



# Uncertainty-Aware COVID-19 Detection from Imbalanced Sound Data

Tong Xia, Jing Han, Lorena Qendro, Ting Dang, Cecilia Mascolo

Department of Computer Science and Technology, University of Cambridge, UK

tx229@cam.ac.uk

## Abstract

Recently, sound-based COVID-19 detection studies have shown great promise to achieve scalable and prompt digital pre-screening. However, there are still two unsolved issues hindering the practice. First, collected datasets for model training are often imbalanced, with a considerably smaller proportion of users tested positive, making it harder to learn representative and robust features. Second, deep learning models are generally overconfident in their predictions. Clinically, false predictions aggravate healthcare costs. Estimation of the uncertainty of screening would aid this. To handle these issues, we propose an ensemble framework where multiple deep learning models for sound-based COVID-19 detection are developed from different but balanced subsets from original data. As such, data are utilized more effectively compared to traditional up-sampling and down-sampling approaches: an AUC of 0.74 with a sensitivity of 0.68 and a specificity of 0.69 is achieved. Simultaneously, we estimate uncertainty from the disagreement across multiple models. It is shown that false predictions often yield higher uncertainty, enabling us to suggest the users with certainty higher than a threshold to repeat the audio test on their phones or to take clinical tests if digital diagnosis still fails. This study paves the way for a more robust sound-based COVID-19 automated screening system.

**Index Terms:** COVID-19, sound-based digital diagnosis, ensemble learning, uncertainty estimation

## 1. Introduction

Since the outbreak of the Coronavirus Disease 2019 (COVID-19) in March 2020, over 100 million confirmed cases and 2 million deaths have been identified globally. Frequent and massive screening with targeted interventions is of vital need to mitigate the epidemic [1, 2]. Currently, the most common screening tool for COVID-19 is the Reverse Transcription Polymerase Chain Reaction (RT-PCR), which is limited by cost, time, and resources [3, 4]. To fight against the virus, researchers' effort has gone into exploring machine learning for fast, contactless, and affordable COVID-19 detection from sounds on smartphones [5]: COVID-19 is an infectious disease, and most infected people experience mild to moderate respiratory illness [6]. To validate the effectiveness of these approaches, sound data are normally collected with self-reported COVID-19 testing results or more trustworthy COVID-19 clinical testing codes.

Although recently great progress have been witnessed on cough [7, 8, 9, 10, 11] and speech-based [12, 13, 14] COVID-19 detection, there are still unsolved issues which hinder the rollout of this technology to the masses. First, the collected sound data are generally imbalanced, with a small proportion of COVID-19 infected or tested positive participants [11, 12, 8]. Such imbalance in training makes the classifier biased to the majority class for a relatively small loss, but it does not mean distinguishable COVID-19 features can be learned from human

sounds [15, 16]. This issue is even more detrimental in deep learning as the data are limited and thus balancing solutions are insufficient [17]. In addition, even if machine learning can achieve high precision, difficult diagnosis cases (e.g., out of training data distribution, noise, sound distortion) are unavoidable [18, 19]. In this respect, information about the reliability of the automated COVID-19 screening is a key requirement to be integrated into diagnostic systems for better risk management. Yet, none of the existing works on sound-based COVID-19 detection takes into consideration the uncertainty in the machine learning prediction.

In this paper, we propose an ensemble learning-based framework to tackle the training data imbalance and uncertainty estimation challenges, simultaneously. Briefly, when training deep learning models for COVID-19 detection, a number of balanced training sets are generated from the imbalanced data to learn multiple ensemble units. During inference, decisions are fused to maximize data utilization and improve generalization ability. More importantly, we make use of the disagreement across the learned deep learning models as a measure of uncertainty. Softmax probability may indicate the confidence of the prediction, but it tends to overestimate confidence and requires further calibration [20]. Instead, disagreement from deep ensembles as the uncertainty estimation was proven to better represent the overall model confidence [21]. With uncertainty, predictions with low confidence can be identified and these samples could be excluded from digital screening. The users who produced these samples could be asked to repeat smartphone testing or referred on for different types of testing. As a consequence, this method improves the system performance and patient safety at the same time [22].

