Smart Health 26 (2022) 100332

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

An explainable COVID-19 detection system based on human sounds

Huining Li [a], Xingyu Chen [c], Xiaoye Qian [b], Huan Chen [b], Zhengxiong Li [c], Soumyadeep Bhattacharjee [a], Hanbin Zhang [a], Ming-Chun Huang [d,e], Wenyao Xu [a,*]

[a] *Department of Computer Science and Engineering, University at Buffalo, United States*
[b] *Department of Electrical, Computer, and Systems Engineering, Case Western Reserve University, United States*
[c] *Department of Computer Science and Engineering, University of Colorado Denver, United States*
[d] *Department of Data and Computational Science, Duke Kunshan University, China*
[e] *Suzhou Huanmu Intelligence Technology Co., Ltd., China*

ARTICLE INFO

Keywords:
COVID-19
Acoustic
Explainable
Accurate

ABSTRACT

Acoustic signals generated by the human body have often been used as biomarkers to diagnose and monitor diseases. As the pathogenesis of COVID-19 indicates impairments in the respiratory system, digital acoustic biomarkers of COVID-19 are under investigation. In this paper, we explore an accurate and explainable COVID-19 diagnosis approach based on human speech, cough, and breath data using the power of machine learning. We first analyze our design space considerations from the data aspect and model aspect. Then, we perform data augmentation, Mel-spectrogram transformation, and develop a deep residual architecture-based model for prediction. Experimental results show that our system outperforms the baseline, with the ROC-AUC result increased by 5.47%. Finally, we perform an interpretation analysis based on the visualization of the activation map to further validate the model.

1. Introduction

At face-to-face clinical visits, clinicians often leverage human body sounds (e.g., breathing, heart, digestion) to diagnose disease or evaluate disease progression, such as respiratory disease, Parkinson's disease (Erdogdu Sakar, Serbes, & Sakar, 2017; Pramono, Bowyer, & Rodriguez-Villegas, 2017; Zhang et al., 2019). With the development of mobile health technology, researchers investigate digital biomarkers from these human sounds collected by microphones for facilitating telemedicine (Baghai-Ravary & Beet, 2017; Cummins et al., 2015).

As the pathogenesis of COVID-19 is increasingly indicating impairments in the respiratory system, recent studies start to explore the digital biomarker from respiratory sounds (e.g., coughs, breathing and voice) to assist the detection of COVID-19 in an effortless and remote manner. Huang et al. (2020) used a remote electronic stethoscope to capture the lung auscultation characteristics as the indicator of COVID-19. Imran et al. (2020) developed a smartphone app to collect cough sounds for studying the difference of pathomorphological changes in the respiratory system caused by COVID-19 infection and other respiratory infections. Brown et al. (2020) developed an automatic diagnosis of COVID-19 approach based on crowdsourced breath and cough data. However, these works only scratch the surface of the potential of such respiratory sound data, rather than providing insightful explanation on how these data contribute to COVID-19 diagnosis.

[*] Corresponding author.
E-mail addresses: huiningl@buffalo.edu (H. Li), XINGYU.CHEN@UCDENVER.EDU (X. Chen), xxq82@case.edu (X. Qian), hxc556@case.edu (H. Chen), ZHENGXIONG.LI@UCDENVER.EDU (Z. Li), sbhattac@buffalo.edu (S. Bhattacharjee), hanbinzh@buffalo.edu (H. Zhang), mh596@duke.edu (M.-C. Huang), wenyaoxu@buffalo.edu (W. Xu).

https://doi.org/10.1016/j.smhl.2022.100332
Received 29 September 2022; Accepted 29 September 2022
Available online 19 October 2022

In this paper, we develop an explainable automatic COVID-19 detection system based on speech, cough, and breath data. We first summarize our design space considerations from the aspect of data size, data dimension, and model. A preliminary feature study is also conducted to investigate the data variance between COVID subjects and healthy subjects based on our dataset. After gaining the basic domain knowledge, we first apply a set of data augmentation techniques including time shifting, pitch shifting, and Gaussian noise injection, to promote the variance of the data during training phase. Then, we transform the 1D audio data to 2D Mel-spectrogram for preserving pathological information with a high-resolution. Next, these Mel-spectrograms are fed to a deep residual architecture for the automatic features selection and the COVID risk estimation. Finally, a feed-forward network is applied to fuse the results from Cough Net, Speech Net, and Breath Net to make the final decision.

