FAIRNESS AND UNDERSPECIFICATION IN ACOUSTIC SCENE CLASSIFICATION: THE CASE FOR DISAGGREGATED EVALUATIONS

Andreas Triantafyllopoulos[1,2], *Manuel Milling*[2], *Konstantinos Drossos*[3], *Björn W. Schuller*[1,2,4]

[1]audEERING GmbH, Gilching, Germany
[2]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]Audio Research Group, Tampere University, Tampere, Finland
[4]GLAM – Group for Audio, Language, & Music, Imperial College, London, UK
atriant@audeering.com

ABSTRACT

Underspecification and *fairness* in machine learning (ML) applications have recently become two prominent issues in the ML community. Acoustic scene classification (ASC) applications have so far remained unaffected by this discussion, but are now becoming increasingly used in real-world systems where fairness and reliability are critical aspects. In this work, we argue for the need of a more holistic evaluation process for ASC models through *disaggregated evaluations*. This entails taking into account performance differences across several factors, such as city, location, and recording device. Although these factors play a well-understood role in the performance of ASC models, most works report single evaluation metrics taking into account all different strata of a particular dataset. We argue that metrics computed on specific sub-populations of the underlying data contain valuable information about the expected real-world behaviour of proposed systems, and their reporting could improve the transparency and trustability of such systems. We demonstrate the effectiveness of the proposed evaluation process in uncovering underspecification and fairness problems exhibited by several standard ML architectures when trained on two widely-used ASC datasets. Our evaluation shows that all examined architectures exhibit large biases across all factors taken into consideration, and in particular with respect to the recording location. Additionally, different architectures exhibit different biases even though they are trained with the same experimental configurations.

Index Terms— acoustic scene classification, evaluation, fairness, ethics, transparency

1. INTRODUCTION

Acoustic scene classification (ASC) has been established as a central task of artificial auditory intelligence, as exemplified by its prominent place in the DCASE challenge and workshop series [1, 2] and a generally broad accumulation of literature [3, 4, 5, 6, 7]. Overall, model performance has substantially improved through the years, and datasets have accordingly evolved to accommodate new challenges by incorporating factors shown to impact model performance. For example, the exact geographical location of the recordings was identified as an important factor early on, with datasets accordingly adapted by keeping data from the same location in the same partitions [1, 2]. The TUT Urban Acoustic Scenes 2018 Mobile dataset additionally introduced the recording device as a separate factor [8], with the development set consisting of multiple recording devices, and the evaluation set including an extra, unseen device. Finally, the TAU Urban Acoustic Scenes 2019 dataset highlighted the importance that the city of origin has by introducing data from two additional cities in the evaluation set [8].

In general, the community is aware of the influence that recording devices and location have on model performance [9, 10]. Most works approach these factors from the perspective of *domain mismatch* [11]: different cities, locations, and devices, result in slightly different input representations, and the difference needs to be accounted for to improve overall performance. Several approaches have been proposed to mitigate the problem, largely drawing from the wide literature of domain adaptation techniques [11] adapted for the ASC problem [12, 13, 14], or specifically taking steps to mitigate the effects of city and device [15, 16].

In this work, we adopt a different perspective: we argue that those factors deserve a prominent place in the evaluation of ASC systems as they reveal important insights about the behaviour of trained models. To do that, we adopt the language of recent works in the machine learning (ML) *fairness* literature. In particular, we propose *disaggregated evaluations*, a concept highlighted by Mitchell *et al.* [17] as a means to expose the effects that these underlying factors have on system performance. Disaggregation, which corresponds to breaking down an evaluation to more fine-grained levels of analysis, can be done both in a *unitary* (how performance is affected by each factor independently) and in an *intersectional* way (how performance is affected by combination of factors). For the task of ASC, we consider the three aforementioned factors, namely location, city, and device, as warranting a closer investigation. This choice is primarily motivated by availability (the existing metadata is already there) and community awareness (past works take them into account).