To help with this research, we launched a mobile app in April 2020 to crowdsource sound data including breathing, cough, and speech with self-reported COVID-19 testing results<sup>1</sup>. In conclusion, our contributions in this paper are summarised as follows,

- To handle the limited and imbalanced training data problem, we propose a deep ensemble learning-based framework for COVID-19 sounds analysis, yielding higher AUC and sensitivity compared to other balancing approaches.
- To the best of our knowledge, we are the first to investigate the uncertainty of deep learning for sound-based COVID-19 detection, leading to a more reliable and robust automated diagnosis system.
- We perform experiments on our collected data with 469 tested positive and 1 526 healthy control samples, achieving an AUC of 0.74 against 0.62 from an SVM baseline. With the estimated uncertainty, the AUC is further improved up to 0.85.

<sup>1</sup>[www.covid-19-sounds.org](http://www.covid-19-sounds.org)

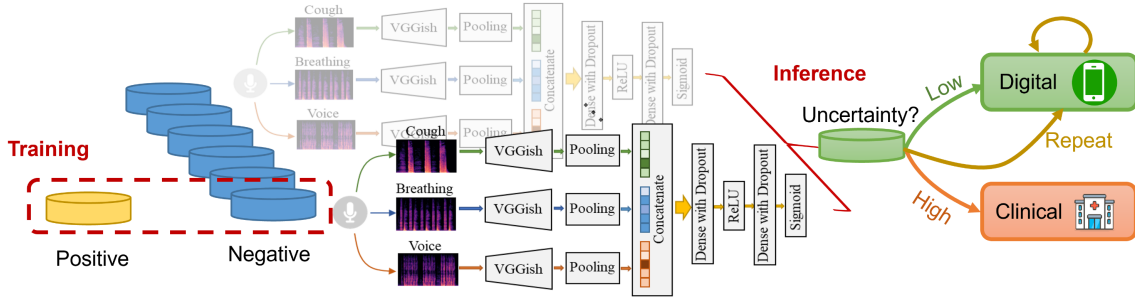


Figure 1: *Uncertainty-aware deep ensemble learning-based COVID-19 detection framework. Balanced training sets are generated to train multiple CNN-based models, and probabilities for a testing sample are fused to form the final decision. Simultaneously, the disagreement level across these models as a measure of uncertainty is obtained and used to indicate the reliability of digital diagnosis.*

## 2. Related Works

In [7, 8, 10, 11, 23], respiratory sounds, especially coughs, have been analysed and detectable features have been discovered from spectrograms to distinguish COVID-19 coughs from healthy or other disease coughs. At the same time, researchers have exploited speech signals [12, 13, 14, 24], which are more natural and informative. All those studies demonstrate the promise of detecting COVID-19 from sounds, in a non-invasive, inexpensive, and largely available manner. However, all the above works have suffered from imbalanced data problem. For instance, 92 COVID-19 tested positive and 1079 healthy subjects were included in [8, 25]. The imbalance was either dealt with down-sampling, up-sampling or generally not yet tackled in the research [10, 26, 23, 11, 27]. However, down-sampling, by discarding samples from the majority class, results in an even smaller size of training data and might lose important discriminative information, while up-sampling, by repeating samples from minority class, might change the original distribution of the data and lead to model overfitting. Moreover, although Synthetic Minority Oversampling Technique (SMOTE) is a more advanced up-sampling method [28], it is inherently dangerous as it violates the independence assumption and blindly generalizes the minority area without regard to the majority class.

There has been a consensus in the literature that uncertainty estimation could be used to aid automated clinical diagnosis, for example for clinical imaging analysis [18]. In terms of COVID-19 diagnosis, one work has obtained uncertainty from CT (computerized tomography) scans to achieve interpretable COVID-19 detection [29]. Though softmax probability may indicate the confidence of the prediction, to some extent, it only captures the relative probability that an instance is from a particular class, compared to the other classes, instead of the overall model confidence. Furthermore, it tends to overestimate confidence and thus requires further calibration [20]. In general, Bayesian Convolutional Neural Network [30] and Monte Carlo Dropout (MCDrop) [31] & Monte Carlo Dropweights [29] have been exploited to estimate uncertainty. However, Bayesian estimation is computationally expensive, which is not an optimal solution for our task with limited training data, while Dropweights in [29], keeps dropout on during inference, reducing the model capacity and may leading to lower accuracy [21]. With evidence suggesting that uncertainty from deep ensembles outperforms MCDrop [21, 32, 33], this paper proposed the ensemble learning framework which tackles the data imbalance and uncertainty estimation simultaneously within a unified framework.

