


Anomaly Detection in Sound Activity with Generative Adversarial Network Models

Wilson A. de Oliveira Neto   [Amazonas Federal University | wilson.oliveira@icomp.ufam.edu.br]

Elloá B. Guedes  [Amazonas State University | ebgcosta@uea.edu.br]

Carlos Maurício S. Figueiredo  [Amazonas State University | cfigueiredo@uea.edu.br]

 Research Group on Intelligent Systems, Amazonas State University. Av. Darcy Vargas, 1200, Manaus, AM, Brazil.

Received: 03 December 2023 • Accepted: 28 June 2024 • Published: 05 September 2024

Abstract In state-of-art anomaly detection research, prevailing methodologies predominantly employ Generative Adversarial Networks and Autoencoders for image-based applications. Despite the efficacy demonstrated in the visual domain, there remains a notable dearth of studies showcasing the application of these architectures in anomaly detection within the sound domain. This paper introduces tailored adaptations of cutting-edge architectures for anomaly detection in audio and conducts a comprehensive comparative analysis to substantiate the viability of this novel approach. The evaluation is performed on the DCASE 2020 dataset, encompassing over 180 h of industrial machinery sound recordings. Our results indicate superior anomaly classification, with an average Area Under the Curve (AUC) of 88.16 % and partial AUC of 78.05 %, surpassing the performance of established baselines. This study not only extends the applicability of advanced architectures to the audio domain but also establishes their effectiveness in the challenging context of industrial sound anomaly detection.

Keywords: Anomaly Detection, Sound Activity, Generative Adversarial Networks, Deep Learning.

© Published under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0)

1 Introduction

The objective of anomaly detection is to discern and discriminate anomalous samples from those that are representative of typical behavior [Chandola *et al.*, 2009]. The primary purpose of anomaly detection is to differentiate between the expected and unexpected behaviour of a system, enhancing robustness and reliability [Sabuhi *et al.*, 2021]. In contrast to problems involving majority or regular patterns, anomaly detection focuses on minority, unpredictable, and rare events, introducing unique complexities, such as: (i) unknownness, such as instances exhibiting abrupt behaviors, unfamiliar data structures, and unpredictable distributions; (ii) heterogeneous anomaly classes, resulting in the distinct possibility that one class may manifest entirely different abnormal characteristics compared to another class; and (iii) rarity and class imbalance, for being difficult, if not impossible, to collect a large amount of labeled abnormal instances [Pang *et al.*, 2021].

Anomaly detection holds paramount significance in the domains of intelligent environments and Industry 4.0 applications. Noteworthy applications include, but are not limited to, the identification of faults in industrial machinery [Purohit *et al.*, 2019], detection of road accidents [Rovetta *et al.*, 2020], multimodal approaches using video analysis [Kittler *et al.*, 2018], and the identification of faults in 5G signal transmission [Zhou *et al.*, 2021].

Within the realm of sound activity anomaly detection, autoencoders have emerged as a compelling approach [Zhou and Paffenroth, 2017; An and Cho, 2015]. Autoencoders

(AEs), characterized as neural networks, exhibit the capacity to acquire a condensed and high-fidelity representation of input data [Schmidhuber, 2015]. This capability stems from their inherent design, wherein they are explicitly tailored to reconstruct the original input while minimizing information loss within the compressed representation. Training an AE on a dataset comprising normal sound instances enables it to adeptly reconstruct these customary patterns [Xu *et al.*, 2021]. Consequently, anomalies are difficult to be reconstructed from the resulting representations and thus have large reconstruction errors. In this way, unknown patterns in audio should generate reconstruction errors, which will be used to identify and quantify the anomaly.

Proposed by Goodfellow *et al.* [2014], Generative Adversarial Networks (GANs) represent a model architecture comprising two neural networks: the discriminator D and the generator G . The latter is tasked with discerning authentic data from synthetic data, while the former learns to produce artificial data that closely approximates authentic data, thereby thwarting the discriminator's ability to distinguish it as counterfeit [Langr and Bok, 2019]. Although initially employed for content generation, predominantly in the realm of images, the adversarial training paradigm has facilitated the extension of GANs into diverse applications, notably anomaly detection [Sabuhi *et al.*, 2021]. The fundamental concept revolves around the assumption that normal data instances can be better generated than anomalies from the latent feature space of the generative network in GANs [Pang *et al.*, 2021]. Thus, discriminators trained to differentiate between real and fake entities can be repurposed to discern the

typicality of data, identifying anomalies in the latter scenario.

