Bridging the Gap: GANs as a Solution for Data-Scarce Industrial Audio Classification

Pitchapa Ngamthipwatthana[1], Saichand Gourishetti[1], Andrew McLeod[2], Sascha Grollmisch[1]
[1] *Industrial Media Applications, Fraunhofer IDMT;Ilmenau, Germany, E-mail: {first.last}@idmt.fraunhofer.de*
[2] *Semantic Music Technologies, Fraunhofer IDMT;Ilmenau, Germany, E-mail: {first.last}@idmt.fraunhofer.de*

Introduction

In machine listening, data scarcity is a significant challenge, particularly in Industrial Sound Analysis (ISA) [1], where openly available datasets are limited. Most existing datasets consists of laboratory recordings that mimic real world scenarios [1, 2, 3]. Acquiring accurately labeled data, essential, is resource-intensive, requiring substantial time and effort. This challenge is exacerbated in real-world applications, where suitable training data is often only available on-site.

As a possible solution, data augmentation can be applied to reduce data sparsity by adding small amounts of noise to the available data points [4]. Although data augmentation is effective in domains such as computer vision and speech recognition, these methods often fail in cases of extreme data scarcity, where the available data do not cover the full input space [5, 6].

In this work, we propose using Generative Adversarial Networks (GANs) [7] to generate synthetic training data. Specifically, we train a GAN on limited available data and use it to synthesize mel-spectrograms, significantly expanding the dataset. We then train a classifier on the combined dataset. Unlike traditional methods, our approach avoids synthesizing audio, as the GAN generates mel-spectrograms directly, and the classifier is trained on these spectrograms without the need for audio reconstruction.

We hypothesize that the proposed approach, akin to data augmentation, introduces realistic variations by learning possible alterations within each class.We evaluate our approach against a baseline, both with and without additional data augmentation, using two distinct ISA datasets. To simulate real-world data scarcity, we artificially limit the training data in our experiments. Our results highlight the conditions under which our method is effective and identify areas for further improvement.

Related Work

GANs [7] have shown remarkable success in generating high-fidelity data in various audio domains, e.g., speech synthesis [8], music generation [9], and sound effect production [10]. A GAN comprises two sub-networks—a generator and a discriminator—with opposing goals: the generator transforms a a random latent vector into realistic data, while the discriminator distinguishes between real and synthesized data. These sub-networks can employ diverse architectures, such as Convolutional or Transformer-based models.

WaveGAN [10] generates raw audio waveforms using a convolutional architecture with stacked 1D convolutions in both the generator and the discriminator. However, due to the length of the waveforms and the limited receptive field of convolutions, it is constrained to producing only 1-second clips. The same work introduced SpecGAN, which generates magnitude spectrograms using 2D convolutions. While SpecGAN produces realistic spectrograms, converting them to audio via the Griffin-Lim algorithm [11] and inverse STFT results in lower audio fidelity compared to WaveGAN.

GANSynth [9] addresses the limitations of WaveGAN and SpecGAN by generating high-fidelity mel-spectrograms for longer audio sequences. It adopts the Progressive GAN framework [12], gradually increasing resolution during training to produce high-quality outputs. The generator and discriminator use stacked 2D convolutional layers, with the generator incorporating a one-hot class label vector, inspired by Conditional GAN [13]. Additionally, a classifier is appended to the discriminator, leveraging Auxiliary Classifier GAN [14] to enhance training. This class-label mechanism enables the generation of diverse, class-specific outputs, making it suitable for our application.

Previous studies have explored GANs for generating training data in audio classification. Aswathy et al. [15] used WaveGAN for environmental sound classification, demonstrating performance improvements over standard augmentation methods. Similarly, Yang et al. [16] reported a 10% improvement in acoustic scene classification using WaveGAN-based augmentation in the DCASE 2018 Challenge Task 1A. However, both studies faced limitations, such as generating only 1-second audio clips at 16 kHz. To address this, Aswathy et al. [15] upsampled and extended the audio, while Yang et al. [16] stacked 2-second snippets to achieve 10-second segments. Later, Aswathy et al. [17] improved output length by incorporating transposed convolutions, enabling the generation of longer audio clips (around 4 seconds).

GANSynth offers advantages over WaveGAN, such as training a single GAN for an entire dataset using class-label vectors, unlike WaveGAN, which typically requires one GAN per class. Additionally, GANSynth generates mel-spectrograms directly, avoiding noisy audio conversions, as classifiers are trained on these spectrograms.

Text-to-audio methods have also been explored for data generation [18, 19], but their applicability to ISA is limited due to the domain-specific nature of industrial sounds, which are unlikely to be accurately captured by general models.