## 3. Methodology

In this paper, we focus on developing an uncertainty-aware audio-based covid-19 prediction model. In particular, the basic unit integrates information from three different sound types, i. e., breathing, cough, and speech. Then, ensemble learning is exploited to handle the highly unbalanced data. Furthermore, an uncertainty estimation can be obtained from the ensemble framework. The proposed framework is illustrated in Figure 1.

### 3.1. COVID-19 Detection Model

Three different modalities are adopted to develop the deep learning-based COVID-19 detection model - the basic unit of ensemble learning. Following previous work [10], the CNN model named Vggish [34] is applied and adapted as the feature extractor, which is trained on a large-scale audio dataset for audio event classification. In particular, Mel-spectrums are extracted from each modality and fed into Vggish, which yield a 128-dimensional embedding for each 0.96-second audio segment. It is followed with average pooling along the time-axis to obtain a fixed-sized feature vector from length-varying inputs. The feature vectors of three modalities are then concatenated as the combined features. Finally, another two dense layers with a softmax function are utilized as a binary classifier, the outputs of which can be interpreted as the probability of COVID-19 infection.

### 3.2. Ensemble Training and Inference

Many machine learning approaches struggle to deal with real-world health data because it is common to have imbalanced datasets where the healthy users are a significant majority of the whole set. This is also the case for COVID-19 sound-based detection, where the users who tested positive are the minority class. To tackle this problem, as described in Figure 1, we generate a series of training bags  $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)\}$  with an equal number of positive and healthy users by re-sampling from the majority healthy class, where  $X_i$  denotes the input audio samples within the  $i$ -th bag and  $Y_i$  are the corresponding labels. Note that since some healthy users can be re-used, we can generate numerous bags. Consequently, based on a fixed and shared validation set, we train  $N$  neural networks, as introduced in Sec. 3.1, via cross-entropy loss, independently. During inference, we pass the testing sample  $x$  into the ensemble suite and obtain probability-based fusion [35] as the final output, formulated as follows,

$$p(y) = \frac{1}{N} \sum_{i=1}^N P(y = 1|x, X_i, Y_i, \theta_i), \quad (1)$$

where  $P(y|x, X_i, Y_i, \theta_i)$  is the predicted softmax probability of  $i$ -th model. If  $p(y)$  is higher than 0.5, the given testing sample will be predicted as COVID-19 positive.

### 3.3. Uncertainty Estimation

No matter how good the deep learning model is, difficult cases to diagnose are unavoidable: this could be due to many reasons, including very noisy samples. This highlights the importance of uncertainty estimation for the digital screening. Considering the incapability of softmax probability to capture model confidence, we define the disagreement level across models within the ensemble suite as the measure of uncertainty. Uncertainty from deep ensembles has also been shown to be more effective than other estimation approaches [21]. To be specific, we use the standard deviation across the  $N$  models as the measurement of uncertainty as follows,

$$\sigma(y) = \sqrt{\frac{1}{N} \sum_{i=1}^N (P(y = 1|x, X_i, Y_i, \theta_i) - \mu)^2}, \quad (2)$$

where  $\mu$  is the averaged probability, as  $p(y)$  in Eq. (1).

If the uncertainty  $\sigma(y)$  is higher than a predefined threshold, it implies that the model is confident enough with its prediction during digital pre-screening. Under this circumstance, the system can first request a second or even more repeated audio testing on smartphones. If the uncertainty is still high, this particular sample could be then referred for further clinical or another testing. As a consequence, both system performance and patient safety can be improved.

## 4. Evaluation

### 4.1. Dataset

Given the great potential of audio-based COVID-19 digital screening, we launched an app, namely *COVID-19 Sounds App*, to crowdsource data for research. In the initial registration, users' basic demographic and medical history information are required. Then, users are guided to record and submit breathing, coughing, and short speech audios, together with the latest coronavirus testing results and associated symptoms, if any, every other day. To be more specific, audios are three to five inhalation-exhalation sounds, three voluntary cough sounds, and the participant reading a standard sentence from the screen three times.

In this study, we focus on the group consisting of users who declared to have tested positive and ones who declared they have tested negative and declared no symptoms: we call these users "healthy". To avoid any language confounders in the voice recordings, only English speakers were retained. After a manual quality check, 330 positive users with 469 samples and 919 healthy users with 2021 samples were selected. Overall, 58% of the users are male and more than 60% are aged between 20 to 49. Demographic and medical history distributions are similar in the two classes. As for pre-processing, we resample all the recordings to 16 kHz mono audios, and then remove the silence period at the beginning and the end of the recording as in [26]. Finally, audio normalisation by calibrating the peak amplitude to 1 is applied to eliminate the discrepancy across recording devices.

