Classification of COVID-19 from Cough Using Autoregressive Predictive Coding Pretraining and Spectral Data Augmentation

John Harvill[1], Yash R. Wani[2], Mark Hasegawa-Johnson[1], Narendra Ahuja[1], David Beiser[2], David Chestek[3]

[1]University of Illinois at Urbana-Champaign
[2]University of Chicago
[3]University of Illinois at Chicago

harvill2@illinois.edu, yashwani@uchicago.edu, jhasegaw@illinois.edu,
n-ahuja@illinois.edu, dbeiser@medicine.bsd.uchicago.edu, dchest2@uic.edu

Abstract

Serum and saliva-based testing methods have been crucial to slowing the COVID-19 pandemic, yet have been limited by slow throughput and cost. A system able to determine COVID-19 status from cough sounds alone would provide a low cost, rapid, and remote alternative to current testing methods. We explore the applicability of recent techniques such as pre-training and spectral augmentation in improving the performance of a neural cough classification system. We use Autoregressive Predictive Coding (APC) to pre-train a unidirectional LSTM on the COUGHVID dataset. We then generate our final model by fine-tuning added BLSTM layers on the DiCOVA challenge dataset. We perform various ablation studies to see how each component impacts performance and improves generalization with a small dataset. Our final system achieves an AUC of 85.35 and places third out of 29 entries in the DiCOVA challenge.

Index Terms: COVID-19 classification, cough, pretraining, data augmentation, DiCOVA

1. Introduction

Coronavirus Disease of 2019 (COVID-19) caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2) virus has led to a sudden and dramatic loss of human life since its identification in December 2019. The virus has spread swiftly around the globe and as of March 2021, officials have confirmed over 2.6 million deaths and 116 million cases [1].

A strategy that has effectively slowed the spread of COVID-19 has been physically distancing and isolating infected individuals. The self isolation strategy's success has heavily depended on the rapid and accurate diagnosis of COVID-19. Unfortunately, the current gold standards for early diagnosis, viral and serology tests, can be expensive and hard to scale due to a scarcity of personnel to administer the tests and slow throughput [2, 3]. Artificial intelligence and machine learning methods have jumped to conquer the many challenges presented by the virus, especially the rapid and accurate diagnosis of COVID-19. A variety of data has been used by the machine learning community for this purpose, such as CT scans, X-rays, and laboratory features [4, 5, 6]. However, many of the types of data collected can only be obtained using large and expensive equipment, and would not be effective at diagnosing patients before a visit to the hospital. In contrast, an AI-based diagnostic test of cough sounds could be used to efficiently and remotely screen individuals almost instantly at virtually no cost.

In addition to efficiency and accessibility, pursuit of a cough-based diagnostic system might be worthwhile for two key reasons. The first is that past studies have shown coughs from patients with different respiratory syndromes to exhibit acoustic differences based on a variety of factors including nature of disease, location, and type of irritant [7, 8]. Furthermore, a study by Smith et al. showed that health professionals performed poorly at identifying a clinical diagnosis from cough sounds, obtaining a correct diagnosis only 34% of the time [9]. Therefore, a machine learning model capable of differentiating between disease-specific latent acoustic features would be able to accurately diagnose COVID-19 from cough sounds alone and support clinicians' assessments of patients. The second reason is that behind the presence of a fever, a cough is the most prevalent symptom of COVID-19, occurring in 57% of patients [10]. Additionally, the COVID-19 virus predominantly spreads through airborne means such as the forceful expulsion of mucosalivary droplets in coughs [11]. Thus, identifying coughing individuals infected by COVID-19 for isolation would serve to directly slow the spread of the virus.

In this paper we describe our approach for the DiCOVA COVID-19 classification challenge [12]. We demonstrate the effectiveness of using autoregressive predictive coding [13] as a pretraining technique for classification of COVID-19 from coughing. Due to the limited number of data samples in the DiCOVA challenge dataset, we find data augmentation and pretraining to be critical to improved generalization performance on both validation data and blind test data. We demonstrate the effectiveness of the popular spectrogram augmentation technique proposed by Park et al. [14] and show that always augmenting the data leads to the best performance on test data.