Datasets

In order to thoroughly evaluate our proposed method, we use two datasets with different signal characteristics. Basic features of each dataset (file duration, number of classes, etc.) can be found in Table 1. We do not report the size of each training set because we varied it in our experiments to evaluate different levels of data scarcity.

IDMT-ISA-METAL-BALLS (MB) [1]: This contains a collection of audio recordings of three differently-coated metal balls rolling down a metal slide. The classification is to detect the surface of the ball: eloxed, coated, or broken. The sounds in the MB dataset are continuous (rather than impulses).

IDMT-ISA-PUCKS (Pucks) [20] consists of one-minute audio recordings capturing impulse responses from plastic pucks 3D printed using four different materials. The classification task involves identifying the material of the puck in a recording, with an additional fifth class representing recordings without puck impulses. For training, we use files with quieter background noise (vol_050), while evaluation is performed on files with louder noise (vol_100). Unlike continuous signals, the Pucks dataset contains impulse sounds, testing our method's ability to synthesize such signals. To address memory constraints, we extract 5-second clips with 50% overlap from the 1-minute recordings, resulting in 23 clips per file for training.

Table 1: Dataset details.

Dataset	Classes	# Files (Test)	File duration	Sample rate
MB	3	171	0.4 s	44.1 kHz
Pucks	5	75	60 s	44.1 kHz

Proposed Method

Similar to real-world applications with limited data, we assume the availability of only a small amount of labeled real data for training. Our method uses this data to train a GAN capable of generating a much larger amount of labeled synthetic data. We then use the GAN to generate such data in the form of Mel-spectrograms, combining the synthesized spectrograms with those from the original real data. This combined dataset is finally used to train a classifier, with possible data augmentation applied to all data points (we investigate the effect of including data augmentation in our experiments). An overview of our method can be seen in Figure 1.

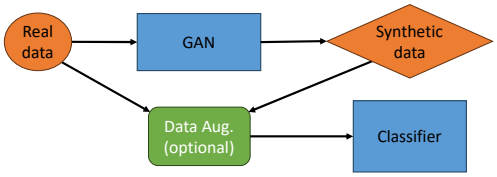

Figure 1: An overview of our proposed method.

GAN

Our method employs, GANSynth [9] to generate labeled spectrograms, treated as single-channel images. We retain most of the original hyperparameters but modify four key settings, detailed in Table 2. First, we reduce *Images per Stage* from 800k to 8k to shorten training time from a week to a few hours. However, this setting produced poor results for the Pucks dataset, so we increased it to 80k, yielding better outcomes.

The remaining hyperparameters relate to GANSynth's progressive resolution training. We use the original spectrogram settings: a Mel-spectrogram with an FFT size of 2048 samples and a step size of 512 samples, resulting in 1024 frequency bins. Audio files are adjusted to the nearest power-of-two length, determining the *Final Resolution*. The number of training stages (*Resolutions*) is set to one less than the base-2 logarithm of this resolution, with the *Start Resolution* being two frames.

Additionally, We modify GANSynth one-hot target vector, originally used for pitch control, to represent dataset class labels instead. This adaptation ensures the model generates class-specific spectrograms suitable for our application.

Table 2: GANSynth hyperparameters used for each dataset.

GANSynth parameter	MB	Pucks
Images per Stage	8,000	80,000
Final Resolution	(32, 1024)	(512, 1024)
Resolutions	5	9
Start Resolution	(2, 64)	(2, 4)

Classifier

We use the classifier that had previously shown state-of-the-art results on each dataset. In all cases, that was a Convolutional Neural Network (CNN)-based classifier, with a single feedforward layer on top with softmax activation and dimensionality equal to the number of classes in the dataset. Categorical cross-entropy loss is used during training, and all classifiers were trained with the Adam optimizer [21] with a learning rate of 0.001. MB used a batch size of 256 and trained for 70 epochs, while Pucks used 256 for 1,500.

For MB, we use the CNN420 ResNet architecture from [6]. We use average pooling after each ResNet block, while max global pooling is applied after the final convolutional layer. For Pucks, we use a 3-layer CNN with 3x3 convolutions on each layer as in [20]. During training, the classifier takes as input only 5-second clips: both extracted from the original files with 50% overlap and synthesized by the GAN. At inference time, when run on the real 1-minute recordings from the Pucks test set, we first cut each 1-minute input into 23 5-second clips (with 50% overlap) and run the classifier on each clip. Those 23 classifier outputs are then averaged in order to generate the final output for a given 1-minute audio file, as in [20].

