Exploring Semi-supervised Learning for Audio-based COVID-19 Detection using FixMatch

Ting Dang[1‡], *Thomas Quinnell*[1‡], *Cecilia Mascolo*[1]

[1]Department of Computer Science and Technology, University of Cambridge, UK
td464@cam.ac.uk, tq215@cam.ac.uk, cm542@cam.ac.uk

Abstract

While there has been recent success in audio-based COVID-19 detection, challenges still exist in developing more reliable and generalised models due to the limited amount of high quality labelled audio recordings. With a substantial amount of unlabelled audio recordings available, exploring semi-supervised learning (SSL) may benefit COVID-19 detection by incorporating this extra data. In this paper, we propose a SSL framework which adjusted FixMatch, one of the most advanced SSL approaches, to audio signals and explored its effectiveness in COVID-19 detection. The proposed framework is validated with a crowd-sourced audio database collected from our app, and showed superior performance over supervised models with a maximum of 7.2% relative improvement. Furthermore, we demonstrated that the proposed framework significantly benefits model development using imbalanced datasets, which is a common challenge in clinical data. It can also improve model generalisation. This potentially paves a new pathway of utilising unlabelled data effectively to build more accurate and reliable COVID-19 detection tools.

Index Terms: COVID-19 detection, audio, semi-supervised learning, Fixmatch, VGGish

1. Introduction

The outbreak of COVID-19 in 2020 has caused considerable socioeconomic impact and threatened human life. Policy responses, vaccine development, and effective test tools have greatly reduced the spread and brought the pandemic under control. While the most commonly used test tools for COVID-19 detection such as polymerase chain reaction (PCR) tests [1, 2] and lateral flow device antigen (LFD) tests [3] are effective, digital technologies that employ machine learning using different biomarkers also demonstrated great potential for scalable, flexible and fast detection.

Extensive attention has been paid to using audio biomarkers for COVID-19 detection, such as cough and speech, due to its numerous advantages (e.g. flexibility in data collection and convenience in a home monitoring context). A variety of studies have demonstrated its potential in detecting COVID-19 infections using deep learning techniques [4, 5, 6, 7, 8]. However, most of the work is validated in a relative small dataset [4, 8, 6], which may struggle to generalise and cannot be employed for unseen data. The extensive amount of annotated audio recordings required for reliable data analysis and model development is generally infeasible, as it requires experts to label the data with a huge labor force and may also get delayed by prioritising system development over data gathering. As it is easy to collect unlabelled audio recordings at a large scale, exploring semi-supervised learning (SSL) for COVID-19 detection is of great interest, combining the unlabelled audio recordings in conjunction with a small amount of labelled data to improve the model performance.

A variety of SSL schemes have been investigated across a variety of different tasks, while the most commonly adopted algorithms include pseudo-labelling [9], mean-teacher [10], MixMatch (MM) [11] and its variant ReMixMatch (RMM) [12]. However, they are either highly dependent on the reliability of the supervised model, or suffer from high computational cost. FixMatch (FMM) was recently proposed for image recognition tasks, and showed superior performance while significantly simplifying existing SSL methods [13]. However, it has been mainly investigated in the image domain, and not been well explored in audio-related tasks.

In this paper, we proposed a SSL framework which explores the potential of semi-supervised learning for audio based COVID-19 detection tasks, with FixMatch adjusted to audio signals. We compared the proposed approach with supervised and other SSL approaches, and showed a 6.2% relative improvement in terms of ROC-AUC over the supervised model. Furthermore, we demonstrated how our approach can benefit COVID-19 detection in different sub-tasks, e.g. distinguishing symptomatic/asymptomatic positive users from symptomatic/asymptomatic negative users. The results showed great advantages of the proposed approach in dealing with imbalanced datasets, with a maximum relative improvement of 7.2% further validating its potential in developing more accurate and generalised detection tools.

