UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Positive-pair redundancy reduction regularisation for speech-based asthma diagnosis prediction

Georgios Rizos, Rafael A. Calvo, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Rizos, Georgios, Rafael A. Calvo, and Björn W. Schuller. 2023. "Positive-pair redundancy reduction regularisation for speech-based asthma diagnosis prediction." In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4-10 June 2023, Rhodes Island, Greece*, edited by Petros Maragos, Kostas Berberidis, Petros Boufounos, Aggelos Pikrakis, and Thomas Feuillen, 1–5. Piscataway, NJ: IEEE.
https://doi.org/10.1109/icassp49357.2023.10097087.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
Deutsches Urheberrecht
Weitere Informationen finden Sie unter: / For more information see:
https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/

POSITIVE-PAIR REDUNDANCY REDUCTION REGULARISATION FOR SPEECH-BASED ASTHMA DIAGNOSIS PREDICTION

Georgios Rizos[1], *Rafael A. Calvo*[2], *Björn W. Schuller*[1,3]

[1]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[2]Dyson School of Design Engineering, Imperial College London, UK
[3]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
georgios.rizos12@imperial.ac.uk

ABSTRACT

Asthma affects an estimated 334 million people worldwide, causing over 461 000 deaths. Exacerbations or asthma attacks can be predicted with new sensor technologies. We explore how recordings of human voice, and machine learning can provide better diagnostics for pulmonary diseases like asthma, as well as tools for helping patients better manage it. Past studies have focused on data collection processes that either mimic traditional auscultation, or make multi-sensor measurements, where the application of specialised recording hardware is required, possibly by expert personnel. This is costly and places limits on the size of the studies (e. g., number of study participants, and recording devices). In this paper, we consider another avenue, that of modelling self-recorded voice samples made using regular smartphones, along with self-reported clinical diagnosis annotations; specifically of asthma. We propose the usage of self-supervised learning that aims to reduce within-class representation redundancy among heterogeneous samples as an auxiliary task to promote robust, bias-free learning. The application of our method achieves an absolute increase of 1.80% in area under the Precision-Recall curve, compared to not using it, and a total of 3.54% compared to our baseline.

Index Terms— Asthma, speech-modelling, self-supervised-learning, redundancy-reduction, dataset-bias-reduction

1. INTRODUCTION AND PRIOR WORK

Given the potential of machine learning (ML) based audio modelling in disease diagnostics [1], a number of studies have attempted to predict pulmonary function of subjects with respiratory illness [2, 3, 4, 5, 6, 7]. In this study, we focus only on voice recordings made using the embedded microphone of personal smartphone devices for predicting whether a subject has been previously diagnosed with asthma, an illness that is estimated to affect 334 million people worldwide [8].

Such recorders have been shown to be sensitive enough to be able to mimic a spirometer in evaluating pulmonary function when used as part of an ML framework [9]. Most importantly, personal smart device data collection allows for large-scale crowdsourcing of audio data along with self-reported clinical annotations, as well as a plethora of other informative metadata (e. g., language, sex, smoking-status, recording hardware type, etc.) [10]. We believe that such an approach can be conducive to a more robust, realistic, and challenging exploration of the asthma diagnosis prediction problem.

1.1. Related work

In a related domain-mismatch study [4], a model trained on data recorded from one device type (e. g., smartphone, or smartwatch) does not necessarily perform well on data recorded using another without a feature adaptation step. The authors offer a solution in which the adaptation step is supervised, i. e., we need to know the recording device used. We believe that assuming any domain knowledge of the test set is a limiting requirement in the case of app-based, crowdsourced data collection. Other studies have side-stepped such design challenges by focusing on subsets of a greater dataset, e. g., on non-smokers [6] or English speakers [7].

The data recording process we use is not mimicking auscultation from the trachea and chest that aims to detect crackles, rhonchi, and wheezes [5], which requires specialised biosensor hardware. We are interested in users actively recording themselves without the need for even a dedicated microphone [11], in the interest of widespread application.