As per the literature, GANs have exhibited exceptional proficiency in generating realistic instances, enabling the identification of abnormal instances that exhibit suboptimal reconstruction from the latent space. Furthermore, numerous GAN-based models and theories can be repurposed for anomaly detection [Sabuhi *et al.*, 2021]. Nevertheless, the majority of existing research in this domain has been developed and assessed primarily for images [Schlegl *et al.*, 2019; Liu *et al.*, 2021; Akcay *et al.*, 2019a,b]. These researchers have conjectured the potential use of such models for audio data, a point that has been remarked upon by some of them. However, these studies do not detail the process of adapting the proposed models to this new media type, nor do they provide reference evaluation metrics for this specific case.

In this context, we highlight the importance of exploring methodologies to customize GAN models for audio anomaly detection, aiming to assess the feasibility of this approach in the specified application domain. To harness the insights into the task under consideration, this study presents experimental results derived from adapting prominent anomaly detection models from the literature. The comparison is conducted using well-established challenge datasets, including *DCASE 2020* [Tanabe *et al.*, 2021; Kawaguchi *et al.*, 2021], *Urban Sound* [Salamon *et al.*, 2014], and *AudioSet* [Gemmeke *et al.*, 2017]. This is an extension of previous work [Neto and Figueiredo, 2023] with improved descriptions, better results from model customization for each audio class, and discussions. The main contributions of this work are:

1. Establishment of a methodology for the adaptation of GAN models to the audio domain;
2. Standardization of architectures of GAN models for the assessment of Area Under the Curve (AUC) of the Receiver Operating Characteristic and partial-AUC (pAUC) metrics;
3. Results of the modified architectures are presented and also introduced as a baseline for future research and comparisons in the audio domain;
4. Performance benchmarks for these adapted GAN architectures are provided using the best hyperparameters.

The subsequent sections of this paper are structured as follows: Section 2 provides an overview of the related work. Section 3 delineates the proposed approach, while Section 4 outlines the experimental results. Lastly, Section 5 elucidates the conclusions and provide directions for future research.

2 Related Work

The task of anomaly detection can be approached through various learning paradigms. In the realm of supervised learning, the assumption is made that a training dataset is available, comprising labeled instances for both normal and anomaly classes. This scenario poses two major challenges: (i) the scarcity of anomalous instances compared to normal instances in the training data, and (ii) the difficulty in obtaining accurate and representative labels, particularly for the anomaly class [Chandola *et al.*, 2009]. In unsupervised

learning anomaly detection, the main challenge of this task is to detect unknown anomalous instances under the condition that only typical samples have been provided as training data [Koizumi *et al.*, 2020b].

Audio anomaly detection is typically approached through the latter strategy, where training data includes normal instances, while anomaly samples are unavailable. This is reflective of real-world scenarios where anomalous sounds are infrequent and highly diverse. Consequently, creating or collecting comprehensive patterns of anomalous sounds intentionally is impractical or impossible [Koizumi *et al.*, 2020b].

Towards learning directly from an embedding layer, Wilkinghoff [2023] proposed an Anomalous Sound Detection (ASD) system. Utilizing two distinct feature representations from raw waveforms, it incorporates magnitude spectrograms and spectra of entire signals. The neural network consists of two sub-networks, each specialized for different representations and trained to discriminate machine types, sections, and attributes. The design choices prevent learning trivial mappings, ensuring sensitivity to anomalies. Concatenating sub-network outputs yields single embeddings for each file. Anomaly scoring strategies in source and target domains utilize k-means clustering and cosine distance calculations. Implemented with mixup data augmentation. However, the processing of 2D spectrograms entails significant computational overhead due to the large input dimensionality and convolutional operations.

In the same context, Chen *et al.* [2024] proposed an enhancement of the previous work, utilizing the same neural network as a backbone. They designed and added an attention module called the Multi-Dimensional Attention Module (MDAM), which focuses on specific frequency bands to retrieve semantic information. This module infers attention along three independent dimensions: time, frequency, and channel. However, like the previous work, this approach also suffers from the limitation of using a 2D spectrogram as input.