To evaluate our system, we construct a baseline that implements a respiratory sound-based COVID diagnosis work (Brown et al., 2020) on our dataset. The baseline is based on a machine learning classifier (e.g., SVM) with handcrafted features and learnable features as input. ROC-AUC results show that our work can outperform the baseline with a single breath input, a single speech input, and a fusion input. To validate the interpretability of our model, we further visualize the activation map and identify that the regions of formants, some unvoiced phonemes, and inhales after unvoiced signals can contribute to the COVID prediction. The visualization of the model's regions of interest can provide more insights and explainability for researchers and clinicians to investigate the correlation between human acoustic characteristics and COVID-19 onset, specifically for COVID patients with no symptoms. It also helps explore a brand new vocal biomarker, which can identify the difference between respiratory diseases and COVID-19. To summarize, it boosts the transition from existing expensive molecular testing methods to non-contact, effortless, and fast testing schemes.

To summarize, our contributions are three-fold:

- We investigate a set of fundamental theories to select the most suitable feature design space and model design space for disease detection.
- We develop an explainable COVID-19 detection system based on speech, cough, and breath, which consists of data augmentation, Mel-spectrogram transformation, and a residual deep learning network.
- We evaluate our system on 172 COVID-19 subjects and 793 healthy subjects, and achieve above 70% ROC-AUC result. An interpretation analysis is performed to further validate our model.

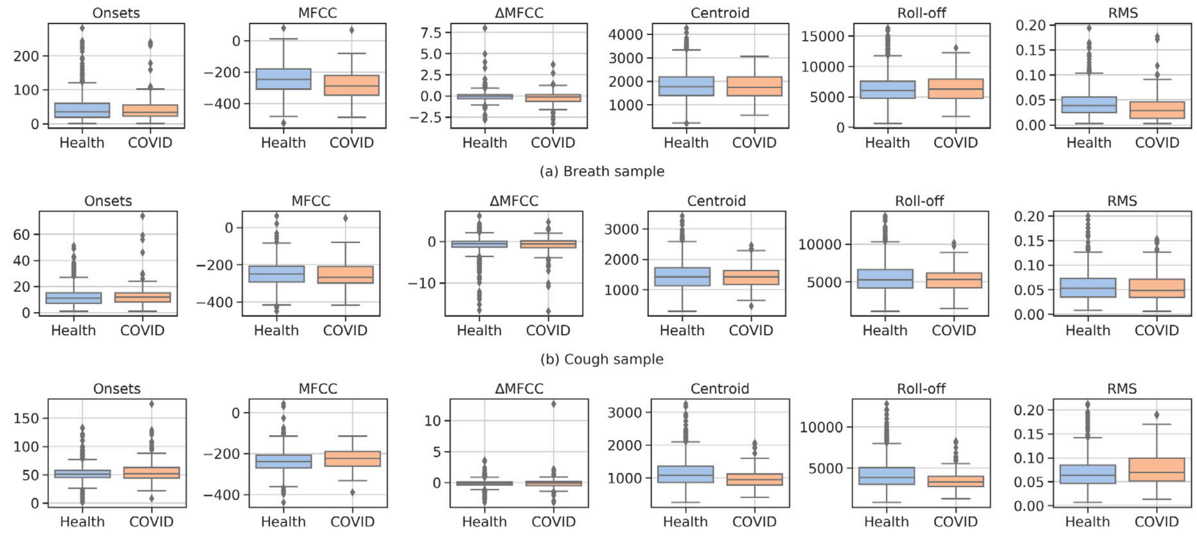
2. Design space consideration

2.1. Data consideration

Data size. The key bottleneck to developing the audio-based classification model using deep learning technology is limited labeled data. Recent studies have shown that increasing the amount of data can promote the generalization ability as well as the model performance (Wen et al., 2020). Data augmentation is an effective tool to increase the size and improve the quality of the training data, and solve the imbalanced class issue. The basic idea of data augmentation is to generate a synthetic dataset that can cover unexplored input space while keeping correct labels. To design a suitable audio data augmentation approach, we need to consider: (1) the intrinsic properties of time series audio data, such as temporal dependency; (2) the data augmentation methods should be task-dependent. For example, the data augmentation methods effective for classification problems may not be valid for time series anomaly detection.

