ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097238

SEMI-SUPERVISED LEARNING WITH PER-CLASS ADAPTIVE CONFIDENCE SCORES FOR ACOUSTIC ENVIRONMENT CLASSIFICATION WITH IMBALANCED DATA

Luan Vinícius Fiorio[⋆], *Boris Karanov*[⋆], *Johan David*[†], *Wim van Houtum*[⋆†],
Frans Widdershoven[†], *Ronald M. Aarts*[⋆]

[⋆] Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands
[†] NXP Semiconductors, High Tech Campus 46, Eindhoven 5656 AE, The Netherlands

ABSTRACT

In this paper, we concentrate on the per-class accuracy of neural network-based classification in the context of identifying acoustic environments. Even a fully supervised learning framework with an equal amount of data for each class can lead to significant differences in class accuracies. This is then amplified by semi-supervised learning using naturally imbalanced data. To address this problem, we propose an adaptive method for pseudo-label selection via a straightforward optimization of the validation accuracy per class, aimed specifically at reducing the variance between different classes. The proposed method is general and can be applied for both maximum probability and entropy-based confidence criteria. Compared to fully supervised learning as well as state-of-the-art methods for pseudo-labeling, it achieves the lowest variances of per-class accuracy and the highest accuracies of the minority classes when tested on common publicly available environment sound databases.

Index Terms— Semi-supervised learning, deep learning, neural networks, pseudo-labeling, sound classification

1. INTRODUCTION

Research on machine learning for audio processing has drawn growing attention in recent years [1, 2, 3]. Such techniques allowed for more powerful denoising and dereverberation methods [4, 5], as well as beamforming [6]. Especially for sound classification, feedforward, recurrent, and convolutional neural networks represent the most used structures [7]. Recently, environmental sound classification using machine learning had become relevant for medical applications such as hearing aids [8, 9], avoiding the amplification of ambient noises, making up for better speech comprehension and overall listening experience.

For sound classification tasks, labeled data is usually scarce and the labeling process can be time-consuming and expensive, which is especially critical when creating larger datasets [10]. Unlabeled data is also easier to collect, e.g., on a device [11]. In these situations, semi-supervised learning (SSL) [12] can be applied, which aims to use both labeled and unlabeled data during training. The lowest complexity SSL method is known as pseudo-labeling [13] and is of particular interest for applications in power-constrained devices.

However, unlabeled data is often imbalanced, causing strong biases in the neural network model [14]. For pseudo-labeling, some approaches have tried to compensate for imbalanced data by using adaptive probability thresholding [15], where, though, a strong dependence on data in the majority class is still maintained. Others tried to achieve a better classification by using the entropy of the network's softmax output [16], but not considering per-class adaptability.

In this paper, we focus on SSL for environmental sound classification. We propose a low-complexity adaptive framework of defining per-class confidence scores where each class is independently treated according to a specified metric. We show that this approach can be applied to both probability and entropy-guided implementations of pseudo-labeling. It achieves better class balance and higher accuracy as well as F1 score for the minority classes when compared to other available methods.

2. SEMI-SUPERVISED LEARNING FRAMEWORK WITH ADAPTIVE CONFIDENCE SCORES

In the supervised learning problem, the data used for training is composed of input and desired (labeled) output, while for unsupervised learning, no information about the desired output is available. The *semi-supervised learning* framework combines the two cases [12]. Here we focus on pseudo-labeling, the most elementary SSL method.

2.1. Pseudo-labeling

Pseudo-labeling [13] uses the prediction of an iteratively updated model for the classes of unlabeled data, which, combined with a small labeled dataset, can be used for (semi) supervised optimization. During the optimization, the loss function, which is minimized by gradient descent, is given by $\mathcal{L} = (1-\alpha)\mathcal{L}_L + \alpha\mathcal{L}_U$, where α is a weighting factor between labeled ($\mathcal{L}_L$) and unlabeled ($\mathcal{L}_U$) losses. The labeled data loss is defined as $\mathcal{L}_L = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C}\ell(y_i^m, f_i^m)$, with $\ell(\cdot)$ being the cross-entropy loss function. Here, C is the number of classes, f_i^m the network's probability output at the i-th class, $i \in \{1, ..., C\}$, for the m-th data sample, $m \in \{1, ..., n\}$, with one-hot encoded label y_i^m, for n labeled data examples. Similarly, for the unlabeled part of the data, the loss function is $\mathcal{L}_U = \frac{1}{n'} \sum_{m=1}^{n'} \sum_{i=1}^{C} \ell(y'^m_i, f'^m_i)$, where y' are the pseudo-labels and the apostrophe indicates that the variables are related to the unlabeled dataset.