2. Related work

Existing studies have shown the effectiveness of audio signals for COVID-19 detection [4, 5, 6, 7, 8, 14]. Coughs are first explored using deep learning techniques [7]. One of the studies investigated the dynamics of the glottal flow waveform during speech production (e.g. phonemes) to identify COVID-19, given the evidence that infection affects the respiratory system which in turn affects the speech [8]. Further, different sound types including cough, breathing and speech were combined to improve the detection performance [4, 5, 6]. Various machine learning techniques have also been validated, ranging from traditional Support Vector Machines [6] to more advanced deep learning approaches such as pre-trained VGGish [5] and ResNet [4] models. However, most of these models were developed using a small dataset ranging from 19 participants [8] to 355 participants [4], making it hard to generalise to participants unobserved by the model. Furthermore, asymptomatic positive patients may not volunteer to get tested, thus, the dataset used for model development may lack these samples. This makes it hard or even impossible for the model to recognize asymptomatic

[‡] Equal contributions, listed in alphabet order

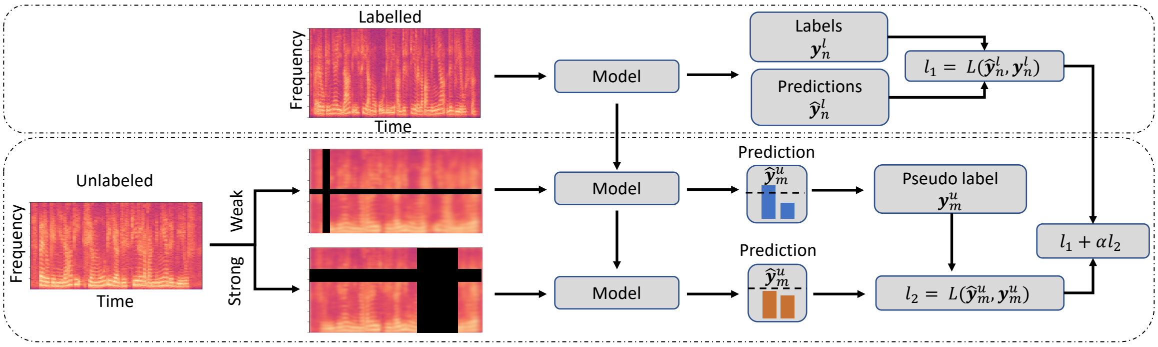

Figure 1: *An overview of the proposed SSL framework using FMM for COVID-19 detection. The model is trained using labelled samples, and employed to weakly augmented unlabelled spectrograms to select reliable unlabelled samples. Their predictions are converted to pseudo-labels for strongly augmented spectrograms. The final model is dynamically optimized using the loss combining both labelled (l_1) and selected unlabelled (l_2) samples.*

patients. However, their audio recordings can be available (no associated test label) within a large amount of unlabelled audio recordings. This provides a great portion of data with rich and valuable information to improve performance.

Semi-supervised learning (SSL), combining unlabelled data with labelled data, has attracted tremendous attention. Pseudo-labelling is one of the most widely adopted techniques [9], and serves as part of the pipeline for many advanced SSL algorithms. Recent SSL approaches which employ consistency regularization on unlabelled data show improved results [10, 11, 12, 13, 15, 16]. This minimises discrepancies between predictions for weakly and strongly deformed unlabelled samples, forcing the model to be versatile when faced with outliers and benefiting model development. FixMatch (FMM), one such algorithm that simply combines consistency regularisation and pseudo-labelling, demonstrated superior performance.

3. Methods

3.1. FixMatch for COVID-19 detection

An overview of the model pipeline using FMM for COVID-19 detection is shown in Figure 1. Labelled samples are first used to develop the supervised model, which is then adopted to gather the predictions for the weakly augmented unlabelled samples. Those with the predicted probability above a threshold for each class are selected as the confident samples. Their predictions are served as the artificial labels for the corresponding strongly augmented samples, which are combined with the labelled dataset to further optimise the model.