Table 1: Basic statistics of COVID-19 sound data.

	Positive		Healthy	
	#Users	#Samples	#Users	#Samples
Train	231	327	820	1871
Validation	33	44	33	56
Test	66	98	66	94

### 4.2. Experimental setup

To evaluate the proposed framework, for the positive group, we hold out 10% and 20% of users as validation and testing sets, and then use the remaining data for training. Correspondingly, we select the same of healthy users for validation and testing (see Table 1). Furthermore, to generate a balanced training set for each ensemble unit  $(X_i, Y_i)$ , 231 users are randomly selected from 820 negative tested users. The number of hidden units of dense layers in our model is set to 64 and 2, respectively. When training, our batch size is 1, the learning rate is 0.0001 with a decay factor of 0.99 and we use cross-entropy loss and Adam optimiser. Early stopping is applied on the validation set to avoid over-fitting. All experiments are implemented in Python and TensorFlow. Feature extractor is initialised by pre-trained VGGish model and then updated with the following dense layers jointly.

Moreover, baselines from the latest literature are conducted for performance comparison. In addition to deep models, acoustic feature-driven SVM achieved state-of-the-art performance in robust-based COVID-19 detection due to its effectiveness and robustness in small data learning [10, 12, 26]. Therefore, we repeat the experiments in [26] by using the openSMILE toolkit to extract 384 acoustic features and SVM with linear kernel and complex constant  $C = 0.01$  as the classifier. For both SVM and deep models, to test the superior performance of ensemble learning, we compare training one single model with all samples, with balanced samples by down or up-sampling, against training  $N = 10$  models in an ensemble learning manner. For down-sampling balancing, we randomly discard some healthy users, while for up-sampling, synthetic minority over-sampling techniques (SMOTE) [28] is performed to generate synthetic observations of the minority class. This is the most commonly adopted technique for imbalanced data.

To justify the performance of the proposed framework for COVID-19 screening and diagnosis, we calculate the following metrics: **Sensitivity**, also named true positive rate or recall defined by  $TP/(TP + FN)$ , **Specificity**, also referred to as true negative rate formulated by  $TN/(TN + FP)$ , and **ROC-AUC**, the area under receiver operating characteristics curve, which measures the overall sensitivity and specificity at various probability thresholds. Besides, for both the baseline and our proposed methods, the mean and standard deviation of the aforementioned metrics across 10 runs are reported.

### 4.3. Results

#### 4.3.1. Performance of Ensemble Learning

The overall comparison is presented in Table 2. First, deep learning is more vulnerable than the SVM model with imbalanced training data. When the CNN model achieves an ROC-AUC of 0.69, against 0.60 of SVM, the sensitivity is very low because the model is biased to the healthy class, and the very high specificity of 0.98 is practically meaningless. Second, re-sampling can improve the performance, especially in terms of sensitivity for both SVM and deep learning, since a balanced

Table 2: Performance comparison with Mean(Std) for ROC-AUC, Sensitivity, and Specificity reported for Single model (SM) and Ensemble model.

		ROC-AUC	Sensitivity	Specificity
SM imbalanced data	SVM	0.60	0.54	0.57
	CNN	0.69	0.15	0.98
SM down-sampling	SVM	0.60(0.03)	0.68(0.05)	0.45(0.05)
	CNN	0.68(0.04)	0.62(0.04)	0.63(0.06)
SM up-sampling	SVM	0.62(0.02)	0.53(0.02)	0.63(0.02)
	CNN	0.70(0.04)	0.52(0.02)	0.77(0.05)
<b>Ensemble model</b>	SVM	0.66(0.04)	0.63(0.05)	0.62(0.04)
	<b>CNN</b>	<b>0.74(0.03)</b>	<b>0.68(0.05)</b>	0.69(0.06)

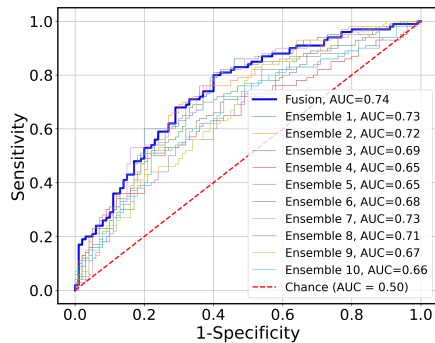


Figure 2: ROC curves of COVID-19 detection, where the curve of each ensemble unit model is shown separately. The ROC curve after probability-based fusion is shown in bold.