2. Related Work

Alongside the dramatic rise in translational research focused on diagnosing COVID-19 from saliva, blood, and other contact-based sources, researchers have recently been looking to non-contact methods for COVID-19 classification. A healthy body of research in audio-based cough classification has paved the way for COVID-19 diagnostics based on cough sounds alone. Within this growing field, several acoustic features from cough sounds have been proposed for the classification of respiratory illnesses. For example, Abeyratne et al. [15] train logistic regression models on time series statistics, formant-frequency tracking, and general time-frequency representations to differentiate between bronchitis coughs, asthma coughs, and pneumonia coughs. The productivity of coughs, a higher level assessment that aids clinicians in a differential diagnosis, had also been explored by Swarnkar et al. in [16] using a variety of sig-

nal processing features including bispectrum score, formant frequencies, log energy, kurtosis, and mel-frequency cepstral coefficients (MFCC). The features were passed to a logistic regression model to classify wet vs dry coughs.

Additionally, a variety of deep learning architectures have been used on different configurations of cough sounds, including original time series audio data, processed time series data, and time-frequency representations. For example, MFCC and spectrogram images were used to train convolutional neural networks (CNN) to classify cough audio into COVID and non-COVID categories by Bansal et al. [17], achieving a 71% and 81% test accuracy and sensitivity, respectively. Imran et al. [18] train a tri-pronged classifier using a dataset of 48 COVID-19, 102 bronchitis, 131 pertussis, and 76 normal cough sounds. The classification algorithm proposed requires three independent support vector machine (SVM) or CNN classifiers using mel spectrograms or MFCCs as input features to agree before making a final classification. Pahar et al. [19] found that a Resnet50 CNN architecture best discriminated between COVID-19 positive and healthy coughs while an LSTM classifier was better able to discriminate between COVID-19 positive and COVID-19 negative coughs in ill patients. Laguarta et al. [20] make use of several pre-trained feature extraction systems, which they refer to as biomarkers, to classify the presence or absence of COVID-19 from cough samples. Biomarkers include muscular degradation, vocal cords, sentiment, and lungs and respiratory tract.

3. Data

We pretrained our system on the COUGHVID dataset, a collection of over 20,000 crowd-sourced cough recordings [21]. Cough audio data and relevant metadata were collected from participants using a web interface. Metadata collected included age, gender, presence of respiratory condition, presence of fever or muscle pain, and COVID status. The collected data was filtered using a cough detection algorithm, and the surviving samples were subsequently annotated by experts. Expert annotations included assessments of the quality of cough recording, cough type (wet or dry), dyspnea, wheezing, stridor, choking, congestion, diagnosis, and severity. Cough audio recordings were sampled at 48 kHz and all continents were represented in the subject data. Our system used these data for unsupervised pre-training only; none of the available metadata or expert annotations were used. 23% of the collected samples were either COVID-19 symptomatic or COVID-19 positive.

Our system was fine-tuned on the DiCOVA cough dataset, provided through the DiCOVA Challenge [12]. Data was recorded through a web interface where more than 90% of the audio clips were recorded at a sampling rate of 48 kHz. The collected audio was converted to 44.1kHz and distributed in the FLAC format. All audio samples were manually annotated for type of cough (shallow or heavy). The provided dataset consisted of 1040 cough audio files recorded from unique subjects, with 75 COVID-19 positive samples and 965 non COVID-19 samples. Metadata included COVID status, gender, and nationality (Indian or Other). The dataset was split into five folds such that all samples are held out for validation at least once. The blind test data consisted of 233 raw cough audio files with no attached metadata. The blind test data was used for leaderboard rankings for the challenge and can be found at `https://competitions.codalab.org/competitions/29640`.