3.1.1. Model pipeline

The supervised model f is developed and optimized using cross-entropy loss L:

$$l_1 = \frac{1}{N}\sum_{n=1}^{n=N} L(\hat{\mathbf{y}}_n^l, \mathbf{y}_n^l) \tag{1}$$

where $\hat{\mathbf{y}}_n^l$ and $\mathbf{y}_n^l$ represent the prediction and test label for the n_{th} labelled sample. The model is also applied to weakly augmented unlabelled spectrograms to obtain pseudo-labels $\hat{\mathbf{y}}_m^u$ as:

$$\hat{\mathbf{y}}_m^u = f(\phi_w(\mathbf{x}_m^u)) \tag{2}$$

where $\phi_w(\cdot)$ represents the weak augmentation, and $\hat{\mathbf{y}}_m^u$ represents the predicted probabilities for the m_{th} unlabelled sample. To discard potentially incorrect pseudo-labels, only the unlabelled samples with the predicted probability $\hat{\mathbf{y}}_m^u$ of one class larger than a threshold τ are selected. The predictions of these selected samples are converted to a one-hot pseudo label $\mathbf{y}_m^u$ by binarilising $\hat{\mathbf{y}}_m^u$ using the larger probability (e.g. argmax). They are then used as the artificial labels (i.e. assumed positive or negative test results in our task) for the corresponding strongly augmented spectrograms.

The loss function for the selected unlabelled dataset is estimated as:

$$l_2 = \frac{1}{M_1}\sum_{m_1=1}^{m_1=N} L(f(\phi_s(\mathbf{x}_m^u)), \mathbf{y}_{m_1}^u) \tag{3}$$

where $\phi_s(\cdot)$ represents the strong augmentation and M_1 is the number of selected unlabelled audio samples after weak augmentation. The final loss computed over both the labelled and unlabeled dataset is:

$$l = l_1 + \alpha l_2 \tag{4}$$

where α controls the relative weight of the loss for the unlabelled data.

3.1.2. Data Augmentation

SpecAugment is used as the augmentation method, which shows success in automatic speech recognition [17] and acoustic scene classification [18]. Time masking and frequency masking are applied to the spectrogram. For each spectrogram of size $T \times F$, time masking is first applied to a range of consecutive time frames $[t_1, t_1 + \Delta t]$ by replacing these elements with 0, where t_1 and Δt are randomly selected from a uniform distribution to introduce randomness. Similarly, frequency masking is applied along the frequency dimension. To produce the weakly and strongly augmented spectrograms, we mask an equal or larger number of random bins in the strongly augmented spectrograms than in the weakly augmented ones.

3.1.3. Training strategies and model structure

Two different training strategies are proposed, referred to as static and dynamic training. Static training selects the unlabelled samples using the supervised model f once, and include these samples in the training data to further optimize the model. The reliability of the selected unlabelled samples are highly dependent on the accuracy of the supervised model, and the errors

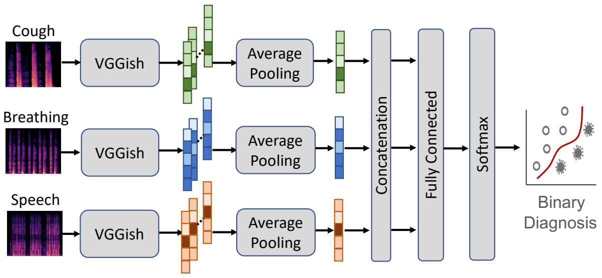

Figure 2: *Model structure. Three different modalities are used, and VGGish is used for feature extraction. These features from different modalities are concatenated and processed by fully connected layers for binary classification.*

introduced in the selected unlabelled samples may propagate and disrupt the model development.

The dynamic training scheme instead selects the unlabelled samples dynamically at each iteration and combines them with the labelled data. Specifically, it iterates through samples, to select the confident unlabelled samples and combine them with labelled samples. For each iteration, the selected samples will be different as the model's parameters are optimised. Through experiments, we have observed more unlabelled samples are included as the model's confidence improves over time.

The model structure is shown in Figure 2. Three different audio modalities are used: cough, breathing and speech. A pre-trained VGGish network [19], optimized for acoustic event detection, is used as the feature extractor. Two fully-connected layers are employed as the classifier for COVID-19 detection.

3.2. Tasks

According to participants' clinical symptoms, a series of binary classification tasks are explored, to provide more insights into how FMM aids in detecting different subgroups of patients.

- **Task 1**: Distinguish positive participants from negative (healthy) participants, which is the general case and referred as 'Pos-Neg'.
- **Task 2**: Distinguish symptomatic positive participants who reported at least one symptom from asymptomatic negative participants. This is expected to be a simple task as the audio sounds may show clear difference between the two subgroups. This task is referred as 'sPos-aNeg'.
- **Task 3**: Distinguish symptomatic positive participants from symptomatic negative participants, refereed as 'sPos-sNeg'.
- **Task 4**: Distinguish asymptomatic positive participants from asymptomatic negative participants, refereed as 'aPos-aNeg'.