Apart from auscultation-based breath recordings, the authors of [2] model smartphone-based cough sounds to predict the effectiveness of inhaler usage on 55 Chronic Obstructive Pulmonary Disease (COPD) patients. In the study performed in [6], recordings from 26 non-smoking subjects with mild atopic asthma that undergo a specialised methacholine inhalation challenge [12] before being recorded reciting a text, are segmented into speech and breathing clips, and undergo ML modelling to predict abnormal lung function. More comprehensively, the authors of [3] describe a smartphone and smartwatch based dataset that includes various audio recordings, like tidal breathing, coughing, sustained vowels, and both spontaneous and read speech, as well as spirom-

etry measurements (including the methacholine challenge [12]), breath count annotations, and smartphone accelerometer recordings from a total of 228 subjects (asthmatic, COPD, and healthy) used for breath rate prediction. The number of subjects in the above studies is somewhat limited for a robust, widely representative modelling attempt, even when the number of samples per subject is high.

Towards robust learning of representations encoding sequences, a number of Self-Supervised Learning (SSL) methods encourage minimising a measure of distance between representations of samples that are supposed to encode similar content. In completely unsupervised SSL, the distance is between a sample representation, and the representation after having the original sample distorted via data augmentation. For example, the study performed in [13] aims to minimise the mean squared error of keyword speech representations, and the Barlow twins method [14] aims to reduce representation redundancy by encouraging the correlation matrix of the representations to be close to identity, as recently applied in speech modelling [15, 16]. We consider such approaches to be orthogonal to ours, and are also influenced by domain adaptation SSL techniques [17, 18] and bias-free learning [19] in treating various speaker/recording characteristics as bias factors. An example is SelfReg [17], which aims to reduce squared Euclidean distance between the representations of *same-class, different-domain* samples, i. e., disregarding any markers unrelated to asthma recognition. Such a technique requires knowledge of the class to be predicted, but is unsupervised with respect to the domain categories.

1.2. Contributions

We explore past asthma diagnosis prediction via modelling self-recorded speech using personal smartphones and using self-reported labels from a large-scale dataset that is heterogeneous in terms of the clinical, personal, and recording-device metadata of the subjects. We propose a within-class, cross-sample representation redundancy reduction auxiliary task to regularise our main supervision task, influenced by Barlow twins [14], and show its efficacy in a comparative study among recent approaches. The code to preprocess the data and replicate the experiments can be found in: `https://github.com/glam-imperial/asthma-within-class-barlow`.

2. SPEECH MODELLING FOR ASTHMA

The prediction of whether a speech sample has been self-recorded by a person that has been diagnosed in the past with asthma is treated as a binary classification task. A deep learning model $f_W(\cdot)$ with parameter weights W is trained to make such predictions by the supervised minimisation of a binary cross-entropy cost function $\mathcal{L}_{sup}$. We further consider an auxiliary SSL task that acts as a regulariser (with corresponding weight λ_{SSL}, and cost function $\mathcal{L}_{SSL}$) to the main supervised task. The total loss is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_{SSL}\mathcal{L}_{SSL} \tag{1}$$

2.1. Acoustic model definition

What follows is a description of the core model that will perform best in terms of experimental results in Section 5, compared to other baselines. Let $x_i, y_i \in \mathcal{D}$ be the i-th sample/asthma-label pair from the dataset $\mathcal{D}$. y_i is either 0 or 1, denoting a sample from a non-asthmatic or an asthmatic subject, respectively, and x_i is a sequential audio feature representation, in this study a LogMel-Spectrogram. We follow [10] in using a pre-trained VGGish [20] model to process this input, which yields a sequence of latent frame embeddings of length T : $\{h_{i,t}\}$. We use global average pooling to aggregate this sequence into a single embedding h_i that is representative of the voice sample. A further submodel consisting of two dense layers separated by a ReLU nonlinearity yields the logit corresponding to the positive class, and a final sigmoid activation produces the model-estimated probability that sample x_i is positive, denoted by $\hat{y}_i = f_W(x_i)$.