The rest of this document is organised as follows. In Section 2, we formulate our research question by discussing fairness and underspecification for ASC. Our methodological approach, including a description of the data and deep learning (DL) architectures used in our experiments, is outlined in Section 3. The results and a discussion of our disaggregated evaluations are presented in Section 4. Finally, we summarise our findings in Section 5.

2. FAIRNESS AND UNDERSPECIFICATION IN ASC

The success and increased usage of ML, and in particular DL, systems in commercial applications has led to rising concerns towards discriminating biases exhibited by ML applications, for instance

based on race [18]. Especially in the case of DL, a lack of interpretability can often be observed [19], thus posing additional challenges to discover and mitigate said biases. Even though ASC models are not widely considered high risk applications, their increasing usage in smart city [20], security [21], elderly monitoring [22], and autonomous driving [23] applications means they may soon (or already) be part of critical decision making systems, thus making fairness a critical consideration for those algorithms.

Of the three factors, the recording device is perhaps the most benign; it is hard to justify why an ASC system that only works for specific devices should raise ethics concerns, although low-income groups could be excluded if data are only collected with high-end equipment. On the other hand, city and location (which corresponds e. g. to specific neighbourhoods) pose potentially bigger problems; a security application should work *equally well* for all citizens irrespective of where they reside, and autonomous driving systems should maintain a standard of performance irrespective of where the vehicle currently is. There is a already a rich body of work in social sciences discussing inequality across different neighbourhoods on income, health, and other socioeconomic factors [24], which an unreliable system may inadvertently exacerbate. This could have adverse effects against people living in those neighbourhoods, and may disproportionately affect minorities in demographically segregated communities. Therefore, we anticipate that explicitly communicating disaggregated performance with respect to all three factors would enhance trustability in ASC systems used in real-life environmental sensing applications.

Disaggregated evaluations can also be viewed under the perspective of recent research on the *underspecification* of ML architectures [25], which corresponds to the fact that several architectures yielding similar in-domain performance nevertheless exhibit different behaviour during system deployment. This undesired property may have negative consequences on the reliability and trustability of ASC systems. For example, if a person using an ASC system observed substantially different performance when visiting different neighbourhoods of the same city, they might eventually lose their trust in system performance and stop using it. As ASC architectures increasingly find their way into more real-life applications, the need to address this issue becomes more pressing. Our evaluation reveals that different architectures yielding almost equivalent performance in standard aggregated evaluations exhibit different behaviour across different sub-populations of the herein examined datasets, thus illustrating that underspecification is also a problem for ASC applications. This shows that disaggregated evaluations can be a useful tool for practitioners that need to select among a pool of candidate models.

3. METHODOLOGICAL APPROACH

Our approach consists of the following steps. First, we train several deep neural network (DNN) models on the training set of each of the datasets examined here. Each model is trained for 60 epochs using stochastic gradient descent (SGD) with a Nesterov momentum of 0.9, a learning rate of 0.001, and a batch size of 64. For all experiments, we use log Mel spectrograms with 64-bins as input features, extracted with a window size of 32 ms and a hop size of 10 ms. These hyper-parameters were fixed a priori for all models and not optimised during our experiments. Each model is trained with 5 random seeds to mitigate the effect of randomness.

Our experiments are conducted on the TUT Urban Acoustic Scenes 2018 and TUT Urban Acoustic Scenes 2018 Mobile data sets [8], which will be henceforth referred to as TUT-Urban and TUT-Mobile for brevity. Both datasets contain data from 10 acoustic scenes recorded across several locations of 6 different European cities. TUT-Urban contains 8640 stereo samples recorded at 48 kHz with a single high-quality recording device (Soundman OKM II Klassik/studio A3), whereas TUT-Mobile additionally contains 720 samples from each of two additional low-quality recording devices (Samsung Galaxy S7 and iPhone SE). In the case of TUT-Mobile all data are stored as mono recordings at 16 kHz.