Experiments and Results

Experimental Design

To systematically evaluate our method on each dataset, we imposed various levels of data scarcity by only using small randomly chosen subsets of each set for training, ranging from four to 360 real data files per class depending on the dataset. This setup is designed to evaluate

whether our method is able to compensate for the lack of available real data. We trained a GAN on each of these subsets, and used it to generate large amounts of corresponding synthetic data. Specifically, we generated 3,000 files per class for MB and 1,000 per class for Pucks[1]. Finally, the classifier was trained five times on this synthetic data plus that exact subset of real data (i.e., the same subset that was used to train the corresponding GAN). We report the mean and standard deviation of the classification accuracy on each dataset's pre-defined test set across these five trials[2]. As a baseline, we also trained each classifier five times on only that real data subset.

To evaluate the effectiveness of our method in comparison to, and in combination with, data augmentation, we also included a version of the baseline and our method both with and without data augmentation. For our method, this data augmentation was applied to both the real and the synthetic data. We used different combinations of augmentations and hyperparameters for each dataset, taken from the best-performing classifier on each dataset from previous work. Specifically, for the MB dataset, we applied the augmentation technique proposed in [5] (see Table 2 therein for a full list of the augmentations used) with a probability of 0.5 every time a data point was fed to the model for training. For the Pucks dataset, we applied mixup (with a probability of 1) as well as two image augmentations randomly chosen from a pre-defined set, each with a random magnitude and probability of 0.5. These data augmentation settings are identical to those used in the state-of-the-art classifier on the Pucks dataset [6] (see Appendix B therein for the full list of augmentations and associated parameters).

Results

Fig. 2 presents our method's classification accuracy on the MB dataset. The x-axis refers to the number of real audio files per class used to train the GAN, and likewise included when training the classifier. It is clear that the baseline classifier struggled to perform well with fewer than 108 training data points per class, but reached up to 90% and higher after that point. Data augmentation improved the baseline in all cases, with its benefit decreasing as more real data was added. Our proposed method outperformed the baseline (also with data augmentation) in cases of severe data scarcity ($\leq$ 18 files per class), while our method with data augmentation performed the best of all methods with $\leq$ 72 files per class. With more than that, all methods outperformed the baseline without augmentation to a similar degree. Overall, on MB, our method does indeed show a marked performance improvement over what is possible with just real data given a moderate or greater level of data scarcity.

The results on the Pucks dataset are shown in Fig. 3,

[1]We found that adding more data would generally improve performance, but there was a point of diminishing returns. For each dataset, we kept adding additional synthetic data until this point was reached.

[2]for MB, we repeated this entire process three times (i.e., we trained three GANs) and averaged the accuracy values across the resulting 15 trials. We found almost no difference between the three GANs and therefore only trained one GAN for Pucks dataset.

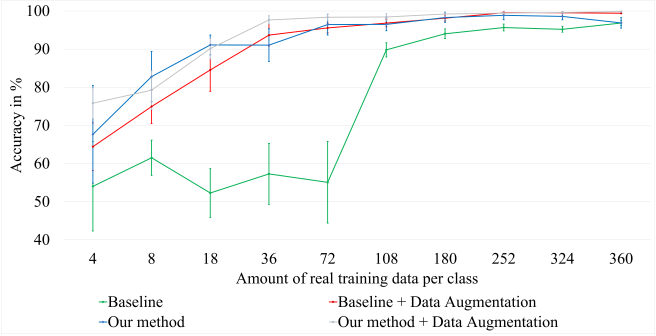

Figure 2: Mean classification accuracy on the MB dataset with varied levels of data scarcity. Vertical bars indicate standard deviation.

where the baseline outperformed all other methods. "10" training data points refers to the number of 5-second clips per class used for training, extracted from the original training files. These were extracted from one, two, or three original training files for the points 10, 20, and 30 on the x-axis, respectively. Here, adding data augmentation actually shows a trend towards decreasing performance, especially when combined with our method. The GAN struggled to produce realistic data for each class, and adding additional noise through the augmentations therefore only served to further exacerbate the issue.

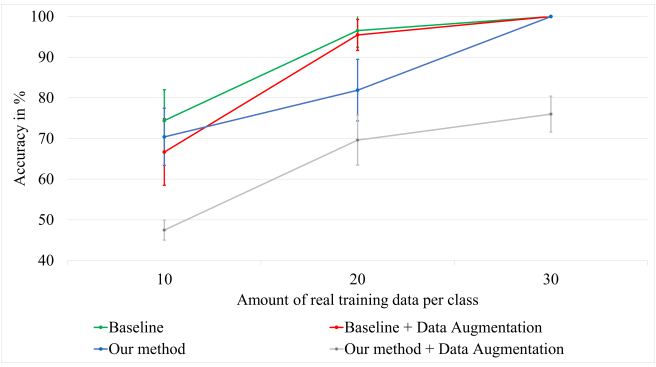

Figure 3: Mean classification accuracy on the Pucks dataset with varied levels of data scarcity. Vertical bars indicate standard deviation.