2.2. Reducing within-class representation redundancy

We are inspired by the Barlow twins method [14], as well as by recent SSL-based methods designed for improving cross-domain generalisation [17, 18], like SelfReg [17].

We work on the latent space of audio sample representations h_i, and our goal is to bring closer and reduce redundancy between same-class audio representations, albeit with potentially differing bias factors (i. e., sex, language, smoking status, etc.). This is a 'supervised' SSL approach, in the sense that we are utilising the main task class labels to form positive pairs (see SSL taxonomy in [18]), as well as non-contrastive, as it does not consider negative pairs. Denoting a batch-wise cross-correlation matrix by $C_{k,l}$, our method follows the Barlow twins [14] approach of penalising the model for cross-correlation matrices that deviate from the identity $\mathcal{I}$; that being said, we do not use it for audio representations that arise from different distorted versions of the same sample, and instead encourage bringing closer in a non-redundant manner the representations of potentially heterogeneous data, similar to SelfReg [17]. The $C_{k,l}$ SSL loss are defined as follows:

$$C_{k,l} = \frac{\sum_b h^A_{b,k} h^B_{b,l}}{\sqrt{\sum_b h^A_{b,k}}\sqrt{\sum_b h^B_{b,l}}}, \tag{2a}$$

$$\mathcal{L}_{SSL} = \sum_{k=l}(1 - C_{k,l})^2 + \lambda_{RR}\sum_k \sum_{l \neq k} C^2_{k,l}, \tag{2b}$$

where k, l are the representation dimension indices, b is the index of the batch, and λ_{RR} is the regularising factor for the

Partition	Positives	Negatives	All
training all-three	1,694	3,352	5,046
training all-speech	2,140	3,582	5,722
training all-breath	1,944	3,505	5,449
training all-cough	2,056	3,480	5,536
development	862	1,624	2,486
testing	852	1,618	2,470

Table 1. Partition sizes for the asthma diagnosis prediction dataset. These partitions are user-independent. We denote by *all-three* the count of instances where users recorded all three modalities. We further include counts of all available samples of a particular modality. In development and testing we only include instances with all three modalities available.

redundancy reduction term. The way we implement this, is that we split each batch b into two equal size parts A, B. The cross-correlation matrix is calculated between these two half-batches. The effective reduction in batch size is acceptable as the success of the redundancy reduction approach does not depend on large batch sizes [14].

3. SELF-REPORTED ASTHMA DATASET

We base our analysis on the dataset that was collected in the context of [10] for COVID-19 modelling based on human speech, breath, and cough; we reformulate the partitions and focus, instead, on predicting which samples are made by asthmatic users. We wanted to avoid training with imbalanced data, and, in fact, preliminary experiments showed diminished performance when all negatives were used, even with upsampling and/or positive class up-weighting. We, thus, elected to use a subset of the asthma-negative users in roughly a 1:2 positive-negative user ratio, for a total of 6,774 users. The dataset statistics are summarised in Table 1.

The dataset from [10] contained numerous metadata related to the users, apart from their asthma diagnosis status. The ones we considered as related to this study are: age group, sex, smoking status, language, the outcome of COVID-19 testing (if any), COPD diagnosis, other pulmonary disease, and other symptoms (possibly of COVID-19), such as dry or wet cough, sore throat, and short breath, as well as whether they have used the android or the iOS version of the app to make their recording. We did not use the samples from the web app described in [10], because there was no user identification, and we wanted to avoid placing samples from the same user in different partitions.