All models are first evaluated in the standard, aggregated way by computing a single accuracy value, and subsequently assessed using unitary and intersectional evaluations as described below. We begin with unitary evaluations, where each factor is considered in isolation. For city and device, where we have only 6 and 3 different groups, respectively, we simply report the accuracy for each group. The location factor is more complicated, as we have 83 different locations in the test set, thus making it hard to visualise results. Moreover, whereas for each city and device we have all classes available, each location corresponds to exactly one class, thus making accuracy an inappropriate metric for that evaluation. To overcome these problems, we compute the F_1 score, which is the harmonic mean of precision and recall, for the class corresponding to each location and further normalise the per location F_1 score, F_1^l, by the overall F_1 score for that architecture.

Intersectional evaluations are in turn conducted by taking into account two, or more, factors. Due to space limitations, we only consider results for two pairs of factors: the variation of cities across different devices and the variation across locations in different cities. For the first case, we report the accuracy for each combination of factors. For the latter case, we compute the F_1^l score for each location as in the unitary case, but now normalise over the F_1 score for each city, F_1^c.

As DL architectures, we use 5 standard DNN models that belong to different architecture families. **FFNN**: as the most simple architecture we choose a feed-forward neural network with three hidden layers of decreasing sizes, 300, 200, and 100 units with a rectified linear unit (ReLU) activation function. The inputs of of the network are the flattened log Mel spectrograms. **TDNN**: we further employ a time-delay neural network (TDNN) architecture. First introduced by [26] with the aim of learning temporal relationships, TDNNs have recently seen great success in the field of speaker identification [27]. Our TDNN architecture is identical to the DNN architecture described as the *x-vector system* in [27]. **CNN6**, **CNN10**, **CNN14**: the final architectures considered in our experiments are CNN-based and were recently introduced by Kong *et al.* [28] in the context of audio pattern recognition. The three architectures have a total of 6, 10, and 14 layers, respectively, excluding pooling layers after convolutional layers, and take Mel-spetrograms as inputs. The final two layers of each network are fully connected.

4. RESULTS AND DISCUSSION

Our unitary evaluation results for different cities are presented in Table 1, along with the standard aggregated metrics. We show model accuracy for each factor in isolation, and also report the standard deviation over all factors. F_1 results for different locations in TUT-Urban are shown in Figure 1, where we show box-and-whisker plots of the normalised F_1 scores. We omit unitary results for different devices as they can be inferred from the intersectional results in Table 2; as expected, all architectures perform best on the high-quality device A, for which we also have the most data, while doing worse

Table 1: Aggregated and unitary disaggregated evaluations considering different cities in isolation. For the aggregated evaluation, we show accuracy[%] for all test data for TUT-Urban and TUT-Mobile. For the unitary disaggregated evaluations, we show accuracy[%] on different cities for each architecture, as well as its standard deviation (σ) over the different cities. Results are averaged across 5 different runs.

	TUT-Urban					**TUT-Mobile**				
	FFNN	**TDNN**	**CNN6**	**CNN10**	**CNN14**	**FFNN**	**TDNN**	**CNN6**	**CNN10**	**CNN14**
Aggregated	52.8	57.2	**68.7**	67.7	66.3	53.1	54.7	**66.8**	66.3	63.6
City	**Disaggregated evaluations**									
Barcelona	52.9	61.7	**64.8**	60.9	58.9	55.9	56.4	**61.3**	57.9	57.6
Helsinki	56.1	61.3	**70.2**	67.3	63.5	50.8	57.1	**67.9**	66.7	58.3
London	51.1	61.7	71.6	**74.2**	71.5	52.7	59.2	70.1	**72.0**	70.8
Paris	45.5	53.8	62.0	61.0	**62.4**	45.8	54.1	60.1	59.7	**60.8**
Stockholm	53.0	47.4	**73.1**	68.4	68.2	56.2	46.9	**72.5**	68.3	67.3
Vienna	59.8	57.9	69.9	**74.0**	73.0	58.8	54.3	68.0	**72.4**	65.7
σ	4.4	5.2	**3.9**	5.4	5.1	4.3	**3.9**	4.5	5.6	4.9

Table 2: Intersectional evaluations considering recording device and city in combination for the TUT-Mobile dataset. We show accuracy[%] for each combination of city and device. Cites are Barcelona (B), Helsinki (H), London (L), Paris (P), Stockholm (S), and Vienna (V). The best performing architecture value per city and device is marked by boldface. Results are averaged across 5 different runs.