Data dimension. How to effectively represent the data for feeding to the machine learning model is a critical issue. We need to consider the representation dimension and the information in each dimension. For the time series audio data, it can be abstracted in the time domain, frequency domain, or both.

Preliminary feature study. We conduct a preliminary study to investigate the data variance between COVID subjects and healthy subjects based on our dataset. A set of handcrafted features (Brown et al., 2020) is extracted from each cough, breath, and speech sample, including pitch onset, root-mean-square (RMS) energy, spectral centroid, roll-off frequency, Mel-Frequency Cepstral Coefficients (MFCC), ΔMFCC. Specifically, pitch onset characterizes the peak from a pseudo syllable strength envelope, which can be achieved by adding the positive first-order difference across each Mel band. RMS energy is the root-mean-square of the spectral magnitude of a short-time Fourier transform. The spectral centroid locates large peaks corresponding to formants' positions and pitch frequencies and is related to the sound brightness. The roll-off frequency is the center frequency for a spectrogram bin. MFCC can characterize the shape of the vocal tract, which manifests in the envelope of the short-time power spectrum. As shown in Fig. 1, most features show a slight difference between COVID subjects and healthy subjects on speech samples except for onset and ΔMFCC. However, for breath samples and cough samples, most feature distributions are nearly the same across healthy subjects and COVID subjects, only MFCC displays a slight difference on breath samples, and only spectral centroid has a minor difference on cough samples. To summarize, we can obtain that *pathological information are not well preserved in these hand-crafting features, especially for breath and cough signals in our dataset.*



(a) Breath sample

(b) Cough sample

(c) Speech sample

Fig. 1. The comparison of the mean features of breath, cough, and speech across healthy subjects an COVID subjects.

2.2. Model consideration

Machine learning technology is boosting various applications nowadays. To select the most appropriate machine learning model for a specific application, we need to consider: (1) Input representation type. For example, 1-D input and 2-D input should choose different network architectures; (2) Generalization ability. The labeled data collected in real-world applications are often limited. How to design a network architecture that can be trained using small amounts of data without overfitting is an ongoing problem. (3) Explainability. Most machine learning models are black boxes, which may influence the model extendability in similar applications. Therefore, model explainability is required for the design.

3. Methodology

3.1. Data augmentation

We first apply a set of augmentation techniques on the time series audio data of each input batch in the training process. These can increase the variance of the data during training to build a more generalized and robust model.

Although the dataset has a large amount of cough, breath, and speech data from COVID-19 patients, it still has an unbalanced issue. Therefore, we intentionally over-sample the audio samples of COVID-19 subjects and healthy subjects following a rate of 7:1 in the training process. We combine the augmentation operations for each sample, which are summarized as follows:

Shifting. A recent study shows that pitch shifting and time shifting are effective in cough data augmentation for training a cough detection model (Xu et al., 2021). Inspired by this work, we apply basic pitch shifting and time shifting on the raw audio signals. Specifically, time shifting is set as a rolling basis from −0.5 s to +0.5 s to emulate different signal window positions in the audio stream. In the same while, we lower or raise the pitch of the audio sample while maintaining the duration unchanged. The pitch of each sample is randomly shifted from −4 semitones to 4 semitones.

Noise Injection. Noise injection is a method that injects a small amount of noise/outlier into time series signals without changing the corresponding labels. To reduce the generalization error and improve the structure of the mapping problem, we choose to add random noise in time-series audio data. It is intuitive to expect that noise might degrade the model performance in the training process. In contrast, recent work has suggested that training a model with noise is equivalent to a form of regularization by adding an extra term to the loss function (Bishop, 1995), which can reduce the generalization error. Moreover, noise injection to raw data can enlarge the training dataset size and is proven to be an effective augmentation approach for audio data. When the training sample is fed to the model, random noise is injected into the input variables which makes them different each time but the labels are unchanged. Specifically, we add Gaussian noise to the audio signal with the amplitude ranging from 0.001 to 0.015, and SNR ranging from 0 to 35 dB, while maintaining the label unchanged.