Recent advancements in this field have notably enhanced the efficacy of generative models for this task. Typically, this process entails training a deep AE model to reconstruct input data and then utilizing the reconstruction error for anomaly detection [Cheng *et al.*, 2021]. The works of Koizumi *et al.* [2020a] and Suefusa *et al.* [2020] fall into this category, using regular AE architecture, and the work of Müller *et al.* [2021] uses memory cells. To amplify the differentiation between the reconstruction errors of authentic data and anomalies, trained discriminators from GANs were integrated into the architectures of AEs [Schlegl *et al.*, 2019]. The aforementioned methodology has demonstrated success in the domain of image anomaly detection, and some works are better described as follows.

The Efficient GAN-Based Anomaly Detection (EGBAD) proposed by Zenati *et al.* [2018] is an example of GAN-based model in which the AE network is utilized to acquire the latent representation of the data. Diverging from conventional GANs, the Discriminator in EGBAD also considers the latent representation in addition to the image. The underlying hypothesis posits that anomalous data exhibit latent vectors that lack correlation with the corresponding images. During the computation of anomaly scores, the author employs a com-

posite metric involving Discriminator classification and the distance between the generated and real images.

GANomaly [Akçay *et al.*, 2019a] represents another instance of a combined AE and GAN architecture. A notable aspect is the design of the generator G network, incorporating a blend of AE networks with two outputs: the generated image and a corresponding latent representation. The authors illustrate that the latent space of the generated image exhibits significant distinctions from the latent space of the original image when the original image deviates from normal data. SKIP-GANomaly [Akçay *et al.*, 2019b] emerges as an advancement over the prior related efforts. Its generator network adopts a U-NET architecture [Ronneberger *et al.*, 2015], enabling the recreation of input images with enhanced detail compared to a conventional AE network, courtesy of residual connections within the model layers. This approach underscores the benefits of adversarial training in contrast to solely training a standard AE network.

Table 1 compiles a summary of these noteworthy contributions in the anomaly detection domain. It becomes apparent that the prevailing trend involves the application of generative architectures primarily in the context of images, while AE architectures find predominant usage in audio applications.

In reviewing the body of literature related to the subject matter, some observations emerge:

1. The use of GANs for anomaly detection in acoustic events and scenes has been explored less comprehensively compared to the advancements achieved in anomaly detection in images;
2. Although studies in this area are still limited, existing research suggests that the application of GANs in the acoustic domain may also hold promise;
3. The combination of encoder and generative architectures has proven effective in creating latent representations and generating realistic audio data;
4. A greater research effort is needed to explore and adapt GANs to the specific nuances of acoustic events and scenes, aiming to achieve more accurate and reliable results in anomaly detection within this context;
5. It is important to emphasize that the application of GANs for anomaly detection requires careful consideration and specific practices when training models in different domains, such as images and audio. While training techniques for images are widely studied and established, adapting these approaches to the audio domain presents additional challenges. One of the key considerations is the appropriate selection of loss functions, taking into account the characteristics and nature of audio signals;
6. It is necessary to adjust the hyperparameters and architectures of GANs to effectively capture the nuances of the acoustic domain, such as temporal representation and spectral complexity.

A previous work Neto and Figueiredo [2023] showed that this approach is feasible in audio domain by presenting competitive results when compared to literature. This work extends that research by increasing experiments and parameter optimization to achieve better performance metrics.

3 Proposed Solution

The proposed solution within the scope of this study involves the adaptation of GAN-based models from the literature [Zenati *et al.*, 2018; Akçay *et al.*, 2019a,b] for sound anomaly detection. While there are established results regarding their performance in the domain of images, a gap exists in understanding their effectiveness in the audio domain, as well as the requisite adaptations for this specific data type. The following stages are comprised in the proposed solution:

1. **Pre-processing.** It involves standardizing audio instances into feature vectors. This not only enables GANs to process this type of information but also facilitates the utilization of the same datasets across various GAN architectures. This step also encompasses the partitioning of training and testing examples;
2. **Architectures Adaptation and Training.** In this phase, the architectures are initially instantiated for the type of input data and trained solely with examples of typical audio behavior. We will consider the approach of creating a model for each class of each considered architecture. During training, the generators G faithfully reproduce input sets, while the discriminator networks D assess the similarity of the generated data to the originals. As a result of each unsupervised training, two neural models are obtained: one for generating data close to typical instances and another for classifying them;
3. **Validation and Evaluation.** The objective is to assess the efficacy and effectiveness of sound anomaly detection GAN-based models. To achieve this, each generator model G takes the audio inputs from the test partition and produces an output. The distance between the input and its generated representation is then calculated. Considering the entire test set, the average errors will be measured to derive metrics such as the Area Under the Curve (AUC) and partial AUC (pAUC).