In the interest of a fair experimental comparison, such that the models do not end up memorising such biasing characteristics in training, and such that we do not explicitly exclude any of them from the development and testing partitions, we aimed for the formation of fair partitions that preserve the percentage mixtures of the above metadata in all partitions. We used a recent algorithm initially designed for generating stratified partitions of a multilabel dataset [21]. However, instead of using it as intended, i. e., with multiple prediction labels per data sample, we supplied it with the multiple metadata categories that are associated with each user, thus generating *stratified user partitions*, with respect to metadata.

	Method	AU-PR	AU-ROC	F1	R
VGGish	all-speech	*41.55*	*58.27*	*57.08*	*57.28*
	all-breath	36.00	52.07	52.02	52.81
	all-cough	38.30	55.48	53.36	54.72
baseline	VGGish+mt	39.22	56.29	54.94	55.33
	VGGish+gr	39.32	55.93	53.98	54.97
	ResNet	40.44	57.57	56.93	56.90
hom	VGGish	*43.29*	59.46	*57.99*	58.21
	VGGish+mt	42.63	58.87	57.23	57.27
	VGGish all-3	42.46	*60.30*	57.52	*58.60*
SSL	Norm (.1)	44.61	60.06	58.64	58.68
	RR (.1)	**45.09**	**60.47**	**59.00**	**59.04**
	RR (.2)	43.73	60.14	57.74	57.67

Table 2. Results on the asthma diagnosis test set.

4. EXPERIMENTAL SETUP

We calculate LogMel-Spectrograms with 64 Mel filterbanks, as per [10]. For data augmentation, we use: SpecAugment [22] with 2 time and 1 frequency masks of size 24 and 16, respectively, and input jitter sampled from a zero-centred normal distribution with standard deviation equal to 1e-6. We use a batch size of 8 (split into two sub-batches of 4 for the redundancy reduction task). We use λ_{RR} equal to 5e-3.

On the development set, we monitor Area Under the Precision- Recall curve (non-interpolated AU-PR) of the positive class with a patience of 200 epochs for model selection, and use this model in testing. We also report Area Under the Receiver Operator Characteristic curve (AU-ROC), and macro-averaged F1 (F1) and recall (R) scores. For F1, we identify the probability class threshold that maximises F1 on the development set, and we utilise the same one for the test set. This way, we observe at least as good performance on F1, of up to 4 absolute points. In all cases, we performed 3 trials for which we report the mean outcome.

5. RESULTS & DISCUSSION

All experimental results are summarised in Table 2. In the upper block (**VGGish**), we make a comparison between single modalities, using the pre-trained VGGish based architecture described in Sub-section 2.1. We use all available samples per modality, as shown in Table 1. We see that the usage of the **speech** modality is the most informative, and continue as such, unless specified.

Next, we attempted a **baseline** comparison with another

architecture (i. e., instead of the pre-trained VGGish, we used an untrained *ResNet* model that performed well in [23]), as well as two multi-task regularisation attempts. *VGGish+mt* also attempts to model the various user metadata (summarised in Sub-section 3) pertaining to the samples; the sum regularisation weight for all these tasks is conservatively set to .2, and they all share the same weight amongst them. *VGGish+gr* is a similar setup, albeit there is a gradient reversal layer before the prediction blocks for the metadata, as proposed in [19] for bias-free learning. We see that none manage to surpass the VGGish-based baseline.

In order to go through with the SSL auxiliary task related experiments, we need batches that are homogeneous (**hom**) in terms of the main prediction class, i. e., either all positives, or all negatives, so we re-run select baselines in that setup. *VGGish* is the so-far best baseline using all speech samples, *VGGish+mt* refers to the addition of the multitask regularisation, and we also perform an experiment where we use all available samples from all three modalities (*all-3*). We use a separate VGGish model per modality, and we aggregate the three h_i^{mod} via max pooling before applying a common dense prediction block. For *VGGish* and *VGGish+mt* we see that this kind of same-class batching has yielded an improvement in all measures, thus formulating a stronger baseline, for stricter comparison with the SSL methods, although *VGGish* still performs better than the latter. As for *VGGish all-3*, it manages to surpass *VGGish* in two out of four measures (non-strict improvement), albeit at a multiple execution time cost (not just for processing three modalities, but also requiring roughly twice the epochs until patience-based training termination), and, as such, we opted not to continue with it.