Conclusion

In this work, we proposed a method for improving classifiers for industrial audio data with limited training data using Generative Adverserial Networks (GANs). Such scarcity is often present in the Industrial Sound Analysis (ISA) domain due to the difficulty and expense of gathering and labeling large quantities of real-world data. Our method trains a GAN on a small amount of real data, and uses the GAN to generate much larger amounts of synthetic data for each class. We then combine this synthesized data with the original real data, and train a classifier on the resulting set. We hypothesize that this method functions as a "learned" form of data augmentation, where the GAN generates additional variety for each class.

We evaluated our method on two different datasets and compared it to a supervised baseline with and without

data augmentation: Metal Balls (MB), and Pucks. Our proposed method improved the results compared to the baseline on MB, especially when training data is scarce, showing its potential for audio classification. The results on Pucks were mixed. After a thorough investigation, we hypothesize that the time- and frequency-varying signal of Pucks is more difficult for the GAN to reproduce.

Nonetheless, our method's ability to achieve an increase in performance on the MB dataset shows that the approach has merit, and that future work is warranted. We would like to test our method's performance with a wider range of classifiers, including larger and smaller models for each dataset. We also intend to investigate ways to improve the GAN's performance, for example by starting from some pre-trained GAN that is already able to produce realistic audio. We would also like to see whether other synthesis methods besides a GAN, e.g., latent diffusion [22], might lead to better results.

References

[1] Grollmisch, S. et al.: "Sounding industry: Challenges and datasets for Industrial Sound Analysis," in the European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 2019, pp. 1–5.

[2] Purohit, H. et al.: "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), New York University, NY, USA, 2019, pp. 209–213.

[3] Johnson, D. et al.: "Compressed air leakage detection using acoustic emissions with neural networks," in the 49th International Congress and Exposition on Noise Control Engineering (Inter-Noise 2020), Online, 2020, pp. 5662–5673.

[4] Abayomi-Alli, O. O. et al.: "Data augmentation and deep learning methods in sound classification: A systematic review," Electronics, vol. 11, no. 22, 2022.

[5] Johnson, D. and Grollmisch, S.: "Techniques improving the robustness of deep learning models for Industrial Sound Analysis," in the 2020 European Signal Processing Conference (EUSIPCO), Online, 2021, pp. 81–85.

[6] Grollmisch, S. and Cano, E.: "Improving Semi-Supervised Learning for audio classification with FixMatch," Electronics, vol. 10, no. 15, 2021.

[7] Goodfellow, I. J. et al.: "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.

[8] Pascual, S. et al.: "SEGAN: Speech Enhancement Generative Adversarial Network," in the Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden, 2017, pp. 3642–3646.

[9] Engel, J. et al.: "GANSynth: Adversarial Neural Audio Synthesis," in the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 2019.

[10] Donahue, C. et al.: "Adversarial Audio Synthesis," in the International Conference on Learning Representations (ICLR), Online, 2018.

[11] Griffin, D. and Lim, J.: "Signal estimation from modified short-time Fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, 1984.

[12] Karras, T. et al.: "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in the International Conference on Learning Representations (ICLR), Online, 2018.

[13] Mirza, M. and Osindero, S.: "Conditional Generative Adversarial Nets," arXiv preprint arXiv:1411.1784, 2014.

[14] Odena, A. et al.: "Conditional Image Synthesis with Auxiliary Classifier GANs," in the International Conference on Machine Learning (ICML), Sydney, NSW, Australia, 2017, vol. 70, pp. 2642–2651.

[15] Madhu, A. and Kumaraswamy, S.: "Data augmentation using generative adversarial network for environmental sound classification," in 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5.

[16] Yang, J. H. et al.: "Se-resnet with gan-based data augmentation applied to acoustic scene classification technical report," Tech. Rep., DCASE2018 Challenge, 2018.

[17] Madhu, A. and K., S.: "Envgan: a gan-based augmentation to improve environmental sound classification," Artif. Intell. Rev., vol. 55, no. 8, pp. 6301–6320, dec 2022.

[18] Feng, T. et al.: "Can synthetic audio from generative foundation models assist audio recognition and speech modeling?," arXiv preprint arXiv:2406.08800, 2024.

[19] Ronchini, F. et al.: "Synthetic training set generation using text-to-audio models for environmental sound classification," arXiv preprint arXiv:2403.17864v3 [eess.AS], 2024.

[20] Grollmisch, S. et al.: "Plastic material classification using neural network based audio signal analysis," in Sensor and Measurement Science International (SMSI), Online, 2020, pp. 337–338.

[21] Kingma, D. P. and Ba, J.: "Adam: A method for stochastic optimization," in the International Conference on Learning Representations (ICLR), San Diego, USA, 2015.

[22] Rombach, R. et al.: "High-Resolution Image Synthesis with Latent Diffusion Models," in the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10674–10685.