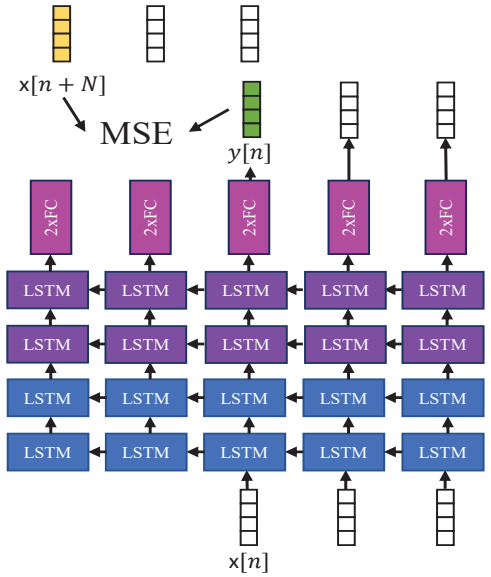

Figure 1: *Unsupervised pre-training: A four-layer LSTM was trained to minimize* $\|y[n] - x[n+N]\|^2$ *(adaptive predictive coding), then the upper layers (purple) were discarded, and the lower layers (blue) were copied into the fine-tuning network shown in Fig. 2.*

3.1. Preprocessing

We first downsampled all audio recordings to 16kHz. We then computed 80-dimensional Mel log spectrograms using a window of 1024 samples and a hop length of 160 samples (10ms). The spectrograms were normalized by setting any components less than or equal to -120dB to -120dB. Then we normalized the spectrogram such that the minimum value is 0 (corresponding to -120dB) and the maximum value is 1 (corresponding to 0dB).

4. Baselines

Muguli et al. [12] provide three baseline system implementations for the DiCOVA challenge. Feature vectors are composed of 39-dimensional mel-frequency cepstral coefficients (MFCC) plus delta and delta-delta coefficients using window size 1024 and hop length 441. Models are trained at the frame level, and the probability of an audio sample being COVID-19 positive is the mean of the probabilities of all frames in the audio. The three baseline methods are described briefly below.

Linear Regression: Classifier is trained for a maximum of 25 iterations with `liblinear` optimizer, regularization strength of 0.01 and l_2 penalty. **Multi-layer Perceptron**: Classifier is composed of one layer of 25 hidden units with the `tanh` nonlinearity applied to the output. l_2 regularization is used with weight 0.001. Examples are sampled during training such that the model is equally exposed to positive and negative samples. **Random Forest**: Classifier uses 50 trees and Gini impurity.

5. Method

Considering the small sample size of the DiCOVA dataset, we found it critical to avoid overfitting in order to improve test performance. We focused on two main methods to avoid this issue: pre-training and data augmentation. We pre-trained on all samples in COUGHVID, a dataset of cough sounds, because we

wanted our system to have a general understanding of cough structures without overfitting the DiCOVA training data. We then proceeded to fine-tune on the augmented DiCOVA data.

5.1. Autoregressive Predictive Coding

Pretraining has shown incredible success in natural language processing with the advent of BERT [22]. Inspired by this success, Chung et al. [13] adapted pre-training to audio in a way similar to BERT. The key similarity is that the pre-training is predictive. No external labels are required, but rather structure is learned by trying to fill in missing parts of the original signal. Chung et al. explore two types of pre-training: (1) autoregressive predictive coding (APC) and (2) contrastive predictive coding (CPC). The authors found superior performance of APC over CPC for both phone classification and speaker verification tasks, so we chose to implement APC.

APC is a simple yet effective form of pre-training for audio. The objective is for the model to predict a future spectral frame given previous frames. If the goal is to predict the N'th future frame, the error for any audio clip becomes:

$$E = \sum_{n=1}^{T-N} \|y[n] - x[n+N]\|_2^2 \quad (1)$$

where y refers to the network output and x refers to acoustic input. By forcing the model to predict the future spectral frame, the underlying structure of the audio signal is learned. We used a uni-directional long short-term memory (LSTM) model for pre-training (bi-directional would break causality). We used 4 LSTM layers with a hidden size of 400 and a dropout of 0.1. Then we used two linear layers with 500 nodes. We applied the hyperbolic tangent nonlinearity after the first linear layer. The output dimension of the model is the same as that of the input since we compute MSE loss. See Figure 1 for a visual description. We split the 4 LSTM layers into "upper" and "lower" layers because we used the output of the lower layers as extracted features during fine-tuning, discussed next.