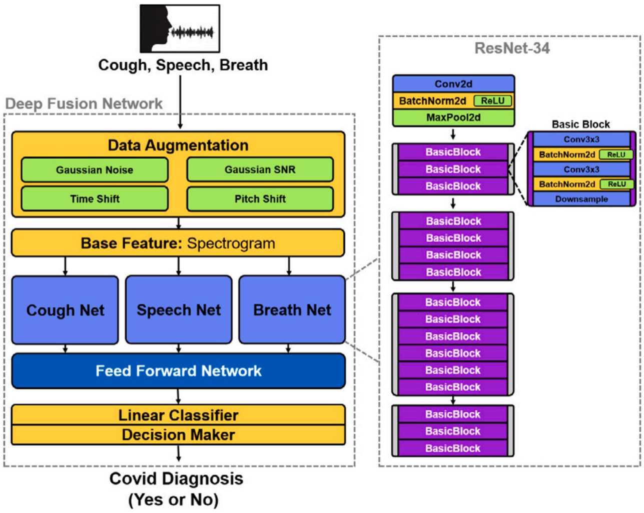


Fig. 2. System overview. Our system consists of feature extraction and SoundNet.

3.2. Mel-spectrogram representation

Challenge: Intuitively, Fast Fourier Transform (FFT) is a useful tool to investigate the frequency properties of audio signals. However, breath and cough sounds are usually non-stationary and thereby an FFT operation is not able to reflect their time-domain characteristics.

Solution & Insights: To solve the above challenge, we partition the human sounds into multiple segments while assuming that the signal in every segment is stationary, and then the FFT is performed on every segment. After that, we can present the data in both time domain and frequency domain. To avoid the truncation effect, we adopt the window function on each segment to reduce the spectrum leakage and enhance the spectral resolution. A recent study has shown that an appropriate window function can effectively prevent the implicit pathological information from the high-frequency interference (Zhang et al., 2019).

Specifically, the augmented human sounds are segmented into a series of small Hanning windows with a length of 25 ms for each and overlapping with 60%. The Hanning window function is formulated as:

$$w[n] = \frac{1 - \cos 2\pi \cdot \frac{n}{N-1}}{2}. \tag{1}$$

After segmentation, 1-D audio signals in these small windows can be transformed to 2-D spectrograms by:

$$\text{Spectrogram}\{x(t)\}(m,\omega) = |\sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}|^2. \tag{2}$$

Since Mel frequency can accurately represent the envelope of the short-time power spectrum associated with the shape of the vocal tract, we map the spectral magnitudes of short-time Fourier transform onto the Mel-scale using the filterbank technology for Mel-spectrogram generation. The mapping from y-axis (frequency) to the Mel frequency is given by (Ittichaichareon, Suksri, & Yingthawornsuk, 2012):

$$Mel(f) = 2595 \log_{10}^{(1+\frac{f}{100})}. \tag{3}$$

The Mel-spectrogram reflects abundant correlation and pattern information, e.g., formants, sound intensity, which are prone to be extracted by a deep learning model. We detail the SoundNet in the next subsection.

3.3. SoundNet

Challenge: CNN is a useful network for image classification problems. However, it needs a large dataset that contains tens of thousands of samples for training to overcome the overfitting issue. Therefore, it is challenging to develop a deep neural network with a small size dataset.

Solution & Insights: We leverage residual architecture (as shown in Fig. 2) in two aspects. (1) *Prevent overfitting.* Firstly, compared with plain networks, the residual architecture does not have extra fully connected layers, which results in few parameters. Secondly, the batch normalization layers in the residual network mitigate overfitting. This is because the normalization operation makes the

Table 1
The performance comparison between our work and baseline.