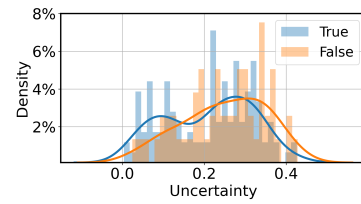
training set is guaranteed. Compared to down-sampling, up-sampling performs better, as expected, because more data samples are available for parameter learning. Third, ensembles can boost the performance of both SVM and CNN. Last but not least, our CNN ensemble framework outperforms all the baselines and achieves an AUC of 0.74 with a sensitivity and a specificity close to 0.7, demonstrating the superiority of exploiting ensemble learning and deep CNN network for COVID-19 detection from imbalanced data. In addition, the commonly used majority voting fusion method [35] was also validated in comparison to our proposed probability-based fusion: an AUC of 0.74 with a sensitivity of 0.62 and a specificity of 0.70 was obtained. Given the same AUC but lower sensitivity, we confirm that probability-based fusion is more promising.

To further inspect the ensemble suite, we visualise the ROC curves for each unit model in Figure 2 for comparison. All ROC curves are above chance level but the model variance is not negligible. A plausible explanation is that we only have a small training dataset for each unit. However, after probability-based fusion, the ROC curve is generally higher than the other curves. All these demonstrate that ensembles can improve the robustness and generation ability of machine learning models.

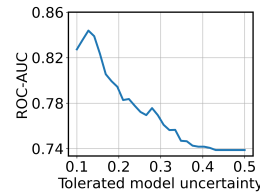
To conclude, by integrating multiple networks trained from different balanced data, our proposed approach achieves state-of-the-art with the AUC of 0.74, the sensitivity of 0.68 and the specificity of 0.69 for COVID-19 detection.

#### 4.3.2. Estimation and Application of Uncertainty

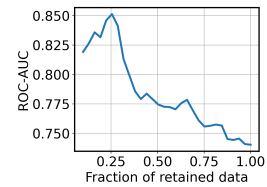
By using the standard deviation across ensemble models as the measure of uncertainty, we are able to understand the confidence of the predictions. We illustrate the uncertainty distribution in Figure 3(a). The density of high uncertainty values for false prediction is significantly higher than that of true prediction, indicating that our approach succeeds in identifying less confident predictions when making an incorrect diagnosis.



(a) Uncertainty distribution.



(b) Uncertainty threshold.



(c) Remained fraction.

Figure 3: Performance for uncertainty-aware referral.

Inspired by the findings, we set up thresholds to exclude some testing cases when the uncertainty is higher than a given value. Results from our dataset are shown in Figure 3(b), from which we can find that when keeping only testing samples with an uncertainty lower than 0.2, the AUC can be improved from 0.74 to 0.79, and when the threshold is 0.12, the highest AUC value is reached at 0.85. Note that the drop on the left side of the curve is caused by a small number of samples to calculate the metric. These results show that studying predictive uncertainty and finding the optimal threshold can significantly improve automatic digital prediction.

Referral for further testing can impose a range of costs which can vary from a cost-negligible “repeat the audio test on your phone” to a clinician inspection or visit. In this respect, we inspect the AUC on different fractions of retained data with the uncertainty lower than a certain threshold (assuming the samples above this threshold would be referred on). Figure 3(c) shows that the AUC climbs from 0.74 to 0.76 when 15% of the samples with the highest uncertainty are excluded. When only remaining 25% data with the lowest uncertainty, the AUC can increase to 0.85. This indicates that our measurement for uncertainty is informative and helpful for a more robust automatic COVID-19 diagnosis system.

## 5. Conclusions

In this work, a sound-based machine learning approach has been proposed to discriminate COVID-19 cases from health controls. Ensemble learning has been explored to solve the imbalanced training data challenge, yielding favorable performance gain on an in-the-wild crowdsourced dataset. In addition, uncertainty has been measured via the disagreement across ensembles, and thus enables confidence-informed digital diagnose. To the best of our knowledge, we are the first to introduce uncertain-aware deep learning approach to sound-based COVID-19 detection studies.

For future work, we plan to gain deeper understanding on the estimated uncertainty, exploring how each modality (breathing, cough, speech) contributes to the overall uncertainty. Our data collection is still ongoing which will yield more data to train our framework for further performance evaluation.

