Techniques Improving the Robustness of Deep Learning Models for Industrial Sound Analysis

David S. Johnson
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany
jsn@idmt.fraunhofer.de

Sascha Grollmisch
Institute for Media Technology
TU Ilmenau
Ilmenau, Germany
sascha.grollmisch@tu-ilmenau.de

***Abstract*—The field of Industrial Sound Analysis (ISA) aims to automatically identify faults in production machinery or manufactured goods by analyzing audio signals. Publications in this field have shown that the surface condition of metal balls and different types of bulk materials (screws, nuts, etc.) sliding down a tube can be classified with a high accuracy using audio signals and deep neural networks. However, these systems suffer from domain shift, or dataset bias, due to minor changes in the recording setup which may easily happen in real-world production lines. This paper aims at finding methods to increase robustness of existing detection systems to domain shift, ideally without the need to record new data or retrain the models. Through five experiments, we implement a convolutional neural network (CNN) for two publicly available ISA datasets and evaluate transfer learning, data normalization and data augmentation as approaches to deal with domain shift. Our results show that while supervised methods with additional labeled data are the best approach, an unsupervised method that implements data augmentation with adaptive normalization is able to improve the performance by a large margin without the need of retraining neural networks.**

***Index Terms*—industrial sound analysis, neural networks, data augmentation, data normalization, transfer learning**

I. Introduction

Manufacturing organizations are making efforts to automate quality control in production lines to reduce human error and the need for specialized worker knowledge. Therefore, production equipment is augmented with sensors to monitor production lines and their output. Not all elements of machinery being monitored can be visually inspected, and require auditory techniques for analysis. While experienced machine operators are often able to hear changes in machinery state indicating the need for maintenance, they cannot constantly monitor the production line while working on their actual task. Hence, there is a the need for more automated approaches. The field of Industrial Sound Analysis (ISA) aims at addressing these issues by automatically analyzing airborne sound recordings in industrial settings using machine learning techniques.

While feed forward neural networks (DNN) have shown success in predictive maintenance scenarios, they are prone to overfit the training data [1], and thus, are not robust to minor changes in setup, such as recording equipment and location. Such environmental changes modify the distribution parameters of the data in a phenomenon known as *domain shift*, *covariate shift*, or *dataset bias* [2]. In an ISA scenario, predictive models need to be robust to domain shift because there is limited control over the conditions of unpredictable acoustic environments, like a production line, or the models may need to be redeployed in new locations with different recording setups.

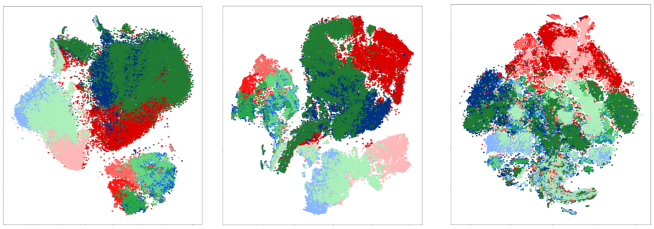
In this work, we contribute a structured evaluation of transfer learning (TL), normalization, and data augmentation techniques to develop robust models for two publicly available ISA datasets[1,2]. The baseline system using DNNs [1] is extended to convolutional neural networks (CNN) to test their general potential for ISA, as well as robustness to domain shift. Furthermore, we propose an unsupervised normalization method in combination with data augmentation that outperformed the state-of-the-art on both domain shifted datasets by a large margin.

To foster research in ISA, Grollmisch et al. [1] published new datasets simulating different ISA related tasks. Results from their baseline detection models showed that DNNs are able to achieve high classification accuracy with controlled environmental configurations. For example, their metal ball surface detection model achieved classification accuracy of 98.9% when the model is trained and tested with data from the exact same recording setup. With small perturbations to the recording configuration of the test data, accuracy drops to 51.4% and 59.2% for two variation sets. Visualizing the datasets without normalization using 2D t-SNE embeddings [3], Fig. 1a, shows that the two variation sets have different distributions from the core data, with the variation sets (lighter shades) distinctly clustered from the core data (darkest shade). Using standard pre-processing normalization (zero mean and unit variance) calculated on the training set and applied to the variation sets does not align the distributions, as seen in 1b. The same behaviour was observed for bulk material detection. The DNN classifier achieved perfect accuracy for distinguishing 10 different bulk materials such as nuts and