	Device A							**Device B**							**Device C**						
Model	**B**	**H**	**L**	**P**	**S**	**V**	σ	**B**	**H**	**L**	**P**	**S**	**V**	σ	**B**	**H**	**L**	**P**	**S**	**V**	σ
FFNN	56.1	53.1	53.0	46.9	56.5	60.5	**4.2**	54.8	38.1	54.8	29.0	56.7	48.7	10.2	54.1	31.0	45.2	46.5	52.0	48.7	7.5
TDNN	57.9	60.4	62.9	57.0	47.5	56.3	4.8	44.4	38.1	26.5	34.8	53.3	41.3	8.3	46.7	30.3	34.2	32.9	31.3	43.3	**6.2**
CNN6	**61.9**	**70.2**	71.2	**63.1**	**73.4**	69.7	**4.2**	**55.6**	**58.7**	**67.7**	38.1	65.3	**61.3**	9.7	**57.8**	44.5	56.8	39.4	**66.0**	54.7	8.8
CNN10	58.8	68.9	**73.7**	62.2	69.1	**74.2**	5.6	53.3	56.1	64.5	41.9	**68.0**	**61.3**	8.5	49.6	**46.5**	53.5	42.6	56.7	**62.7**	6.6
CNN14	58.7	60.5	71.9	62.4	68.5	67.2	4.7	48.1	47.1	64.5	**43.2**	61.3	58.0	**7.9**	51.9	38.1	**60.0**	**55.5**	55.3	56.0	7.0

on the lower quality and less populous B and C devices. Location results on TUT-Mobile are also omitted due to space limitations but exhibit the same trend as those on TUT-Urban.

Table 1 can be read both horizontally, thus emphasising which model works best for a specific factor, and vertically, where we are interested in how a specific model performs across different factors. Overall, CNN6 is showing the strongest performance, followed by CNN10 and CNN14, with TDNN and FFNN performing substantially worse. Furthermore, CNN6 exhibits relative stability across both cities and devices. However, it is not the best choice for all cities; in both datasets, CNN10 is outperforming it for London and Vienna, and CNN14 for Paris, though the latter only marginally.

Of more interest is the vertical interpretation of Table 1. We observe that different architectures exhibit a different ordering when it comes to performance per city. In TUT-Urban for example, different architectures yield their best performance on different cities: FFNN on Vienna, TDNN on Barcelona and London, CNN6 on Stockholm, CNN10 on London and Vienna, and CNN14 on Vienna. Another interesting case is Stockholm, where CNN6 shows its best performance and TDNN its worst. Conversely, Vienna, where FFNN, CNN10, and CNN14 show (near-)best performance for TUT-Urban, is showing mediocre results for CNN6 and TDNN.

For TUT-Mobile, these results are better visualised in Figure 3 which shows the range of F1 scores per location for the different cities. Notable differences exist; TDNN shows worse performance on Stockholm than Paris, whereas all other architectures show the opposite trend. CNN6 and CNN10, which are almost equivalent in terms of aggregated performance, also exhibit differences, in particular for Stockholm and Vienna. Interestingly, TDNN and FFNN deviate substantially from the other three architectures, which are more closely clustered together, indicating that models from the same family exhibit more similar behaviour. These observations illustrate that the inductive biases introduced by each architecture manifest themselves as different behaviours on different strata of each dataset, which is in line with recent research on inductive biases [29, 30].