5.2. Fine-tuning

We passed the output from the lower layers of the APC model as input features to our classification network. We used a network composed of 2 bi-directional long short-term memory (BLSTM) layers followed by three fully-connected layers. The forward and backward summaries were taken from the BLSTM layers and concatenated before being fed through the feedforward layers. We applied dropout with $p = 0.1$ in both the BLSTM and fully-connected layers. To predict the probability of the cough sample coming from a COVID-19 positive patient, we took the softmax at the output and used the cross-entropy loss for training. Note that due to data imbalance, we weighted the loss of positive samples 15 times more than the loss of negative samples. During training, we applied SpecAugment from Park et al. [14] to the cough samples. We found that spectral augmentation was critical to generalization to the test data. See Figure 2 for a visualization of the fine-tuning model.

5.3. Ensembling

During training we validated with area-under-curve (AUC), which measures the performance of a classification system with imbalanced data better than accuracy. AUC is the area underneath the true positive rate (TPR) vs. false positive rate (FPR) curve for different classification thresholds ranging from zero to one. Since we directly optimized the cross-entropy loss and not AUC, we found AUC to vary throughout training. We hypothesized that different sets of model parameters may classify particular samples better or worse than one another, and that an ensemble of several high-performing validation checkpoints may improve test performance. We found this to be true and chose the best three validation checkpoints, taking the mean of their output probabilities for final validation scores for each fold. For the blind test data, we took the mean of the scores from each of the five folds. Thus our test predictions were an average of $5 \times 3 = 15$ model probability scores.

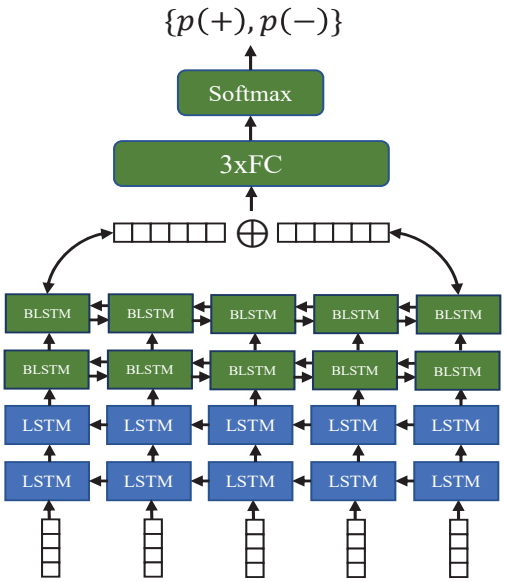

Figure 2: *Fine-tuning: The lower two LSTM layers from the APC network (Fig. 1) were frozen, while the upper BLSTM and fully-connected layers were trained to minimize cross-entropy.*

6. Experiments

Given our pre-training and fine-tuning system, there are several variables that contribute to the overall performance: (1) number of future frames N to predict during pre-training (2) layer from which to extract pre-trained features during fine-tuning (3) percentage of augmented examples seen by the fine-tuning classification network during training. We now discuss the configuration that led to our best score on the DiCOVA challenge leaderboard. We pre-trained with $N = 10$ future frames, used the output from the first 2 of 4 LSTM layers (lower layers) as output, and used spectral augmentation 100% of the time. This means that we never exposed the fine-tuning classifier to an unaltered example of the input spectrogram during training. All experiments were run once for each of the five provided folds.

7. Results

Average validation performance across the five folds for our best configuration compared to that of the three provided baselines is given in Figure 3. While we do not have access to test labels for the DiCOVA challenge, we note that we placed third out of 29 teams on the test data with an AUC of 85.35. This score was less than two points below the top AUC score of 87.07, demonstrating superior test performance over most

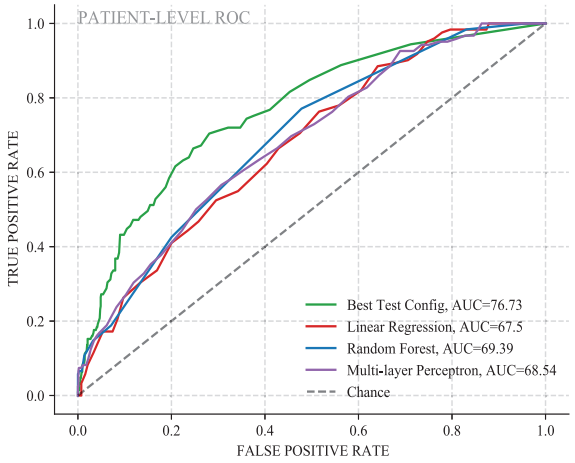

Figure 3: *Best config: We plot ROC curves for our best-performing test configuration against the three baselines provided for the DiCOVA challenge.*