In the best performing variation cluser of the study, we experiment with **SSL** based auxiliary regularisation. Our proposed method, denoted by **RR** is used both with λ_{SSL} equal to .1 and .2. We further compare with a squared Euclidean distance loss (*Norm (.1)*) instead of the one based on cross-correlation, including the use of a projection layer, as per [17]. We see that both *Norm (.1)* and *RR (.1)* outperform the stronger, homogeneous batch baseline in all measures, with *RR (.1)* achieving the best results in this study. We also attempted a higher focus on the SSL auxiliary task, in *RR (.2)*, however it managed to surpass *hom-VGGish* in only two out of four measures. This is promising still, however, indicative that in this auxiliary/regularisation setup, the parameterisation sweet-spot should lean towards the main task.

5.1. What works best for recognising asthmatic speech?

As an outcome of this study, we should take that for predicting whether a voice sample is produced by an asthmatic subject, the speech modality is the most informative. This is in contrast with the study performed in [6], where the use of breath sounds outperformed speech, and also [10], where cough outperformed both speech and breath sounds. That being said, the former study was limited to 323 voice samples from 26 subjects, whereas the latter study was focused on predicting COVID-19. In both these studies and ours, the combination of all modalities was comparable to the performance of the best modality. In terms of bias-free learning, we verify that SSL-based approaches are better than older baselines, like multi-task learning (which may overfit to undesirable characteristics), or gradient-reversal bias-free learning [19] (which may hinder learning of useful features). We show here that the cross-correlation based redundancy reduction approach proposed in the Barlow twins study [14] is indeed useful not just in its initial, completely unsupervised, augmentation-based formulation, but also in a supervised, positive-pairs, non-contrastive learning SSL framework in this audio modelling task. We use a relatively small batch size, which is halved in the way we propose to use redundancy reduction, and the results are still very promising, since redundancy reduction is known to not require large batches.

6. CONCLUSIONS & FUTURE WORK

We first explored the task of predicting whether an audio sample has been produced by a subject suffering from asthma, in a self-reported, self-recording smartphone setup, which allows for the possibility of attracting more, and more heterogeneous users compared to similar studies. Our redundancy reduction based, self-supervised learning approach as an auxiliary task, has exhibited improved performance and shown that such techniques can also be used for addressing such heterogeneous datasets. That being said, our best performing method can still only be considered to address one of the challenges presented by this dataset; that of heterogeneity. Promising avenues for future extensions would be the combination of our SSL approach with existing SSL methods with well-known performance benefits, like augmentation-based Barlow twins for audio [16, 15], or the addition of a full self-supervised pre-training step [14], as well as introducing meta-learning methods for addressing the potential of noisy labels [24] introduced by the self-reported nature of the dataset.

7. ACKNOWLEDGMENTS

This research was made possible by Grant Number EP/W002477/1 from the UK Engineering and Physical Sciences Research Council (EPSRC).

8. REFERENCES

[1] Manuel Milling, Florian B Pokorny, Katrin D Bartl-Pokorny, and Björn W Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in Digital Health*, vol. 4, 2022.

[2] Anthony Windmon, Sriram Chellappan, and Ponrathi R Athilingam, "Evaluating the effectiveness of inhaler use

among copd patients via recording and processing cough and breath sounds from smartphones," in *International Conference on Mobile Computing, Applications, and Services*. Springer, 2020, pp. 102–120.

[3] Md Mahbubur Rahman, Mohsin Yusuf Ahmed, Tousif Ahmed, Bashima Islam, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao, "Breatheasy: Assessing respiratory diseases using mobile multimodal sensors," in *Proc. 2020 International Conference on Multimodal Interaction*, 2020, pp. 41–49.