The probability output of the neural network model is typically combined with a fixed probability threshold μ to determine the class of the unlabeled feature - the *pseudo-label*. In this sense, (one-hot) pseudo-labels for an unlabeled input feature x_m are chosen as

$$y'^m_i = \begin{cases} 1, & \text{if } i = \arg\max(\boldsymbol{f'}^m) \wedge f'^m_i > \mu, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $\boldsymbol{f'}^m$ is the softmax output of the network.

An alternative for a fixed value μ is to use a cosine-scheduled threshold as proposed in [11]. This allows starting at a lower confidence value that increases slowly, according to the number of epochs, thus gradually providing pseudo-labels with higher confidence. A more sophisticated method based on adaptation to the amount of data in the majority class has been proposed in [15], which is a state-of-the-art adaptive-threshold method and we briefly discuss it next.

2.2. Adaptive probability threshold

Imbalanced class distribution in the training data can cause strong performance differences from class to class in machine learning models [17]. In the case of SSL, the bias generated by the class imbalance will degrade the pseudo-label predictions by biasing them toward the majority class [15].

Adaptive (probability) thresholding (Adsh) [15] is a recently proposed method that selects pseudo-labels based on the majority class, where a class-wise threshold s_i is considered. A larger s_i means that more pseudo-labels are selected for class i. C vectors $\boldsymbol{P}_i$ with $i \in \{1, ..., C\}$, where C is the number of classes, are defined containing the probability of the pseudo-labels in descending order, with the classes ordered from majority ($i = 1$) to minority ($i = C$) class. The amount of pseudo-labels m in the majority class that satisfies $\boldsymbol{P}_1(m) > \tau_1$, for $m \in \{1, ..., length(\boldsymbol{P}_1)\}$ is defined as len. Here τ_1 is a hyper-parameter - the minimum desired confidence for the selected pseudo-labels of the majority class (set to 0.96 as in [15]). Once len is obtained, a ratio ρ can be defined as $\rho = len/length(\boldsymbol{P}_1)$, which is used for the adaptive confidence threshold $s_i = \boldsymbol{P}_i(length(\boldsymbol{P}_i) \cdot \rho)$, for $i \in \{1, ..., C\}$. Thus, (1) becomes

$$y'^m_i = \begin{cases} 1, & \text{if } \arg\max(\boldsymbol{f'}^m) = i \wedge f'^m_i > s_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

It is important to note that the Adsh method tries to adapt the per-class threshold with respect to the amount of data in the majority class that satisfies the confidence criteria. This is different from the methods proposed in this paper, described in Subsections 3.1 and 3.2, which treat the threshold optimization for each class separately.

2.3. Entropy-guided pseudo-labeling

The selection of pseudo-labels with a probability threshold only takes into account the softmax score of the class with highest probability. This method does not consider a confidence measure in the other classes. It has been shown that when the predicted probabilities for all classes are used, the classification performance can increase in certain scenarios [16]. For this reason, we also consider entropy-guided pseudo-labeling (ESL) in our investigation. The pseudo-labels are then obtained as

$$y'^m_i = \begin{cases} 1, & \text{if } \arg\max(\boldsymbol{f'}^m) = i \wedge E(\boldsymbol{f'}^m) < v_i, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

with v_i being the entropy threshold for class i. The entropy of the network's output probability $E(\boldsymbol{f'}^m)$ is defined as $E(\boldsymbol{f'}^m) = -\frac{1}{\log C} \sum_{i=1}^{C} f'^m_i \log f'^m_i$ [16]. The authors of [16] propose to use the entropy threshold as $v_i = \max(v*, \text{median}\{\boldsymbol{E}(\boldsymbol{f'})_i\})$, where $\boldsymbol{E}(\boldsymbol{f'})_i$ is the vector of entropies $E(\boldsymbol{f'}^m)$ for all features m assigned to class i and $v*$ is a hyper-parameter such that all samples with an entropy score lower than $v*$ will be selected.