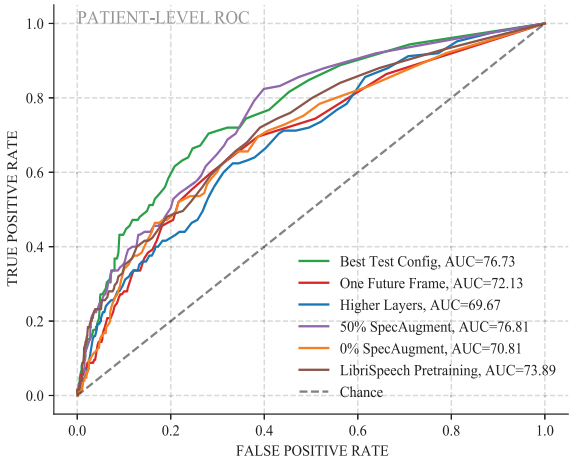

Figure 4: *Ablations: We plot ROC curves for our best-performing test configuration against five ablation methods where we change one setting compared to the best-performing test configuration.*

proposed techniques. Our method also demonstrates significant improvement over the baselines on the validation data, showing that the combination of autoregressive predictive coding and spectral augmentation is useful for determining COVID-19 status from cough sounds.

8. Ablations

To explore how each part of our proposed system influences the overall validation performance, we ran five ablation experiments. In each of the experiments, all hyperparameters stayed fixed compared to the best configuration except for the hyperparameter explicitly under investigation.

Spectral augmentation (two experiments): Perform spectral augmentation 50% or 0% of the time instead of 100% of the time. This means that for each experiment with either 50% or 0% chance, respectively, the fine-tuning model will be exposed to a random spectrally-augmented version of the spectrogram. Otherwise the model will see the unaltered input spectrogram. **Future frames**: Use $N = 1$ future frame for pre-training instead of 10. This ablation is inspired by [13], which found that APC with $N = 1$ was best for speaker verification. **Higher layers**: Use the output from the 4th layer (upper layers) of the pre-trained model. This means we take the output from the purple layers in Figure 1 instead of the blue layers and then pass that representation to the BLSTM layers in Figure 2 for fine-tuning. **Pre-training data**: Use Librispeech [23] for pre-training instead of COUGHVID. Librispeech is a collection of read English speech from audiobooks recorded at 16kHz. We use the `train-clean-100` subset of the data for pre-training.

Average validation performance across the five folds for each ablation experiment plus the best test configuration is provided in Figure 4. Notice that both pre-training hyperparameters and the amount of spectral augmentation are critical for improved validation performance. When applying APC to the cough classification task, there are two big takeaways. First, higher layer representations are not as useful as features for prediction of COVID-19 from cough as the lower layer representations. This makes sense because the higher layer representations correspond more directly to the predictive pretraining task, which inevitably produce a blurred and shifted version of the spectrogram. Second, contrary to the findings for speaker verification [13], predicting the $N = 10$th future frame is more effective than predicting the $N = 1$st future frame as pre-training for COVID-19 cough diagnosis. Also notice that pretraining with speech audio from the Librispeech dataset [23] results in slightly reduced performance compared to the best test configuration. AUC for this ablation method is still above that of all provided baselines in Figure 3. This demonstrates that while generic audio pretraining can be useful for cough classification, pretraining specifically on cough sounds is important for improved performance. This is a hopeful finding, because while it may be difficult to collect large amounts of labeled cough data for classification of the presence of COVID-19, collecting large amounts of unlabeled cough data is much more feasible. Next, notice that the use of spectral augmentation is also important for improved performance. We find very little difference in validation performance between the best test config (100% SpecAugment) and the 50% SpecAugment scenarios, but then see a large decrease when not using spectral augmentation at all (0% SpecAugment). These results demonstrate the importance of both APC pretraining and SpecAugment during finetuning for improving detection of COVID-19 from cough sounds.