This work has been partly supported by the German Research Foundation (BR 1333/20-1, CA 2096/1-1)

[1]Metal ball surface dataset: https://www.idmt.fraunhofer.de/en/publications/isa-metal-balls.html

[2]Bulk material detection dataset: https://www.idmt.fraunhofer.de/en/publications/isa-tubes.html



(a) No Norm (b) Global Norm (c) Adaptive Norm

Fig. 1: t-SNE embeddings of IIMB spectral features with no (1a), *global* (1b), and *adaptive* (1c) normalization methods applied. Each color represents a metal ball surface class. Data from different recording configurations are represented through different color shades.

screws. Applying the trained model on a dataset recorded with a replica of the setup results in the performance of the model dropping to 51.6%. The performance degradation exemplifies the risk of overfitting and domain shift which are common challenges in deep learning. Techniques such as l1 or l2 regularization and dropout [4] have become commonplace over the years to reduce overfitting. However, they were already applied with the baseline DNN classifiers, indicating that they are not solutions for mitigating the effects of domain shift.

II. Related Work

One possible, but expensive, method to overcome domain shift is to record and annotate new data from the modified setup. To avoid the necessity of collecting large amounts of data, an existing model can be fine-tuned on a small number of new data observations, using Transfer Learning (TL) [5]. TL has been successfully applied to research fields such as image classification [6], and music information retrieval (MIR), with tasks such as instrument recognition [7] and genre classification [8]. To the best of our knowledge, this technique has not been investigated in the field of ISA.

Another possible technique to reduce overfitting is data augmentation (DA). DA methods apply various deformations to original data samples to artificially create new data and enlarge the training data. In the field of Detection and Classification of Acoustic Scenes and Events (DCASE)[3], a common image processing augmentation method called Mixup [9] has had success in various DCASE tasks [10], [11]. Adapa [10] implements Mixup, image processing augmentations (random erasing, random rotate, grid distortion), and time domain augmentations (pitch shifting and time shifting) in their winning DCASE challenge submission for urban auto tagging. Mixup and grid distortion had highest performance gains of the techniques when applied separately, but using all techniques together performed the best. An augmentation technique to have success in DCASE acoustic scene classification (ASC) subtask B, in which test data is recorded with a different device than the training data, is SpecAugment [11], [12]. The goal of this work is to find a DA method that makes a model invariant to the nuances of small configuration changes in ISA tasks. Since it is not clear from previous research which augmentation method works best for ISA, we evaluate augmentation techniques used by [10] and [11].

An emerging technique from the field of computer vision for dealing with domain shift is called domain adaption [2]. The goal of domain adaption is to match the distribution of the learned feature representations to the target dataset by adapting the latent representations. While domain adaption has shown to be promising in ASC [13], [14], it is outside the scope of this article and left for future work.

[3]For further information visit: http://dcase.community/

III. Industrial Sound Datasets

For this article, we select two publicly available ISA datasets to use for the evaluation of methods to improve model robustness to domain shift: *IDMT_ISA_METAL_BALLS* (IIMB)[1] for metal ball surface classification and *IDMT_ISA_TUBES* (IIT)[2] for bulk material detection. Both datasets provide variation datasets and reported baselines to compare the influence of the methods evaluated in this work. For extended details on the datasets and their baseline models please refer to [1].

A. Metal Ball Surface Classification

The detection of damages to the surface of metal ball bearings is important to the efficient maintenance of production equipment. Acoustic quality control can be applied to this task since potential material damages are audible but the metal balls within the bearing cannot be easily accessed with other sensors. IIMB was created to acoustically detect the surface condition of a steel metal ball rolling down a slide made out of steel. The dataset has three surface classes, and consists of a training set, a separate test set, and two additional variation sets in which environmental settings, such as the angle of the slide and the position of microphone, were slightly modified. A common method to improve model performance is to extend the dataset. In order to do this, we created and published two additional variation datasets recorded in the same manner as the original data, but with new configuration changes to the system. Variation Set 3 (VS3) contains 624 new recordings with 208 from each class, and Variation Set 4 (VS4) contains 576 new recordings with 192 from each class.

