while (b) yielded the highest score for the non-COVID class. The intensity of the saliency maps, shown on a scale of 0 to 255, indicates the importance of the regions of the spectrogram to the target class.*

In Figure 4, we plot the saliency maps of two samples selected from the blind test set which received the highest prediction scores for (a) the COVID-positive class and (b) the non-COVID class. We can observe that the model is following the same trend when making predictions.

5. Conclusions

This paper describes the building, analysis and interpretation of a system aimed at diagnosing COVID-19 from acoustic cough signals. We showed that significant performance can be achieved with both a feature-engineering system using classical machine learning models as well as a modern CNN-based deep learning approach. We also showed that the combination of these disparate models can provide a modest but relevant improvement in performance. Finally, we addressed the important and often overlooked issue of model interpretation, which is especially relevant in the health-care domain. Given the promising nature of our results (which are based on 1040 samples), larger data sets are needed to train and evaluate such models and to verify if such performance holds in large-scale data sets.

6. References

[1] World Health Organization, "Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19)," 2020.

[2] 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, pp. 1–13, 2020.

[3] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker, "Cough against COVID: Evidence of COVID-19 signature in cough sounds," *arXiv preprint arXiv:2009.08790*, 2020.

[4] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8328–8332.

[5] G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen, and A. Khanzada, "Virufy: Global applicability of crowdsourced and clinical datasets for ai detection of covid-19 from cough," *arXiv preprint arXiv:2011.13320*, 2020.

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

[7] "EPFL Cough for COVID-19 Detection," https://coughvid.epfl.ch/, 2020, [Online; accessed 25-March-2021].

[8] "NYU Breathing Sounds for COVID-19," https://breatheforscience.com/, 2020, [Online; accessed 25-March-2021].

[9] "Cambridge University, UK - COVID-19 Sounds App," https://covid-19-sounds.org/en/, 2020, [Online; accessed 25-March-2021].

[10] A. Muguli, L. Pinto, R. Nirmala, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy, and V. Nanda, "DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," 2021.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[12] E. Edwards, C. Dognin, B. Bollepalli, and M. Singh, "Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech," in *Proc. Interspeech 2020*, 2020, pp. 2197–2201. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2781

[13] M. Jaiswal, S. Tabibu, and R. Bajpai, "The truth and nothing but the truth: Multimodal analysis for deception detection," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 938–943.

[14] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[15] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[16] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[17] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[19] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 24–25.

[20] W. Thorpe, M. Kurver, G. King, and C. Salome, "Acoustic analysis of cough," in *The Seventh Australian and New Zealand Intelligent Information Systems Conference, 2001*. IEEE, 2001, pp. 391–394.

[21] H. Chatrzarrin, A. Arcelus, R. Goubran, and F. Knoefel, "Feature extraction for the differentiation of dry and wet cough sounds," in *2011 IEEE international symposium on medical measurements and applications*. IEEE, 2011, pp. 162–166.

[22] J. A. Smith, H. L. Ashurst, S. Jack, A. A. Woodcock, and J. E. Earis, "The description of cough sounds by healthcare professionals," *Cough*, vol. 2, no. 1, pp. 1–9, 2006.