Figure 1 additionally shows that location is a very important factor when it comes to system performance, with some locations

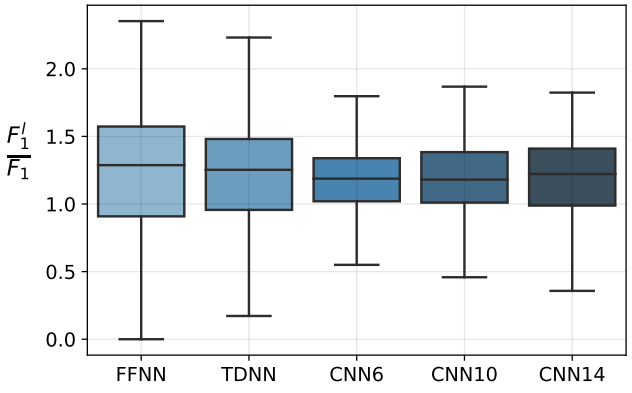

Figure 1: Distribution of relative F_1 score on different locations of TUT-Urban for all architectures. Box plots show median and inter-quartile range of relative F_1 score.

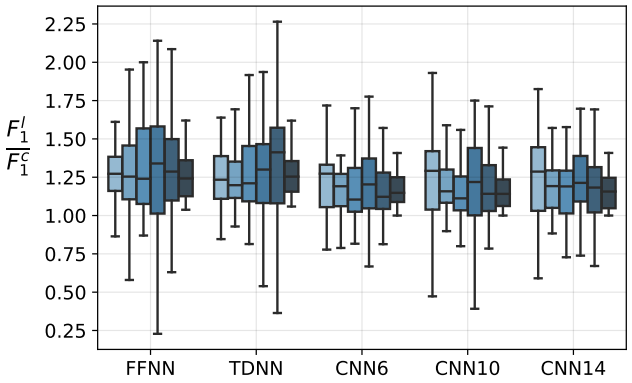

Figure 2: Intersectional analysis of model performance with respect to both city and location for TUT-Urban. For each architecture, cities are (from left to right): Barcelona, Helsinki, London, Paris, Stockholm, Vienna. Box plots show median and inter-quartile range of relative F_1 score with respect to different locations within each city.

exhibiting almost half the aggregated system performance. Such behaviour is highly undesirable because an ASC system deployed across different locations will consistently exhibit subpar performance for some of them, with the risk to equal and fair access to service that this entails. We note that most locations seem to exhibit better than average performance (the F_1 ratio is bigger than 1). This is caused by the fact that the worst performing locations happen to have more samples, thus having a bigger influence on aggregate performance.

Intersectional results are shown in Table 2 for the combination of city and device, and in Figure 2 for the combination of city and location. The differences amongst all cities and all devices were found significant for all architectures using Kruskal-Wallis omnibus H-tests for each factor and architecture, respectively. This shows that, in general, both factors have a large effect on model performance. In addition, Table 2 and Figure 2 both show that different architectures exhibit different behaviour on different strata of the two datasets, even though they were trained on identical settings. Over-

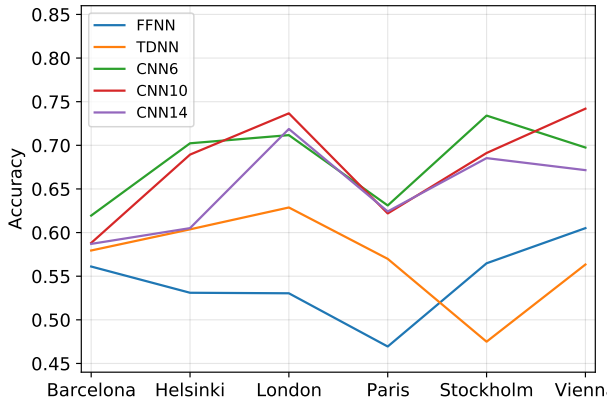

Figure 3: Accuracy for each city and architecture on TUT-Mobile.

all, CNN6 is again showing the strongest performance for most, though not all, combinations, followed by CNN10. In terms of individual factors, Paris is showing the biggest drop in performance when switching from device A to device B for all architectures, indicating that the domain shift introduced by different devices is more adversely impacting this city.

The most interesting case is TDNN, which is showing its best and worst performance on London and Stockholm for device A, respectively, but shows the exact opposite for device B, where the best performance is obtained for Stockholm and the worst for London. In fact, the performance of TDNN on Stockholm is far better for device B than for device A, even though the latter has far more samples and should thus lead to better performance.