Note that here the entropy threshold is not adapted per class. For this reason, in the current manuscript an adaptive thresholding rule, applied to both probability and entropy confidence criteria, is proposed to reduce the difference between the model accuracies for different classes.

3. PROPOSED ADAPTIVE CONFIDENCE SCORES

To showcase our proposed method from a more theoretical point of view, we first concentrate on two training classes (majority and another) from the imbalanced Ambient Acoustic Context (AAC) dataset [19]. Figure 1 shows the probability mass function (PMF) of the (top) maximum probability score and (bottom) entropy obtained at the output of the classifier. Many high probability/low entropy values for the majority class are observed, unlike for the other class. The data whose probability/entropy score is above/below a threshold is selected for model optimization by pseudo-labeling. For the entropy case, the selection process can be described by the cumulative distribution function (CDF) $F_X(x) = Pr[X \leq x] = \sum_{x_i \in X \leq x} Pr[X = x_i]$, where x is the threshold entropy. All data with a score below x is pseudo-labeled. By setting the threshold $x = t_{maj}$ for both classes, we have $F_X^{maj}(x = t_{maj}) > F_X^{other}(x = t_{maj})$. As a result, the number of pseudo-labels will be further biased toward the majority class. Therefore, to mitigate such an effect, our proposed pseudo-labeling process aims to ensure that t_{maj} and

t_{oth} for the different classes are chosen such that $F_X^{maj}(x = t_{maj}) \approx F_X^{other}(x = t_{oth})$. Using the complementary CDF, the same reasoning can be applied for the maximum probability scoring. As a consequence, our per-class adaptive thresholding might be beneficial when applied to the selection of pseudo-labels, as it avoids the introduction of further bias toward over-represented classes in the training data.

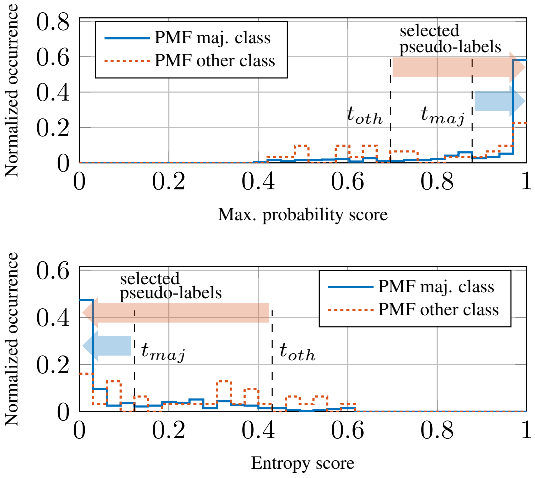

Fig. 1: Normalized occurrence of confidence scores for the majority and another class from the AAC dataset.

3.1. Entropy-based

Our entropy-based adaptive learning (EAD) uses individual credibility criteria for each class for determining pseudo-labels from unlabeled features. For this purpose, an additional optimization of the per-class threshold is performed, aimed at reducing the variance of the validation accuracies.

The pseudo-labels are obtained as in (3), but for this case, the threshold for class i at epoch k is defined as $v_i^k = \text{percentile}(\boldsymbol{E}(\boldsymbol{f}')_i^k, p_i^k)$, where p_i is the p-th percentile and $\boldsymbol{E}(\boldsymbol{f}')_i^k$ is ordered from the lowest to the highest value of entropy. The optimization of the percentile is key to the method and is given as follows. For each class i, the validation accuracy values λ_i are saved on an epoch basis, until a window size w is reached, as $\boldsymbol{\lambda}_i^k = [\lambda_i^1,\ \lambda_i^2,\ ...,\lambda_i^w]$. The average value of $\boldsymbol{\lambda}_i^k$, denoted $\bar{\boldsymbol{\lambda}}_i^k$, is compared to the average value $\bar{\boldsymbol{\lambda}}_i^{k+1}$ of the next window $\boldsymbol{\lambda}_i^{k+1} = [\lambda_i^2,\ \lambda_i^3,\ ...,\lambda_i^{w+1}]$. The percentile used to define the entropy threshold in v_i^k is initialized at 100% (the maximum value), i.e., taking the highest entropy value as the threshold. This results in obtaining pseudo-labels for all samples in (3). The percentile is updated per-epoch as