4. Experimental setup

4.1. Data

We have collected a crowdsourced audio data set for COVID-19 detection via a mobile app (https://www.covid-19-sounds.org), which collects three types of audio recordings for each participant (cough, breathing and speech), along with their demographics, medical history, symptoms, and COVID-19 test results. More details can be found in [20].

We selected a subset of 1000 participants with 1486 labelled samples (734 positive and 752 negative samples). We divide the data into training, validation and test partitions with 70%, 10% and 20% respectively, with relatively balanced gender, age and symptoms to avoid any bias. We have selected 2381 participants with 2778 unlabelled samples in total.

Table 1: *System performance using supervised learning (SL) and SSL of pseudo labelling (PL) and Fixmatch (FMM) for COVID-19 detection. Both static ^{s} and dynamic ^{d} learning schemes are reported. FMM^d outperforms other systems.*

System	ROC-AUC	Sensitivity	Specificity
SL	0.65(0.59-0.71)	0.62(0.54-0.69)	0.56(0.49-0.64)
PL^s	0.65(0.58-0.70)	0.67(0.59-0.74)	0.51(0.43-0.58)
PL^d	0.67(0.61-0.73)	0.80(0.73-0.86)	0.41(0.34-0.48)
FMM^s	0.67(0.61-0.73)	0.68(0.61-0.75)	0.54(0.46-0.62)
FMM^d	**0.69(0.63-0.74)**	**0.65(0.58-0.73)**	**0.63(0.56-0.71)**

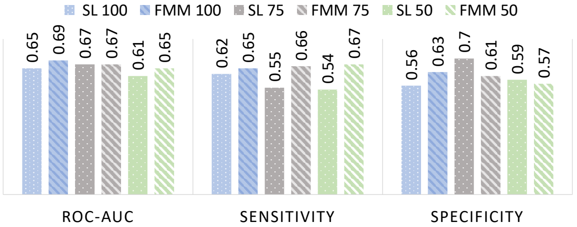

Figure 3: *System performance with different percentages of labelled data within [100% 75% 50%] for SL model and FMM^d. FMM^d outperforms or shows comparable performance as SL (ROC-AUC), and yields higher relative improvements or more balanced sensitivity and specificity with less labelled data.*

4.2. Settings

All the audio recordings were automatically checked using YAMNet [19] to remove the unqualified ones (e.g. noisy background, etc.). Each recording was then resampled to 16 kHz, converted to mono channel and normalised to make the maximum amplitude 1. Silent regions at the beginning and the end of the recordings were removed. We adopted the same model structure (Figure 2) as in [5] due to its comprehensive validations and strong performance using supervised learning for COVID-19 detection task, but the data and the resources used to train the model differ. Our model is trained using one GPU in the high-performance computing clusters.

Two fully connected layers with 64 neurons were used as the classifier. VGGish is fine-tuned jointly with the classifier, with the initial learning rate set as 5e-5 for VGGish and 1e-4 for the classifier.It is decayed by 2% for every 1000 training samples. The model is trained for 20 epochs with the Adam optimiser. The final 5 epochs include the validation data to further boost the model performance. We empirically chose 5% time and frequency masking for the weakly augmented spectrograms, and optimised within the range of $[5, 35]$ percent with a step size of 10 for the strong augmentation strength. The unlabelled loss weight α was set to 0.33 as an approximation to the ratio of labelled and unlabelled data. The threshold τ is empirically chosen as 0.95. The code was implemented using Tensorflow [21].

The model performance was evaluated using ROC-AUC, Sensitivity and Specificity. ROC-AUC shows the overall capability of the model in correctly classifying the positive and negative participants. Sensitivity illustrates the model capability in correctly identifying positive patients, while specificity shows that in correctly identifying healthy participants. A 95% Confidence Interval (CI) for the model performance is estimated using bootstrap [22].