9. Conclusions

We propose a novel approach for the classification of COVID-19 from coughing sounds based on APC pre-training and SpecAugment. We find that our approach is one of the top performers on the test data for the DiCOVA challenge, ranking third out of 29 entries. When studying the effect each training hyperparameter has on overall performance, we find that APC pretraining hyperparameters and SpecAugment are critical for improved performance. We also find that while having a large amount of cough data for pre-training gives the best performance, using related audio like speech can also lead to a substantial improvement over provided baselines. Overall, our approach provides more evidence that a cough-based classification system can assist in the diagnosis of COVID-19 practically instantaneously and at virtually no cost.

10. References

[1] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[2] S. N. Najmabadi and J. Root, "Coronavirus test results in texas are taking up to 10 days," Mar 2020. [Online]. Available: https://tylerpaper.com/covid-19/coronavirus-test-results-in-texas-are-taking-up-to-10-days/article_5ad6c9cc-4fa9-573a-bb4d-ada1b97acf42.html

[3] Y. C. Manabe, J. S. Sharfstein, and K. Armstrong, "The need for more and better testing for covid-19," *JAMA*, vol. 324, no. 21, pp. 2153–2154, 2020.

[4] C. Jin, W. Chen, Y. Cao, Z. Xu, Z. Tan, X. Zhang, L. Deng, C. Zheng, J. Zhou, H. Shi *et al.*, "Development and evaluation of an artificial intelligence system for covid-19 diagnosis," *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.

[5] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in biology and medicine*, vol. 121, p. 103792, 2020.

[6] F. T. Fernandes, T. A. de Oliveira, C. E. Teixeira, A. F. de Moraes Batista, G. Dalla Costa, and A. D. P. Chiavegatto Filho, "A multipurpose machine learning approach to predict covid-19 negative prognosis in são paulo, brazil," *Scientific reports*, vol. 11, no. 1, pp. 1–7, 2021.

[7] Y. Hashimoto, A. Murata, M. Mikami, S. Nakamura, E. Yamanaka, and S. Kudoh, "Influence of the rheological properties of airway mucus on cough sound generation," *Respirology*, vol. 8, no. 1, pp. 45–51, 2003.

[8] K. Dawson, C. Thorpe, and L. Toop, "The spectral analysis of cough sounds in childhood respiratory illness," *Journal of paediatrics and child health*, vol. 27, no. 1, pp. 4–6, 1991.

[9] 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.

[10] M. C. Grant, L. Geoghegan, M. Arbyn, Z. Mohammed, L. McGuinness, E. L. Clarke, and R. G. Wade, "The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (sars-cov-2; covid-19): A systematic review and meta-analysis of 148 studies from 9 countries," *PloS one*, vol. 15, no. 6, p. e0234765, 2020.

[11] L. Bourouiba, "Turbulent gas clouds and respiratory pathogen emissions: potential implications for reducing transmission of covid-19," *JAMA*, vol. 323, no. 18, pp. 1837–1838, 2020.

[12] A. Muguli, L. Pinto, N. R., N. Sharma, K. Prashant, P. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. Chetupalli, S. Ganapathy, and V. Nanda, "Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[13] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[14] 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.

[15] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Annals of biomedical engineering*, vol. 41, no. 11, pp. 2448–2462, 2013.

[16] V. Swarnkar, U. R. Abeyratne, A. B. Chang, Y. A. Amrulloh, A. Setyati, and R. Triasih, "Automatic identification of wet and dry cough in pediatric patients with respiratory diseases," *Annals of biomedical engineering*, vol. 41, no. 5, pp. 1016–1028, 2013.

[17] V. Bansal, G. Pahwa, and N. Kannan, "Cough classification for covid-19 based on audio mfcc features using convolutional neural networks," in *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 2020, pp. 604–608.

[18] 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.

[19] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "Covid-19 cough classification using machine learning and global smartphone recordings," *arXiv preprint arXiv:2012.01926*, 2020.

[20] J. Laguarta, F. Hueto, and B. Subirana, "Covid-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[21] L. Orlandic, T. Teijeiro, and D. Atienza, "The coughvid crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *arXiv preprint arXiv:2009.11644*, 2020.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.