B. Bulk Material Detection

The IIT dataset was created in order to analyze the potential of acoustically distinguishing between different kinds of bulk material during filling processes. The dataset contains audio recordings of various bulk materials rolling down a 3D printed plastic slide inside a tube, and consists of 10 different classes of bulk materials including five types of candies and five kinds of nuts and screws. The recordings were performed on two similarly manufactured tubes with minor variations due to the production process.

IV. Experimental Design

In Fig. 1a it is shown how IIMB data recorded under different configurations clearly suffers from domain shift. In

this work, we evaluate common methods to address this dataset bias through five sets of experiments. Experiments *E1* and *E2* establish upper baselines to address domain shift through supervised techniques with new labeled data. Since annotating new data is costly and should be avoided, the following experiments evaluate improving model robustness without retraining through normalization (*E3*), data augmentation (*E4*), and combined normalization and augmentation (*E5*).

In the first experiment (*E1*), we implement CNN for both datasets and evaluate three different normalization methods. Choosing the right architecture can be crucial for the performance of a CNN [15]. Therefore, we employ Bayesian Optimization (BO) [16][4] to find the optimal CNN hyperparameters. Since the same approach and data was used to generate the original DNN architecture [1], the results between DNN and CNN are comparable. This experiment evaluates the potential of employing the resulting CNN for ISA tasks, and their robustness to domain shift. For IIMB, the classification accuracy on the test set and the average accuracy of all four variation sets are reported. Additionally, we implement 4-fold leave-one-out (LOO) cross validation, in which models are trained using the core training data plus three of the four variations sets and evaluate with the excluded dataset. The reported result for LOO is the average accuracy over all four folds. For IIT, *Tube1* is used as the training data, with the results of 10-fold cross validation reported. Further, the model is evaluated by using one of the variation sets (*Tube1* or *Tube2*) as training and the other for testing. The results of the two variations are averaged for the presented results.

To overcome domain shift in ISA tasks, the best case would be that some new data can be obtained and labeled from the new setup. Then TL can be applied to fine-tune the initial prediction model. In experiment *E2* we evaluate this approach by fine-tuning trained models with the new variation data. The models were retrained with an increasing number of new examples per class (step-wise from 2 to 35) to determine the influence of the amount of new data on the final performance. These examples were excluded from the test data. Furthermore, we evaluate the effect of retraining either only the dense layers or the complete model. For IIMB and IIT, this is performed for each of the variation sets, and the average results are reported. Additionally, as a baseline we report the results of training models from scratch using the same number of samples per class.

The initial baseline DNN for the ISA datasets [1] normalized the training data to zero mean and standard deviation of one for each feature, and applied these normalization values to the test set (called *global* normalization for this article). Due to domain shift the mean and standard deviation of the training data no longer apply to the test set, as seen in figure 1b. For experiment *E3*, we evaluate normalization techniques to help overcome this. First, we propose a simple adjustment of the data by normalizing the test set independently of the training data (called *adaptive* normalization), which more closely aligns the datasets as seen in Fig. 1c. While this method requires some additional data from the new environment configuration, no annotations and retraining of neural networks are needed. In order to limit the need for new data at all, we propose normalizing each CNN input patch individually (called *patch* normalization), as used in [17]. Similar to *E1*, each normalization method is evaluated using the test and variation sets for IIMB and IIT, as well as LOO for IIMB.

In experiment *E4*, common augmentation methods discussed in Section II are evaluated for their potential to improve the robustness of models towards domain shift in ISA tasks. We apply methods in the time domain using pitch shifting (from -2 to 2 semitones) and time stretching (with a factor between 0.9 and 1.1) as well as in the frequency domain on the spectrograms using grid distortion, mixup, random brightness, random erasing, random rotate (limited to a max of 5 degrees in either direction), and SpecAugment (without time warping). Additionally, we perform augmentation randomly combining all methods. For each method, the training data is increased by a factor of four by applying the techniques three times to the original data. For the individual augmentations, the given augmentation method is applied three times, with random augmentation parameters selected each time. When augmenting with all methods, each augmentation method is selected to be applied to a sample with 50% probability to limit the amount of deformation applied to each image. Each augmentation method is evaluated on the variations sets and LOO for IIMB.