Table 2: *System performance for SL and FMM^d for Tasks 2-4. Number of samples for each task is included in parenthesis (training/test). FMM^d shows great advantages in balancing sensitivity and specificity.*

Task	System	Accuracy	ROC-AUC	Sensitivity	Specificity
T2: sPos (*433/150*)-aNeg (*282/88*)	SL	0.69(0.63-0.75)	0.8(0.74-0.86)	0.55(0.45-0.65)	0.83(0.77-0.89)
	FMM^d	0.74(0.68-0.79)	0.78(0.72-0.84)	0.69(0.61-0.76)	0.78(0.69-0.86)
T3: sPos (*433/150*)- sNeg (*336/113*)	SL	0.57(0.51-0.63)	0.62(0.55-0.69)	0.73(0.66-0.8)	0.41(0.32-0.5)
	FMM^d	0.58(0.52-0.64)	0.63(0.56-0.7)	0.57(0.49-0.66)	0.59(0.5-0.68)
T4: aPos (*82/30*)-aNeg (*336/113*)	SL	0.55(0.47-0.65)	0.66(0.55-0.76)	0.27(0.12-0.43)	0.84(0.76-0.92)
	FMM^d	0.54(0.45-0.63)	0.6(0.47-0.71)	0.43(0.26-0.61)	0.72(0.62-0.81)

Table 3: *System performance for SL and FMM^d for Tasks 2-4. Number of samples for each task is included in parenthesis (training/test). FMM^d shows great advantages in balancing sensitivity and specificity.*

Task	System	Accuracy	ROC-AUC	Sensitivity	Specificity
T2	SL	0.69(0.63-0.75)	0.80(0.74-0.86)	0.55(0.45-0.65)	0.83(0.77-0.89)
	FMM^d	0.74(0.68-0.79)	0.78(0.72-0.84)	0.69(0.61-0.76)	0.78(0.69-0.86)
T3	SL	0.57(0.51-0.63)	0.62(0.55-0.69)	0.73(0.66-0.80)	0.41(0.32-0.50)
	FMM^d	0.58(0.52-0.64)	0.63(0.56-0.70)	0.57(0.49-0.66)	0.59(0.50-0.68)
T4	SL	0.55(0.47-0.65)	0.66(0.55-0.76)	0.27(0.12-0.43)	0.84(0.76-0.92)
	FMM^d	0.54(0.45-0.63)	0.60(0.47-0.71)	0.43(0.26-0.61)	0.72(0.62-0.81)

5. Results and discussion

5.1. Comparison with supervised model

The comparison of the proposed system to the supervised learning (SL) model and pseudo-labelling (PL) approach for Task 1 (Pos-Neg) is shown in Table 1. It can be observed that pseudo labelling with either static or dynamic training could not benefit COVID-19 detection, possibly due to the weakness in the supervised model which generates incorrect artificial labels. FMM^s and FMM^d show a 3.1% and 6.2% relative improvement of ROC-AUC over the supervised model respectively. In addition, FMM^d leads to more balanced sensitivity and specificity, and a significant higher specificity with a 12.5% relative improvement, suggesting that FMM is able to select more reliable unlabelled samples, benefiting the task.

5.2. Evaluation for different subtasks

The system performance for Tasks 2-4 (T2-T4) with SL and FMM^d are shown in Table 3. ROC-AUC can be misleading as the number of samples in the positive and negative class for each subtask differ. We additionally reported the balanced accuracy, which is computed as the average of recall obtained on each class and mitigates this problem. SL easily skews to one of the classes for all three subtasks, either with a high sensitivity and low specificity, or vice versa. This is likely due to the imbalanced dataset. FMM^d demonstrates superior performance for T2-T3 in terms of accuracy, showing great advantages in balancing sensitivity and specificity. This suggests that the proposed approach is able to select reliable but unfamiliar unlabelled samples to aid model development. For T4, though FMM^d shows comparable performance in terms of accuracy, it still yields a decreased discrepancy between sensitivity and specificity. The overall unsatisfying performance for T4 could be attributed to the extreme scarce samples in the positive class.