	Data type	Set 0	Set 1	Set 2	Set 3	Set 4	Average	Δ
Baseline	Cough + Breath + Speech	0.7231	0.6451	0.6572	0.6332	0.6268	0.66	
Our Work	Cough	0.6002	0.6181	0.5509	0.6876	0.4238	0.5761	−8.39%
	Breath	0.4479	0.6955	0.7483	0.7558	0.7683	**0.6831**	**2.31%**
	Speech	0.66	0.8727	0.8121	0.7162	0.7417	**0.7598**	**9.98%**
	Cough + Breath + Speech	0.6347	0.6955	0.8269	0.7534	0.6228	**0.7047**	**4.47%**

mean and variance values slightly different from one another on each mini-batch, which can be regarded as injecting noise to every hidden layer's activations and thereby resulting in a slight regularization effect. (2) *Accelerate convergence (He, Zhang, Ren, & Sun, 2016)*. The shortcut connections perform identity mapping by adding their outputs to the outputs of the stacked layers. In this way, the network is able to convey the higher-level understanding of the last layers to the previous layers, which can re-modulate how to understand the input in the training phase.

Implementation: Each type of human sound is labeled by users themselves, and then they are fed to the corresponding Cough Net, Speech Net, and Breath Net, which are based on the 34-layer residual architecture. Each type of Net contains an input block, and [3, 4, 6, 3] residual blocks. Each residual block has 2 convolutional layers, batch normalization is right after each convolution, and the shortcut connection and downsampling are adopted between the residual blocks. Finally, we apply a feed forward network to fuse the outputs from Cough Net, Speech Net, and Breath Net, and get the decision afterward.

4. Evaluation

4.1. Data preparation

In the dataset (Sharma et al., 2020), 172 subjects are COVID-19 positive and 793 subjects are COVID-19 negative. These negative subjects are completely healthy, have respiratory ailments (such as tuberculosis, pneumonia, and chronic lung disease) or COVID-19 like symptoms, e.g., cough, fever, etc. They are in the age group of 15–90 years, and most of them fall in the age of 15–40 years. Among them, 242 subjects are female and the rest are male. Each subject has three audio files, which are speech file, breath file, and cough file. The speech file is number counting from 0 to 20.

Each file longer than 3 s is regarded as a sample. If less than 3 s, we repeat it until it reaches 3 s. We divide all subjects into five sets and perform a 5-fold cross-validation. In each training process, we augment the COVID samples and healthy samples following a rate of 7:1, and then we obtain around 1100 COVID samples and 1200 healthy samples of speech, breath, cough, respectively, for each round. In each test process, we do not augment the samples, and get around 40 COVID samples and 300 healthy samples of speech, breath, cough, respectively, for each round.

4.2. Software Implementation

We implement our Resnet-34 network in PyTorch which is pre-trained with ImageNet (Deng et al., 2009). We adopt Adam optimizer with a mini-batch size of 64. The initial learning rate is set as 0.01.

4.3. Baseline

To objectively evaluate our system performance, we construct a baseline that applies a state-of-the-art respiratory sound-based COVID detection model (Brown et al., 2020) on our dataset. This work develops a simple machine learning classifier (e.g., SVM) with various features (handcrafted and obtained through transfer learning) as input. We repeat this work based on their public available model and perform 5-fold cross-validation on our dataset. Since the baseline is implemented on our dataset, the data quality is controlled, the methodology can be compared in a more fair manner.

4.4. Results

We use the ROC-AUC result as the metric to evaluate the COVID detection performance. As shown in Table 1, our system can achieve above 70% ROC-AUC result. It outperforms the baseline with a single breath input, a single speech input, and a fusion input, where the ROC results are increased by 2.31%, 9.98%, and 4.47%, respectively. Although the system achieves the highest ROC-AUC for speech, the standard deviation of recall and precision is more than 3% higher with single speech as input than those with fusion as input. Therefore, our system requires the end-user to provide all three types of sound to conduct inference.

We further evaluate the impact of gender on system performance because the articulatory organs are different male and female. As observed in Fig. 3, with a single speech input, the detection performance for females is slightly higher (less than 2%). Overall, our system is not biased on gender.