[4] Mohsin Y Ahmed, Li Zhu, Md Mahbubur Rahman, Tousif Ahmed, Jilong Kuang, and Alex Gao, "Device invariant deep neural networks for pulmonary audio event detection across mobile and wearable devices," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 5631–5637.

[5] Honorata Hafke-Dys, Barbara Kuźnar-Kamińska, Tomasz Grzywalski, Adam Maciaszek, Krzysztof Szarzyński, and Jedrzej Kociński, "Artificial intelligence approach to the monitoring of respiratory sounds in asthmatic patients," *Frontiers in physiology*, p. 1980, 2021.

[6] Md Alam, Albino Simonetti, Raffaele Brillantino, Nick Tayler, Chris Grainge, Pandula Siribaddana, SA Nouraei, James Batchelor, M Sohel Rahman, Eliane V Mancuzo, et al., "Predicting pulmonary function from the analysis of voice: a machine learning approach," *Frontiers in digital health*, p. 5, 2022.

[7] Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al., "Sounds of covid-19: exploring realistic performance of audio-based digital testing," *NPJ digital medicine*, vol. 5, no. 1, pp. 1–9, 2022.

[8] Oladunni Enilari and Sumita Sinha, "The global impact of asthma in adult populations," *Annals of global health*, vol. 85, no. 1, pp. 2, 2019.

[9] Eric C Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N Patel, "Spirosmart: using a microphone to measure lung function on a mobile phone," in *Proc. 2012 ACM Conference on ubiquitous computing*, 2012, pp. 280–289.

[10] Tong Xia, Dimitris Spathis, J Ch, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro 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.

[11] Khanh Nguyen-Trong, "An adaptive method for classification of noisy respiratory sounds," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, 2021, pp. 6–11.

[12] Christopher L Grainge, Laurie CK Lau, Jonathon A Ward, Valdeep Dulay, Gemma Lahiff, Susan Wilson, Stephen Holgate, Donna E Davies, and Peter H Howarth, "Effect of bronchoconstriction on airway remodeling in asthma," *New England Journal of Medicine*, vol. 364, no. 21, pp. 2006–2015, 2011.

[13] Jian Luo, Jianzong Wang, Ning Cheng, Haobin Tang, and Jing Xiao, "Speech augmentation based unsupervised learning for keyword spotting," *arXiv preprint arXiv:2205.14329*, 2022.

[14] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12310–12320.

[15] Xin Jing, Meishu Song, Andreas Triantafyllopoulos, Zijiang Yang, and Björn W Schuller, "Redundancy reduction twins network: A training framework for multi-output emotion regression," *arXiv preprint arXiv:2206.09142*, 2022.

[16] Mohammad Mohammadamini, Driss Matrouf, Jean-François A Bonastre, Sandipana Dowerah, Romain Serizel, and Denis Jouvet, "Barlow twins self-supervised learning for robust speaker recognition," in *Interspeech*, 2022.

[17] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee, "Selfreg: Self-supervised contrastive regularization for domain generalization," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9619–9628.

[18] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu, "Pcl: Proxy-based contrastive learning for domain generalization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7097–7107.

[19] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9012–9020.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan 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.

[21] Piotr Szymański and Tomasz Kajdanowicz, "A network perspective on stratification of multi-label data," in *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR, 2017, pp. 22–35.

[22] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," 2019.

[23] Georgios Rizos, Jenna Lawson, Zhuoda Han, Duncan Butler, James Rosindell, Krystian Mikolajczyk, Cristina Banks-Leite, and Björn W Schuller, "Multi-attentive detection of the spider monkey whinny in the (actual) wild.," 2021.

[24] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais, "Meta label correction for noisy label learning," in *Proc. AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 11053–11061.