5.3. Size of labelled training data

We further evaluated the model performance using FMM^d with different percentages of labelled data for Task 1 (a general task). As shown in Figure 3, FMM^d shows superior or comparable performance with the SL model for different percentages of labelled data ranging within [100, 75, 50]. A higher relative improvement of 6.6% using 50% labelled data over a 6.2% improvement using 100% labelled data is observed. Further, FMM^d achieves more balanced sensitivity and specificity over SL using 75% labelled data. This evidence suggests that FMM^d shows greater advantages when less labelled data is available.

5.4. Visualisation in latent space

The latent vectors from the last hidden layer were projected into a 2-dimensional space using t-SNE [23] for both the SL and FMM^d for Task 1, as shown in Figure **??**. SL yields a model which can cluster and separate positive and negative samples accurately for the training samples, but not for test samples, suggesting overfitting possibly due to the limited training dataset size. On the contrary, FMM^d maps a few positive and negative samples in the wrong cluster, but still well separates positive and negative clusters. These few samples might not be wrongly clustered, as the test labels are self-reported which might be noisy and introduce a potential time lag (i.e. recovered but not tested and continuously reported positive test result). The positive and negative samples in the test set are clustered better using FMM^d, indicating a better generalisation capability.

6. Conclusion

A semi-supervised learning framework (SSL) using adjusted FixMatch is proposed for audio-based COVID-19 detection. It is validated in a crowdsourced dataset and the superior performance over supervised learning and commonly adopted pseudo-labelling demonstrates its effectiveness. The improved performance on different tasks further showed that the proposed approach significantly benefits learning in imbalanced datasets, which is common in clinical data. We showed the potential of the proposed framework in developing a more accurate and generalised model incorporating the great source of unlabelled data. Future work includes investigating more advanced augmentation methods for audio signals, and improved fusion strategies of different modalities under SSL scheme.

7. Acknowledgements

This work was supported by ERC Project 833296 (EAR). We thank everyone who volunteered their data.

8. References

[1] M. Cevik, K. Kuppalli, J. Kindrachuk, and M. Peiris, "Virology, transmission, and pathogenesis of SARS-CoV-2," *British Medical Journal*, vol. **371**, pp. 1–6, 2020.

[2] C. B. Vogels, A. F. Brito, A. L. Wyllie, J. R. Fauver, I. M. Ott, C. C. Kalinich, M. E. Petrone, A. Casanovas-Massana, M. C. Muenker, A. J. Moore *et al.*, "Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT–qPCR primer–probe sets," *Nature Microbiology*, vol. **5**, no. 10, pp. 1299–1305, 2020.

[3] I. Torjesen, "Covid-19: How the uk is using lateral flow tests in the pandemic," *bmj*, vol. 372, 2021.

[4] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables covid-19 detection from breath and cough audio: a pilot study," *BMJ innovations*, vol. 7, no. 2, 2021.

[5] J. Han, T. Xia, D. Spathis, E. Bondareva, C. Brown, J. Chauhan, T. Dang, A. Grammenos, A. Hasthanasombat, A. Floto *et al.*, "Sounds of covid-19: exploring realistic performance of audio-based digital testing," *arXiv preprint arXiv:2106.15523*, 2021.

[6] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," *arXiv preprint arXiv:2006.05919*, 2020.

[7] 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.

[8] S. Deshmukh, M. Al Ismail, and R. Singh, "Interpreting glottal flow dynamics for detecting covid-19 from voice," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1055–1059.

[9] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[11] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[13] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[14] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "Covid-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, vol. 135, p. 104572, 2021.

[15] S. Calderon-Ramirez, R. Giri, S. Yang, A. Moemeni, M. Umana, D. Elizondo, J. Torrents-Barrena, and M. A. Molina-Cabello, "Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5294–5301.

[16] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *arXiv preprint arXiv:1903.03825*, 2019.

[17] 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.

[18] H. Wang, Y. Zou, and W. Wang, "Specaugment++: A hidden space data augmentation method for acoustic scene classification," *arXiv preprint arXiv:2103.16858*, 2021.

[19] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[20] T. Xia, D. Spathis, J. Ch, A. Grammenos, J. Han, A. Hasthanasombat, E. Bondareva, T. Dang, A. Floto, P. Cicuta *et al.*, "Covid-19 sounds: A large-scale audio dataset for digital respiratory screening," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[21] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[22] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical science*, vol. 11, no. 3, pp. 189–228, 1996.

[23] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.