In experiment *E5* we augment the training data with all methods and use *adaptive* normalization. The models are evaluated using the variation sets for IIMB and IIT. Additionally, for IIMB we evaluate the combined methods using LOO.

V. Results

A. CNN Design and Evaluation (E1)

The resulting CNN architectures from BO are shown in Fig. 2 and results presented in Table I. For both networks, the input spectral patch is composed of 34 windows of 257 frequency bins with no overlap for a total of 0.4 seconds of audio. The CNN accuracy on the test data is similar to the DNN for

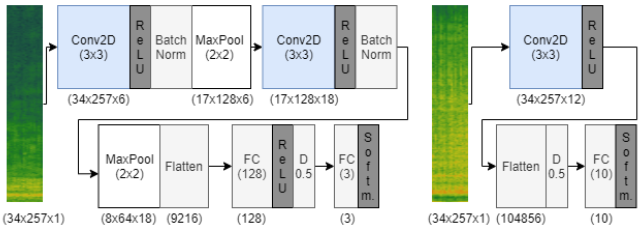


Fig. 2: CNN architectures for IIMB (left) and IIT (right) with input dimensionality, convolutional layers (Conv2D), ReLU activation (ReLU), Batch normalization (BatchNorm), max pooling (MaxPool), Dropout (D), Flatten, and fully connected layers (FC). The final dense layer uses softmax activation (Softm) for classification.

[4]Implemented with http://github.com/fmfn/BayesianOptimization

TABLE I: DNN & CNN classification accuracy (%).

Eval. Set	DNN IIMB	CNN IIMB	DNN IIT	CNN IIT
Test	98.8 ± 0.0	99.6 ± 0.3	100 ± 0.0	99.8 ± 0.1
Variation	58.7 ± 6.1	57.3 ± 7.2	51.6 ± 4.5	47.8 ± 4.0
LOO	71.0 ± 3.0	92.9 ± 2.9	-	-

both datasets, indicating that CNN are viable for ISA. The CNN, however, show limited change compared the DNN on all variations sets. Adding more data improves the robustness of the models, as seen in the results of LOO. Interestingly, the LOO results also show that the CNN is better than the DNN at generalizing to unseen data when the training data has more variability. The CNN is used in the rest of the experiments, as it demonstrates potential for dealing with robustness and making full use of additional variability in the training data.

B. Transfer Learning (E2)

The results of employing TL to fine-tune a CNN trained on the core training set using the variation data are plotted in Fig. 3. The results show that the performance of the TL models improves compared to models trained from scratch with only two new observations per class. For IIT, at 7 samples per class it is similar to train new models. The accuracy of the TL models for IIMB converges around 27 observations at around 98%. In general, it seems better to retrain only the dense layers since it also requires less resources. As expected, TL demonstrates to be a promising method for dealing with domain shift, but is expensive because of the need for additional labeled data. Interestingly, retraining IIT from scratch has similar results as TL with only 7 new observations per class.

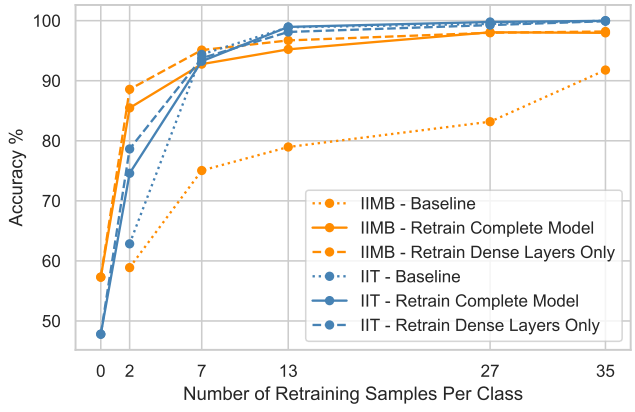


Fig. 3: Transfer learning accuracy (%) for different numbers of new training samples per class, including a baseline model trained from scratch.