$$p_i^{k+1} = \begin{cases} p_i^k - \eta, & \text{if } \bar{\boldsymbol{\lambda}}_i^{k+1} > \bar{\boldsymbol{\lambda}}_i^{k}, \\ p_i^k + \eta, & \text{if } \bar{\boldsymbol{\lambda}}_i^{k+1} < \bar{\boldsymbol{\lambda}}_i^{k}, \\ p_i^k, & \text{otherwise,} \end{cases} \tag{4}$$

where η is the update step, which is a manually tuned hyperparameter. In (4), if the class accuracy increases, the percentile decreases, taking into account only the lower-entropy (higher-confidence) samples, and vice-versa. This reduces the bias toward the majority class since under-represented classes will be able to utilize more samples, while majority classes will have their amount of selected pseudo-labels reduced.

3.2. Probability-based

Alternatively to EAD, we also propose the probability-based adaptive learning (PAD) method. Here the pseudo-labels are chosen as in (2) and the threshold is defined as $s_i^k = \text{percentile}(\boldsymbol{M}_i^k, p_i^k)$, where $\boldsymbol{M}_i^k = max(\boldsymbol{f}')_i^k$ is a vector of maximum (softmax) probabilities of the network's output for input features assigned to class i at epoch k, ordered from the lowest to the highest value. The percentile update for PAD is

$$p_i^{k+1} = \begin{cases} p_i^k + \eta, & \text{if } \bar{\boldsymbol{\lambda}}_i^{k+1} > \bar{\boldsymbol{\lambda}}_i^{k}, \\ p_i^k - \eta, & \text{if } \bar{\boldsymbol{\lambda}}_i^{k+1} < \bar{\boldsymbol{\lambda}}_i^{k}, \\ p_i^k, & \text{otherwise,} \end{cases} \tag{5}$$

which is initialized at 0% (the minimum value), i.e., picking the lowest value of probability as threshold and, thus, selecting all pseudo-labels in (2). If the class accuracy increases in (5), the percentile will also increase, reducing the number of selected pseudo-labels for that class, and vice-versa, for the reduction of the bias toward the majority class.

4. NUMERICAL EXPERIMENTS

We consider two datasets of environmental sounds: Urban Sound 8K (US8K) [18], with the recommended 10-fold cross-validation, where three of the classes are under-represented; and AAC, where we use the same classes as in [11], which are also imbalanced. Since the US8K dataset has more samples per class, we considered a labeled partition of 1% of the total number of samples, while for the AAC, this was 3%.

For scenario 1 (US8K) and scenario 2 (AAC), we balance the labeled part of the data and leave the unlabeled part imbalanced. For an extreme case - scenario 3 (AAC) - we leave the labeled part imbalanced while forcing a strong imbalance on the unlabeled part of the data, done according to [15]: $n'_i = n'_1 \gamma^{-\frac{i-1}{C-1}}$, with imbalance ratio $\gamma = 200$, where n'_i is the number of unlabeled elements assigned to class i.

The model architecture is the same as in [11], a convolutional neural network consisting of four blocks, each with two separate convolutions - on temporal and frequency domains. At each layer, L2 regularization is applied with a rate of 10^{-4}. Max-pooling and spatial dropout rate of 0.1 are used between blocks to reduce the dimension size and avoid over-fitting. ReLU is used in the hidden layers. The model is optimized by the Adam algorithm, with a 10^{-3} learning rate. Cross-entropy is used as the loss function. The model input is a log-Mel spectrogram of the audio file with a window size of 2048 samples, a hop of 512, extracting 64 Mel-spaced frequency bins for each window.

Table 1: Average over 10 independent runs of: validation accuracy (Avg.); variance of validation accuracy (Var.); maximum (Max.) and minimum (Min.) accuracy (class-wise); and minimum F1 score per class (F1); for Scenarios 1, 2, and 3 (in %).