## 6. Acknowledgements

This work was supported by ERC Project 833296 (EAR).

## 7. References

- [1] E. Hunter, D. A. Price, E. Murphy, I. S. van der Loeff, K. F. Baker, D. Lendrem, C. Lendrem, M. L. Schmid, L. Pareja-Cebrian, A. Welch *et al.*, “First experience of COVID-19 screening of health-care workers in England,” *The Lancet*, vol. 395, no. 10234, pp. e77–e78, 2020.
- [2] A. Atkeson, M. C. Droste, M. Mina, and J. H. Stock, “Economic benefits of COVID-19 screening tests,” *National Bureau of Economic Research*, 2020.
- [3] 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.
- [4] 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.
- [5] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, “COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the sars-cov-2 corona crisis,” *Frontiers in Digital Health*, vol. 3, pp. 1–10, 2021.
- [6] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui *et al.*, “Clinical characteristics of coronavirus disease 2019 in China,” *New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.
- [7] W. Wei, J. Wang, J. Ma, N. Cheng, and J. Xiao, “A real-time robot-based auxiliary system for risk evaluation of COVID-19 infection,” in *Proceedings of INTERSPEECH*, 2020, pp. 701–705.
- [8] 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.
- [9] 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:2009.08790*, 2020.
- [10] 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,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2020, pp. 3474–3484.
- [11] J. Laguarda, 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.
- [12] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. Ji, T. Koike, and B. W. Schuller, “An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety,” in *Proceedings of INTERSPEECH*, 2020, pp. 4946–4950.
- [13] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, and V. Aharonson, “Sars-cov-2 detection from voice,” *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 268–274, 2020.
- [14] M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh, “Voice quality evaluation in patients with COVID-19: An acoustic analysis,” *Journal of Voice*, pp. 1–7, 2020.
- [15] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, “On the class imbalance problem,” in *Proceedings of the International Conference on Natural Computation (ICNC)*, vol. 4, 2008, pp. 192–201.
- [16] M. Schubach, M. Re, P. N. Robinson, and G. Valentini, “Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants,” *Scientific Reports*, vol. 7, no. 1, pp. 1–12, 2017.
- [17] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.
- [18] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, “Leveraging uncertainty information from deep neural networks for disease detection,” *Scientific Reports*, vol. 7, no. 1, pp. 1–14, 2017.
- [19] L. Qendro, J. Chauhan, A. G. C. Ramos, and C. Mascolo, “The benefit of the doubt: Uncertainty aware sensing for edge computing platforms,” *arXiv:2102.05956*, 2021.
- [20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330.
- [21] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6405–6416.
- [22] M.-H. Laves, S. Ihler, T. Ortmaier, and L. A. Kahrs, “Quantifying the uncertainty of deep learning-based computer-aided diagnosis for patient safety,” *Current Directions in Biomedical Engineering*, vol. 5, no. 1, pp. 223–226, 2019.
- [23] M. Pahar, M. Klopper, R. Warren, and T. Niesler, “COVID-19 cough classification using machine learning and global smartphone recordings,” *arXiv:2012.01926*, 2020.
- [24] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen *et al.*, “The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates,” *arXiv:2102.13468*, 2021.
- [25] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, P. K. Ghosh, S. Ganapathy *et al.*, “Coswara—a database of breathing, cough, and voice sounds for COVID-19 diagnosis,” in *Proceedings of INTERSPEECH*, 2020, pp. 4811–4815.
- [26] 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 *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 1–5.
- [27] B. W. Schuller, H. Coppock, and A. Gaskell, “Detecting COVID-19 from breathing and coughing sounds using deep neural networks,” *arXiv:2012.14553*, 2020.
- [28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [29] B. Ghoshal and A. Tucker, “Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection,” *arXiv:2003.10769*, 2020.
- [30] M. Teye, H. Azizpour, and K. Smith, “Bayesian uncertainty estimation for batch normalized deep networks,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 4907–4916.
- [31] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 1050–1059.
- [32] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” *arXiv:1906.02530*, 2019.
- [33] X. Wu, K. M. Knill, M. J. Gales, and A. Malinin, “Ensemble approaches for uncertainty in spoken language assessment,” in *Proceedings of INTERSPEECH*, 2020, pp. 3860–3864.
- [34] 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 *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 131–135.
- [35] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, “The 2018 computational paralinguistics challenge: Atypical self-assessed affect, crying & heart beats,” in *Proceedings of the INTERSPEECH*, 2018, pp. 122–126.