The subsequent sections will illustrate each phase of the proposed solution.

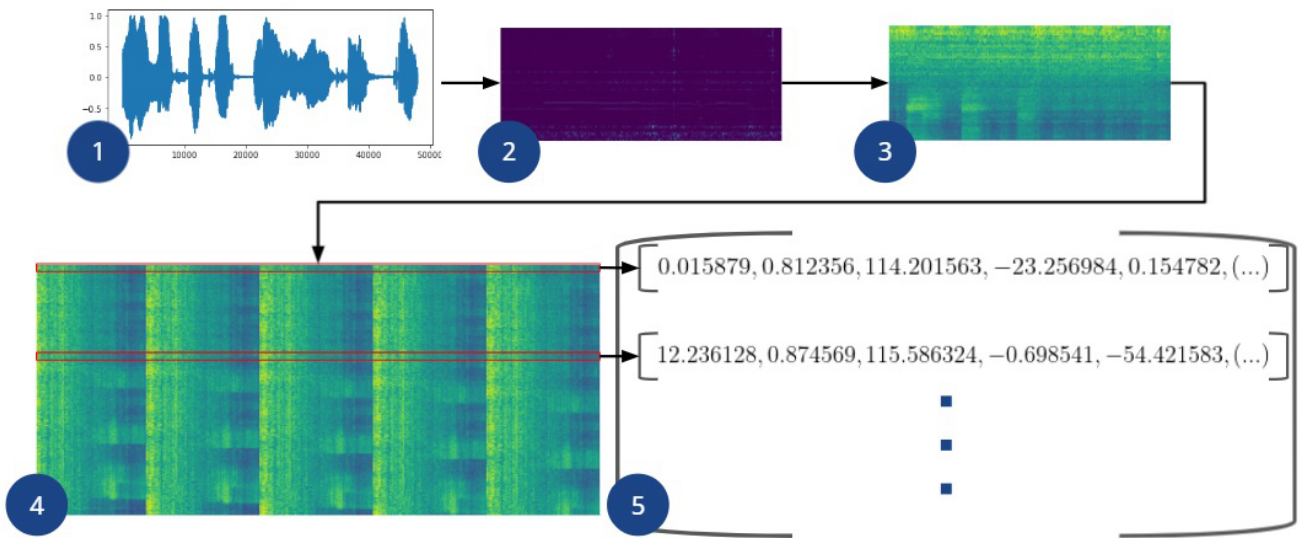
3.1 Pre-Processing Phase

As shown in Figure 1, the pre-processing is conducted through a pipeline consisting of five steps: (1) reading the digital audio signal; (2) applying the Short-Time Fourier Transform (STFT) function to the audio; (3) mapping the signal magnitude to decibels and converting the signal to a logarithmic scale; (4) time-windowed concatenation of generated spectrograms; and (5) transformation of these matrices into vectors. The goal of this phase is to transform audio signals in a standardized vector of features to be processed by all models. This method is based on DCASE challenge for audio anomaly detection [Koizumi *et al.*, 2020a].

In the first stage, the audio is read from the dataset, and the signal is converted into a matrix representing the time and frequency of the signal, known as the short-time spectrogram. In the second stage, feature extraction takes place with STFT which divides the signal into small time segments and, in each segment, calculates the Fourier transform, as shown in Eq. (1) for $x[n]$:

Table 1. Comparison of studies on Anomaly Detection using Deep Learning architectures.