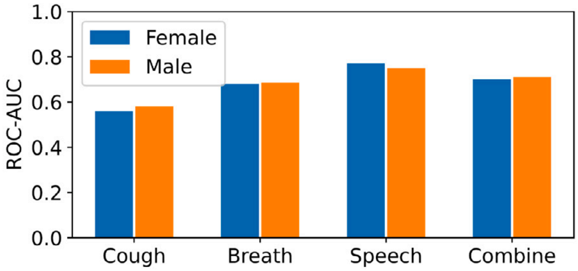
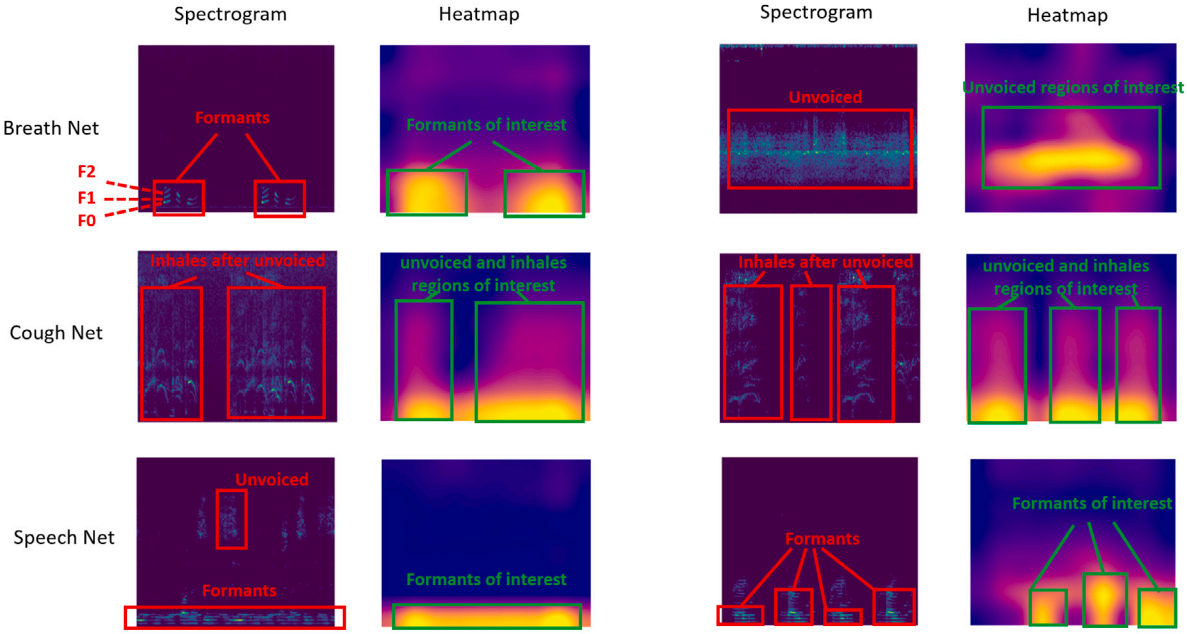


Fig. 3. The impact of gender on system performance.



(a) COVID subject (b) Healthy subject

Fig. 4. Activation map visualization for model interpretation.

5. Interpretability validation and analysis

5.1. Implementation

To understand which time–frequency region of the spectrogram contributes to the final COVID prediction the most, we can apply the weights to each element of each final convolution channel and then average along channels to generate an activation map. The regions with the highest value contributed the most. We then overlay the activation map with the original Mel-spectrogram to visualize.

5.2. Interpretation analysis

Speech Net. The Speech Net captures subjects' formants and pitch information for prediction, whereas ignoring a few unvoiced information, as shown in Fig. 4. The formant frequencies are due to the frequency shaping of vocal source signals by the vocal tract, which involves multiple articulators' cooperation. Thereby, it is likely to contain pathological information and become the focus of the Speech Net.

Cough Net. As observed in Fig. 4, the periodic inhales and unvoiced regions contribute most to the prediction when the subject is continuously performing cough behavior. It is visible that the inhale of a COVID subject following a cough is not clear compared with healthy subjects, which is learned by the Cough Net consequently.

Breath Net. In the Breath Net, our model is interested in both formants and the unvoiced regions as shown in Fig. 4. This is because the strength, rate, and pitch frequency of inhaling and exhaling differ a lot across subjects, which may hide pathological information.