5. CONCLUSION

In this work, we argue for the need of disaggregated unitary and intersectional evaluations for the task of ASC. Our proposed evaluation methodology reveals that several baseline architectures exhibit different behaviour even though they are trained with similar settings. This illustrates that ASC models trained on the examined datasets suffer from the underspecification problem, which heavily impacts the development of reliable and trustworthy systems. In the future, we intend to further investigate this problem under the perspective of inductive biases introduced by each architecture [30].

Moreover, our work raises interesting questions on the fairness of ASC applications. The architectures examined here exhibit a bias with respect to different cities, locations, and devices. If these architectures were deployed in a real-world setting, this would translate to non-uniform behaviour over these different factors. This poses a risk to fair and equitable use of ML resources. We believe this important point needs to be addressed as ASC models are being increasingly integrated in intelligent decision making systems.

6. ACKNOWLEDGMENT

Part of the work leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 957337, project MARVEL.

7. References

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 85–92.

[2] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb. 2018, ISSN: 2329-9290. DOI: 10.1109/TASLP.2017.2778423.

[3] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI Signal Processing*, vol. 20, Apr. 1998. DOI: 10.1023/A:1008066223044.

[4] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[5] A. Rakotomamonjy, "Supervised representation learning for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1253–1265, 2017. DOI: 10.1109/TASLP.2017.2690561.

[6] K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, and B. Schuller, "Wavelets revisited for the classification of acoustic scenes," Nov. 2017.

[7] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, 2018.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Nov. 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840.

[9] ——, *Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups*, New York University, NY, USA, Oct. 2019.

[10] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: Generalization across devices and low complexity solutions," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2020, pp. 56–60.

[11] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, *et al.*, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, p. 137, 2007.

[12] S. Gharib, K. Drossos, E. Cakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Nov. 2018, pp. 138–142.

[13] K. Drossos, P. Magron, and T. Virtanen, "Unsupervised adversarial domain adaptation based on the wasserstein distance for acoustic scene classification," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 259–263. DOI: 10.1109/WASPAA.2019.8937231.

[14] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, "Caanet: Conditional atrous cnns with attention for explainable device-robust acoustic scene classification," *IEEE Transactions on Multimedia*, 2020.

[15] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," DCASE2019 Challenge, Tech. Rep., Jun. 2019.

[16] M. Kośmider, "Calibrating neural networks for secondary recording devices," DCASE2019 Challenge, Tech. Rep., Jun. 2019.

[17] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[18] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[19] N. Burkart and M. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, Jan. 2021. DOI: 10.1613/jair.1.12228.

[20] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*, Springer, 2018, pp. 373–397.

[21] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, IEEE, 2005, pp. 158–161.

[22] R. Mégret, V. Dovgalecs, H. Wannous, S. Karaman, J. Benois-Pineau, E. El Khoury, J. Pinquier, P. Joly, R. André-Obrecht, Y. Gaëstel, *et al.*, "The immed project: Wearable video monitoring of people with age dementia," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1299–1302.

[23] M. K. Nandwana and T. Hasan, "Towards smart-cars that can listen: Abnormal acoustic event detection on the road.," in *Interspeech*, 2016, pp. 2968–2971.

[24] M. Wen, C. R. Browning, and K. A. Cagney, "Poverty, affluence, and income inequality: Neighborhood economic structure and its implications for health," *Social science & medicine*, vol. 57, no. 5, pp. 843–860, 2003.

[25] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *arXiv preprint arXiv:2011.03395*, 2020.

[26] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.

[27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5329–5333.

[28] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[29] G. Ortiz-Jimenez, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "Neural anisotropy directions," *arXiv preprint arXiv:2006.09717*, 2020.

[30] G. Ortiz-Jimenez, I. F. Salazar-Reque, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "A neural anisotropic view of underspecification in deep learning," *arXiv preprint arXiv:2104.14372*, 2021.