	Scenario 1				**Scenario 2**					**Scenario 3**				
	Avg.	**Var.**	**Max.**	**Min.**	**Avg.**	**Var.**	**Max.**	**Min.**	**F1**	**Avg.**	**Var.**	**Max.**	**Min.**	**F1**
Sup	45.46	4.48	87.48	29.30	44.38	2.71	60.00	14.86	—	39.21	3.96	76.42	16.67	—
F05	47.55	5.39	80.44	35.45	45.41	2.65	62.72	14.19	0.2160	37.75	4.90	84.21	11.67	0.0526
F06	47.76	5.32	82.45	31.36	44.93	2.87	62.03	10.95	0.2133	39.66	4.61	84.21	15.27	0.3226
F07	46.45	6.35	86.90	27.00	46.50	2.87	63.29	12.43	0.2352	37.99	4.87	83.46	13.33	0.1600
F08	48.47	4.98	86.85	27.30	45.94	3.19	64.75	11.08	0.1953	39.48	4.56	81.04	10.56	0.0909
F09	47.51	4.94	82.67	32.42	43.76	3.44	67.58	12.03	0.2215	38.87	4.57	83.42	13.65	0.3083
Cos	48.20	5.17	84.76	30.50	44.59	3.03	64.55	13.24	0.2364	38.83	4.76	81.79	13.61	0.3623
Adsh	46.85	5.44	84.28	25.10	45.85	2.80	67.11	12.70	0.2419	39.35	3.66	74.38	13.89	0.3000
ESL	47.33	5.10	83.81	32.60	45.60	2.54	64.30	16.22	0.2697	39.60	4.41	81.63	13.78	0.2857
PAD	47.02	4.13	80.81	33.30	45.10	2.11	59.70	16.49	**0.4384**	39.65	**3.16**	68.37	16.11	**0.4297**
EAD	46.60	**4.08**	82.10	**36.60**	42.35	**1.76**	56.06	**25.14**	0.3792	40.80	3.52	69.17	**20.56**	0.4170

4.1. Fully supervised learning with balanced data

To show that discrepancies between class accuracies can be present even when the data is balanced, we applied supervised learning for training the model described in Section 4 for 200 epochs averaged over 10 independent runs, with the US8K dataset considering only the full classes. Figure 2 shows the normalized occurrence and the validation accuracy per class. It can be seen that the accuracies differ, even though the number of samples is approximately the same.

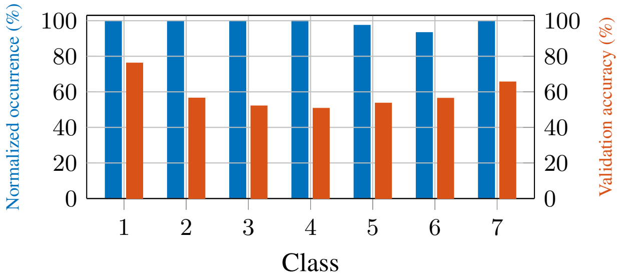

Fig. 2: Normalized occurrence (% over largest class) of the US8K dataset and validation accuracy for fully supervised training, averaged over 10 independent runs.

4.2. Semi-supervised learning

The results are obtained by applying the learning frameworks described in Section 2 and 3 - supervised (Sup), fixed threshold of 50-90% (F05-F09), cosine-scheduled threshold (Cos), adaptive thresholding (Adsh), entropy-guided learning (ESL), probability-based adaptive learning (PAD) and entropy-based adaptive learning (EAD) - for the scenarios described earlier in this section. 200 epochs of training are performed for all cases, where the first 50 are only supervised training.

Our tests showed that η, from (4) and (5), should be monotonically lowered as the training progresses, initially allowing for better exploration on the percentile and later acting as a finer adjustment. In particular, η is chosen to decrease linearly from 20.0 to 1.0 in each epoch from 51 to 150, and from 1.0 to 0.1 in each epoch from 151 to 200. Also, w can affect the validation accuracy variance and convergence. In this case, values of w around 5 and 20 got higher variance, while intermediate values around 10-15 achieved the lowest. Since the window size increases the computation time for the updates, $w = 10$ was chosen.