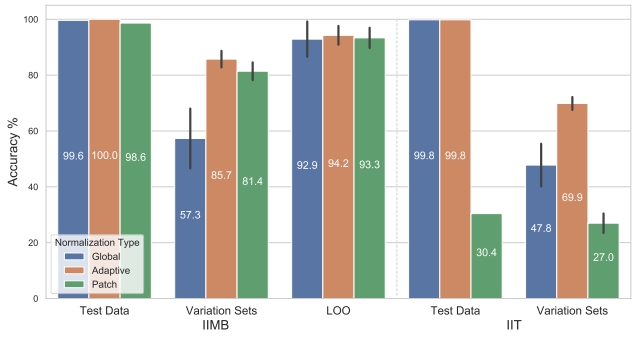


Fig. 4: Classification accuracy (%) for both metal ball (IIMB) and tube (IIT) employing each normalization method.

C. CNN Normalization (E3)

Fig. 4 shows the results of the models with *global*, *adaptive*, and *patch* normalization methods applied to the test and variation datasets. The results show that *global* normalization is problematic on the variation datasets for both IIMB and IIT, which was expected due to the domain shift shown in Fig. 1. Adapting the normalization using *adaptive* and *patch* helps on the IIMB variation sets. On IIT, however, *patch* has a major negative effect. Furthermore, as shown by the LOO results, models trained with a variety of data are less effected by the different normalization techniques. Methods that adapt normalization statistics, such as *adaptive* normalization, seem to be a promising approach to make models robust to environmental configuration changes. While some data is needed to recalculate the normalization values, the data does not need to be annotated making it a cheap alternative to TL.

D. Data Augmentation (E4)

Results of applying the different augmentation methods on the variation sets, and LOO, are listed in Table II. For the IIMB variation sets, all individual techniques show improvement over training without augmentation. On IIT, applying individual augmentation methods has less of an

TABLE II: Classification accuracy (%) of the variation sets for each of the data augmentation methods.

Augmentation	IIMB	IIMB LOO	IIT
No Augmentation	57.3 ± 7.2	92.9 ± 2.9	47.8 ± 4.0
Grid Distortion	59.5 ± 9.2	91.7 ± 1.4	51.2 ± 2.5
Random Brightness	63.0 ± 7.6	91.9 ± 2.0	49.3 ± 3.2
Random Erasing	67.5 ± 5.4	$\mathbf{94.7 \pm 1.9}$	50.8 ± 12.5
Random Rotate	60.0 ± 11.4	93.9 ± 1.5	44.3 ± 3.1
Mixup	64.2 ± 9.9	93.5 ± 2.2	44.2 ± 4.4
SpecAugment	66.5 ± 4.0	93.5 ± 2.4	46.0 ± 1.4
Time Stretch	59.6 ± 4.8	91.4 ± 2.2	53.4 ± 2.6
Pitch Shift	66.9 ± 6.4	87.1 ± 0.9	60.4 ± 4.3
All Methods	$\mathbf{70.5 \pm 7.5}$	92.9 ± 1.3	$\mathbf{66.3 \pm 6.1}$

TABLE III: Classification accuracy (%) of the test and variation sets comparing the effect of *global* and *adaptive* normalization with and without all augmentation.

Aug & Norm	Eval.	IIMB	IIMB LOO	IIT
No Aug & Global	Test	99.6 ± 0.3	-	$\mathbf{99.8 \pm 0.6}$
All Aug & Global	Test	97.9 ± 0.7	-	98.4 ± 1.6
No Aug & Adaptive	Test	$\mathbf{100.0 \pm 0.0}$	-	$\mathbf{99.8 \pm 0.6}$
All Aug & Adaptive	Test	98.8 ± 0.6	-	98.0 ± 2.5
No Aug & Global	Var.	57.3 ± 7.2	92.9 ± 2.9	47.8 ± 4.0
All Aug & Global	Var.	70.5 ± 7.5	92.9 ± 1.3	66.3 ± 6.1
No Aug & Adaptive	Var.	$\mathbf{85.7 \pm 2.4}$	$\mathbf{94.2 \pm 1.1}$	69.9 ± 2.8
All Aug & Adaptive	Var.	85.1 ± 1.4	94.1 ± 1.0	$\mathbf{86.5 \pm 1.2}$

effect, with three of the methods (random rotate, Mixup, and SpecAugment) resulting in decreased performance, although there is substantial improvement with pitch shifting. Applying all methods, however, improves the variation set results for IIMB and IIT by 13.2% and 18.5%, respectively, over the non-augmented baselines. This is surprising for IIT, because of the limited positive effects of the individual methods. The effects of augmentation are less prevalent for IIMB LOO, including the application of all methods, indicating that a dataset that already contains variability may not benefit as much from augmentation. Surprisingly, the intuition that a shift in frequency may lead to a different class is not holding true for the evaluated datasets, except when the training data already has variability.