Work	Architecture	Domain
Akçay <i>et al.</i> [2019a]	AEE+GAN	Image
Akçay <i>et al.</i> [2019b]	U-Net+GAN	Image
Zenati <i>et al.</i> [2018]	AE+GAN	Image
Cheng <i>et al.</i> [2021]	AE	Image
Liu <i>et al.</i> [2021]	U-Net(DSC+CBAM)+GAN	Image
Wilkinghoff [2023]	ResNet+KNN	Audio
Chen <i>et al.</i> [2024]	ResNet+KNN+MDAM	Audio
Koizumi <i>et al.</i> [2020a]	AE	Audio
Suefusa <i>et al.</i> [2020]	AE(IDNN)	Audio
Müller <i>et al.</i> [2021]	AE(DRINK)	Audio

**Figure 1.** Pre-processing pipeline.

$$X(m, k) = \sum_m x[n]W[m - n] \exp\left(-j\frac{2\pi \cdot n \cdot k}{N}\right), \quad (1)$$

where m is the number of time samples, k is the number of frequency samples, $W[n]$ is the window size and N is the number of samples of total frequencies. We assume that $W[n]$ is a fixed time window [Zhao *et al.*, 2015]. As a result, the audio is represented by one or more Mel-Frequency Cepstral coefficients, a set of features developed at MIT during the late 1960s, widely used in tasks such as speech recognition, sound source separation, and sound event detection [Dwivedi *et al.*, 2023]. By providing detailed information about the spectral distribution of audio over time, the Fourier transform and spectrogram play a fundamental role in feature extraction and audio pre-processing. [Chachada and Kuo, 2014].

During the third stage, the signal underwent mapping to decibels, constraining the signal values to audible frequencies for humans, as depicted in Equation (2). Following this, the function $m(\cdot)$ was applied as a threshold for negative frequencies, as illustrated in Equation (3).

$$d(S) = 10(\log_{10}(S) - \log_{10}(|S|)) \quad (2)$$

$$m(x) = \begin{cases} x, & -80 \text{ dB} \leq x \leq 130 \text{ dB}, \\ -80 \text{ dB}, & x < -80 \text{ dB}. \end{cases} \quad (3)$$

This stage was designed to amplify lower frequencies, rendering weaker signals comparable to stronger ones, thereby enhancing feature visualization on the spectrogram.

The fourth step involves the concatenation and time-windowing of the spectrograms. For this to be feasible, the spectrogram needs to have its total size reduced along the time axis, and then this axis is shifted for each new concatenated spectrogram. From *ad hoc experimentation* in order to ensure that each spectrogram has a different temporal characteristic from the others, it was observed that a proper concatenation considers time windows of length $\ell = 5$. Figure 2 illustrates that the spectrograms differ from each other along the y -axis as new spectrograms are added along the x -axis.

Finally, the resulting matrix is separated into vectors, where each row corresponds to 128 Mel-frequency features at $\ell = 5$ different time instances. Each audio instance has a standardized duration of 10 s, divided into frames of 64 ms, with a 50% overlap between frames, using the Hop algorithm. A total of 1024 points were utilized for Short-Time Fourier Transform (STFT), and 128 Mel coefficients were derived for each audio frame. Subsequently, $\ell = 5$ frames were concatenated, yielding a 5×128 matrix that was then

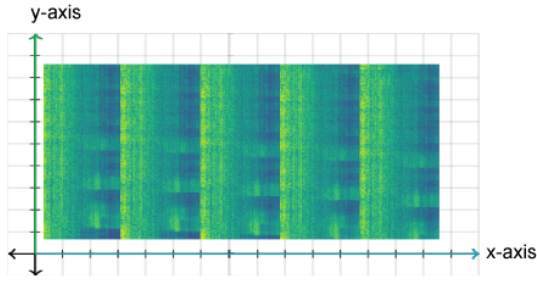


Figure 2. Example of Step 4.

flattened into a 640 feature vector. Following these outlined procedures, the resultant vector serves as a standardized input for all architectures.

3.2 Architectures Adaptation and Training

The adaptations enabling the use of GANs for anomaly detection in audio were based on the model proposed by Koizumi *et al.* [2020a], who introduced an AE network for audio. Consequently, the networks comprising the generator G and the discriminator D present an architecture consisting exclusively of Fully Connected Layers (FCN), encompassing three hidden layers in addition to the input and output layers. Except for the output layer of the decoder, all other layers utilize the Rectified Linear Unit (ReLU) activation function and Batch Normalization. Each hidden layer comprises 128 neurons and has a dimension of 8 in the latent space. This architecture, illustrated in Figure 3, was chosen for its relatively low number of parameters compared to the original networks, reducing computational cost and enabling the handling of audio instances prepared in the previous stage. These adaptations were incorporated to all architectures for comparison and standardization across the same dataset.