Table 1 presents the validation accuracy (averaged over all classes), the variance of the per-class accuracy values, the maximum and minimum per-class accuracy values, and the minimum per-class F1 score, all considering the average value over 10 independent runs for each of the described scenarios. When the proposed techniques, EAD and PAD, are compared to the SSL state-of-the-art methods, Cos, Adsh, and ESL, a lower variance can be clearly seen in all cases, which indicates a more balanced validation accuracy. For EAD, the minority class validation accuracy is always increased if compared to other methods, especially for scenarios 2 and 3 where the improvement is substantial. The minimum F1 score obtained with EAD and PAD was substantially higher than other methods for both imbalanced scenarios, indicating that misclassification is more uniform among different classes, while maintaining the overall accuracy.

Note that the majority class accuracy is reduced when the proposed methods are applied, which is an effect stemming from the forced balance in the per-class accuracies, suggesting that the bias on the majority class is reduced. The average accuracy for PAD is increased in all scenarios in comparison to the baseline (supervised). On the other hand, EAD achieves a higher average accuracy for scenarios 1 and 3, and lower for scenario 2, however with a significantly lower variance.

5. CONCLUSION

We proposed a framework for the per-class adaptation of confidence thresholds for pseudo-labeling. This results in two semi-supervised learning methods, probability-based adaptive learning (PAD) and entropy-based adaptive learning (EAD), with the specific objective of achieving a balance between the validation accuracy of different classes. Our numerical experiments show that the proposed methods are able to reduce the variance of the validation accuracy between classes when compared to supervised learning or state-of-the-art pseudo-labeling methods.

6. REFERENCES

[1] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[2] Thomas Schmitz and Jean-Jacques Embrechts, "Nonlinear real-time emulation of a tube amplifier with a long short time memory neural-network," in *144th Audio Engineering Society Convention*, May 2018.

[3] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[4] Asger Heidemann Andersen, Sébastien Santurette, Michael Syskind Pedersen, Emina Alickovic, Lorenz Fiedler, Jesper Jensen, and Thomas Behrens, "Creating clarity in noisy environments by using deep learning in hearing aids," *Semin Hear*, Aug 2021.

[5] Kun Han, Yuxuan Wang, DeLiang Wang, William S. Woods, Ivo Merks, and Tao Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[6] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florencio, and Mark Hasegawa-Johnson, "Deep learning based speech beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5389–5393.

[7] Juncheng Li, Wei Dai, Florian Metze, Shuhui Qu, and Samarjit Das, "A comparison of deep learning methods for environmental sound detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 126–130.

[8] Po-Jung Ting, Shang-Jang Ruan, and Lieber Po-Hung Li, "Environmental noise classification with inception-dense blocks for hearing aids," *Sensors (Basel)*, vol. 21(16):5406, 2021.

[9] Anusha Yellamsetty, Erol J. Ozmeral, Robert A. Budinsky, and David A. Eddins, "A comparison of environment classification among premium hearing instruments," *Trends in Hearing*, vol. 25, 2021.

[10] Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Björn Schuller, Xiaojie Yu, and Xuan Zhu, "Semi-supervised active learning for sound classification in hybrid learning environments," *PLOS ONE*, vol. 11, no. 9, pp. 1–23, 09 2016.

[11] Vasileios Tsouvalas, Aaqib Saeed, and Tanir Ozcelebi, "Federated self-training for data-efficient audio recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 476–480.

[12] Jesper E. van Engelen and Holger H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, Feb 2020.

[13] Dong-Hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML 2013 Workshop: Challenges in Representation Learning*, 2013.

[14] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 1195–1204, Curran Associates Inc.

[15] Lan-Zhe Guo and Yu-Feng Li, "Class-imbalanced semi-supervised learning with adaptive thresholding," in *Proceedings of the 39th International Conference on Machine Learning*. Jul 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 8082–8094, PMLR.

[16] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez, "ESL: Entropy-guided self-supervised learning for domain adaptation in semantic segmentation," in *CVPR'20 Workshop on Scalability in Autonomous Driving*, 06 2020.

[17] Qi Dong, Shaogang Gong, and Xiatian Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 06, pp. 1367–1381, jun 2019.

[18] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, New York, NY, USA, 2014, p. 1041–1044, Association for Computing Machinery.

[19] Chunjong Park, Chulhong Min, Sourav Bhattacharya, and Fahim Kawsar, "Augmenting conversational agents with ambient acoustic contexts," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, New York, NY, USA, 2020, MobileHCI '20, Association for Computing Machinery.