To summarize, our proposed deep residual architecture is explainable and can effectively learn pathological information from Mel-spectrogram for COVID prediction. The interpretation analysis can help quantify pathological information and build a new vocal biomarker for COVID-19, which can be distinguished from other respiratory diseases.

6. Discussion

Noise Interference. When the user speaks, the generated sound wave propagates in the air media, which is prone to interference from ambient noises. The noise makes our system insensitive to minute changes in human acoustic sounds. We find that the performance of breath-based COVID detection decreases more notably (around 13.2%) than that of cough-based and speech-based models. In the future, we plan to apply software-based noise reduction technology to improve the resilience of our system.

Neural Network Size. We examine the COVID detection performance on a relatively smaller and larger neural network than our current network model. We observe that the performance of ResNet-18 is around 8.7% lower than our model. When the network size is increased to ResNet-50, the performance is almost not changed compared to our current model. This is because our training data size is not large enough to learn the parameters of a deeper network.

Integration of Health History. Since respiratory ailments (such as pneumonia, and chronic lung disease) share certain similar symptoms as COVID-19, the false alarm rate is slightly higher among the subjects with respiratory disease than that of completely healthy people. Self-report of health history can help our system make a more accurate prediction. In the future, we plan to add the self-reporting function to our mobile COVID screening app.

7. Conclusion

In this paper, we are the first to summarize the design space consideration for audio-based classification model in the aspects of data and model. Based on it, we first leverage task-dependent data augmentation techniques to increase the variance of training data, and then perform frequency and time domain modulation to transform the 1D audio data to 2D Mel-spectrogram for preserving pathological information with a high-resolution. After that, the extracted Mel-spectrograms are fed to ResNet-34 for pathological features selection and the COVID risk estimation. Finally, a feed-forward network is adopted to combine the results from Cough Net, Speech Net, and Breath Net to refine the final prediction result. Evaluation shows that our system's ROC-AUC result is 5.47% higher than the baseline. Furthermore, we perform interpretation analysis based on the visualization of the activation map to identify which time–frequency region of the spectrogram contributes to the final COVID prediction the most. Our exploration can provide more insights and explainability for researchers and clinicians to investigate acoustic modality for disease detection, not limited to respiratory disease.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the Coswara dataset. This work was supported by the National Science Foundation, USA under grant No. CNS-2050910, and the 2021–2022 Kunshan Government Research Fund (21KKSGR050). The funders had no role in the study design; data collection, analysis, or interpretation; in the writing of the report; or in the decision to submit the article for publication.

References

Baghai-Ravary, L., & Beet, S. W. (2017). VoiScan: Telephone voice analysis for health and biometric applications. In *International conference on speech and computer* (pp. 799–808). Springer.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural Computation*, *7*(1), 108–116.

Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., et al. (2020). Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. arXiv preprint arXiv:2006.05919.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Erdogdu Sakar, B., Serbes, G., & Sakar, C. O. (2017). Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PLoS One*, *12*(8), Article e0182428.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, Y., Meng, S., Zhang, Y., Wu, S., Zhang, Y., Zhang, Y., et al. (2020). The respiratory sound features of COVID-19 patients fill gaps between clinical data and screening methods. MedRxiv.

Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, M. S., Ali, K., et al. (2020). AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics in Medicine Unlocked*, *20*, Article 100378.

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). Speech recognition using MFCC. In *International conference on computer graphics, simulation and modeling* (pp. 135–138).

Pramono, R. X. A., Bowyer, S., & Rodriguez-Villegas, E. (2017). Automatic adventitious respiratory sound analysis: A systematic review. *PLoS One*, *12*(5), Article e0177926.

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Ghosh, P. K., et al. (2020). Coswara–a database of breathing, cough, and voice sounds for COVID-19 diagnosis. arXiv preprint arXiv:2005.10548.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., et al. (2020). Time series data augmentation for deep learning: A survey. arXiv preprint arXiv:2002.12478.

Xu, X., Nemati, E., Vatanparvar, K., Nathan, V., Ahmed, T., Rahman, M. M., et al. (2021). Listen2Cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(1), 1–22.

Zhang, H., Song, C., Wang, A., Xu, C., Li, D., & Xu, W. (2019). Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds. In *The 25th annual international conference on mobile computing and networking* (pp. 1–16).