3.2.1 EGBAD

The audio anomaly detection for the EGBAD model consisted in changing the generator G , discriminator D and Encoder E as illustrated in Figure 4a. This architecture can map the input instance to a latent representation z during the training of the G and D networks. Unlike conventional GANs, this strategy involves specifying the input to the D network with real data, generated data, and the latent representation.

The training dynamics rely on the discriminator's ability to classify the input instance along with the latent vector. The process begins with sampling from the latent space z' . After that, an artificially generated instance is created from a random latent space z . Finally, the D network takes as input a real or artificial instance along with its corresponding latent vector.

3.2.2 GANomaly

The adaptation can be observed in Figure 4b, where it involves changing the G_E and E networks to the FCN Encoder, while the G_D and D models are converted to the FCN Decoder. The remaining details of the architecture remain unchanged from the original. The training dynamics are based on the assumption that the latent space generated by $E(z')$

accumulates more significant errors than those generated by $G_D(x')$.

Detailing the training process of this network, in the generator network G , the training instance x is encoded into a latent space z and decoded, thus generating an artificial version of the training instance called x' . The next step is the encoding of this artificial instance, called z' . In this way, the optimization can be performed for both the reconstruction of audio features and its latent representation. The discriminator network D classifies the training instances (typical data) and the artificial instances generated in the previous step. This classification occurs synthetically, as the labels for typical and artificial data are always considered as 0 and 1, respectively. The generator converges to create increasingly realistic audio features.

The feature combination of weights in the optimization functions is performed as defined by Eq. (4). We determined the weights of λ_{adv} , λ_{con} , and λ_{enc} through empirical values obtained during hyperparameter adjustments.

$$\mathcal{L}_{gen} = \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{con} \cdot \mathcal{L}_{con} + \lambda_{enc} \cdot \mathcal{L}_{enc}, \quad (4)$$

where \mathcal{L}_{adv} , \mathcal{L}_{con} and \mathcal{L}_{enc} denotes, respectively, the adversarial, contextual and encoder loss functions given in Eqs. (5)-(7). The first one ensures that the G network reconstructs the data as realistically as possible, while the D network correctly distinguishes between real or generated (fake) spectrograms; the second loss function calculates the L_1 distance between the real data x and the generated data; and the encoder loss function calculates the distance of original and reconstructed data representation at latent space.

$$\mathcal{L}_{adv} = E_{x \sim p_x} [\log D(y)] + E_{x \sim p_x} [1 - \log D(\hat{y})] \quad (5)$$

$$\mathcal{L}_{con} = \|x - G(x)\|_1 \quad (6)$$

$$\mathcal{L}_{enc} = \|z - \hat{z}\|_2 \quad (7)$$

3.2.3 SKIP-GANomaly

In SKIP-GANomaly, the networks G_E and G_D have been modified to Fully Connected Network (FCN) Encoder and Decoder networks, respectively. The FCN Encoder network has replaced the model D . Additionally, connections and concatenations have been introduced between the G_E and G_D networks, facilitating significant advantages in information transfer between the layers preserving local and global information. The adaptations in SKIP-GANomaly architecture are illustrated in Figure 5 where dotted lines represent copies of layer content extended to subsequent layers, and red blocks indicate layer concatenation. This architecture aims at enabling reconstructions that are closer to the audio sample.

This architecture's training dynamics are detailed as follows: In the generator network G , the network G_E captures and learns the distribution of the input data x (restricted to typical instances) and then maps them to latent representations z . Simultaneously, the discriminator network D classifies the received instances, differentiating between real images (x) and images generated in the previous step (x').