E. Normalization and Augmentation Combined (E5)

The results of applying *adaptive* normalization with all augmentations, presented in Table III, demonstrate a minor decrease in the performance of the models on the test data but substantial improvement in the accuracy of the variation sets for IIMB and IIT. Improvement is less significant in the case of LOO. While augmentation has limited effect on IIMB and IIMB LOO when the data is normalized with *adaptive*, it greatly improves the performance for IIT. Since using this method slightly reduces accuracy on the test data, developers have to consider a trade-off between the accuracy on current conditions and the robustness of the model to domain shift with the current augmentation methods.

VI. Conclusions

In this article, we present five experiments to evaluate the effectiveness of different approaches for minimizing the effects of domain shift in ISA datasets due changing environmental conditions. CNNs achieved comparable performance to DNNs but displayed a higher potential for capturing additional variability in training data, making them a feasible choice for ISA classification tasks. From the results of our experiments, obtaining new labeled data to fine-tune existing CNNs seems to be the best, but also most expensive, method to overcome domain shift. Instead, there are other unsupervised approaches that developers can apply to improve a model's robustness to domain shift, including data specific normalization and data augmentation. By applying a *adaptive* normalization method and nine data augmentation methods, we were able to increase the performance of two ISA detection models for the IIMB and IIT datasets by 28% and 39% respectively. While individual augmentation methods only have a minor impact, the random combination of augmentations boosts the robustness of the trained model. However, the minor decline of performance on the initial data shows that the evaluated augmentation methods also introduce some unwanted artefacts. One method to integrate *adaptive* normalization in real-world production lines would be to use a running average. Currently, the entire test dataset is normalized at once. For future work, we suggest an evaluation of the number of examples required for a fast and robust adaption. Additionally, a thorough evaluation of domain adaption could be beneficial for ISA applications.

References

[1] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, "Sounding Industry: Challenges and Datasets for Industrial Sound Analysis," in *EUSIPCO*, A Coruña, Spain, 2019.

[2] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *IEEE CVPR*, July 2017.

[3] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, 2014.

[5] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[6] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in *IEEE CVPR*, 2014, pp. 1717–1724.

[7] J. S. Gómez, J. Abeßer, and E. Cano, "Jazz Solo Instrument Classification With Convolutional Neural Networks, Source Separation, and Transfer Learning," in *ISMIR*, Paris, France, 2018.

[8] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," *arXiv:1703.09179*, 2017.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *ICLR*, 2018.

[10] S. Adapa, "Urban sound tagging using convolutional neural networks," DCASE2019 Challenge, Tech. Rep., September 2019.

[11] M. Kosmider, "Calibrating Neural Networks for Secondary Recording Devices," Tech. Rep., 2019.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[13] S. Gharib, K. Drossos, E. Cakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *DCASE*, 2018, pp. 138–142.

[14] K. Drossos, P. Magron, and T. Virtanen, "Unsupervised adversarial domain adaptation based on the wasserstein distance for acoustic scene classification," in *IEEE WASPAA*, oct 2019, pp. 259–263.

[15] S. Grollmisch, E. Cano, F. Mora-Ángel, and G. López Gil, "Ensemble size classification in Colombian Andean string music recordings," in *CMMR*, Marseille, France, 2019.

[16] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *NeurIPS*, Lake Tahoe, Nevada, 2012, pp. 2951–2959.

[17] M. Taenzer, J. Abeßer, S. I. Mimilakis, C. Weiß, M. Müller, and H. Lukashevich, "Investigating CNN-Based Instrument Family Recognition for Western Classical Music Recordings," in *ISMIR*, Delft, The Netherlands, 2019.