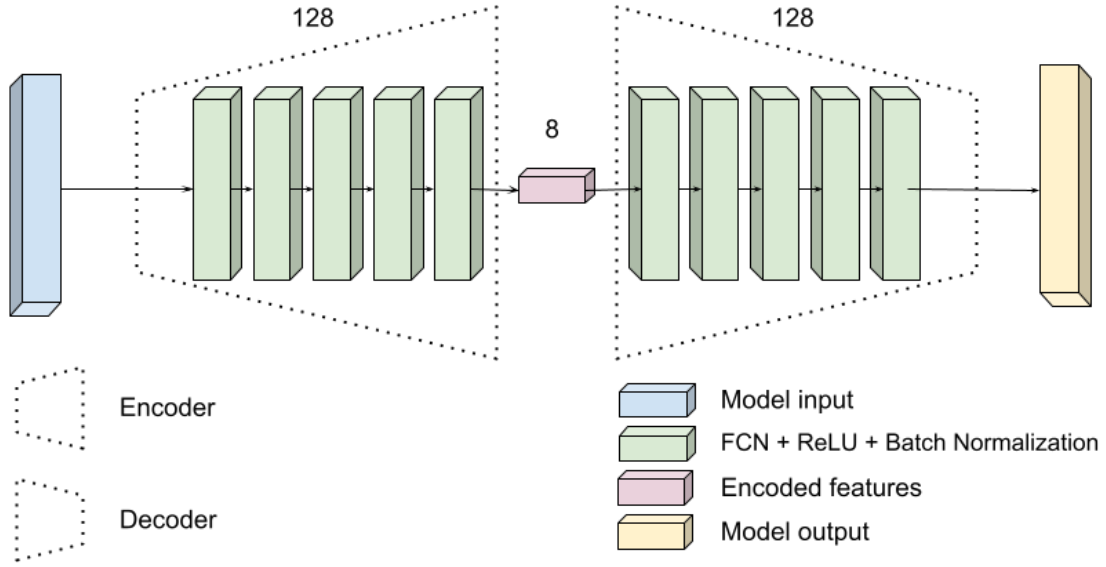


Figure 3. Architecture proposed by Koizumi et al. [2020a].

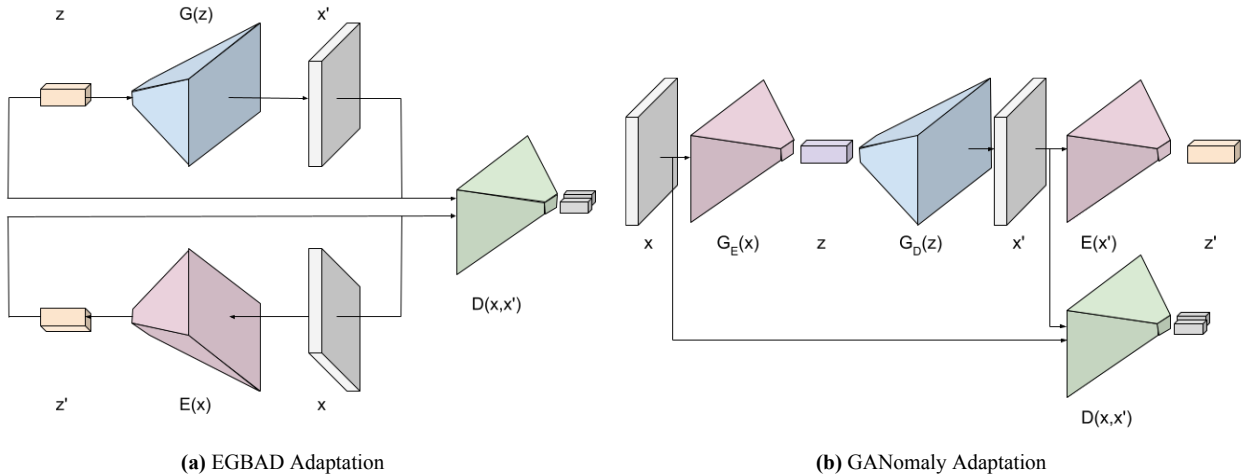


Figure 4. Adaption of GAN-based architectures.

While serving as a classifier, this network also operates as a feature extractor, approximating latent representations between an input audio and a reconstructed audio. This classification process mirrors that of the previous architecture.

Regarding feature combination of weights in the optimization functions, we implemented exactly the same GANomaly method described before and show in Equation 4.

4 Evaluation Metrics

The ROC (Receiver Operating Characteristic) curve is a graphical representation used in binary classification tasks to illustrate the performance of a classifier across different discrimination thresholds. It is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) for various threshold values. TPR and FPR are calculated as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (8)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad (9)$$

where TP, FN, FP and TN denotes the possible outcomes of a binary classification task: True Positive, False Negative, False Positive and True Negative. The ROC curve shows how the TPR changes as the FPR varies [Fawcett, 2006].

The Area Under the ROC Curve (AUC) assesses the separability between classes and the probability of the learning model having greater confidence in the predicted class. The partial AUC (pAUC) quantifies the proportion of the ROC curve over a range of interest. It mainly deals with class imbalance, giving more emphasis to TPR [Fawcett, 2006]. In the present work, this metric will be calculated over the low quantity of FPR, thus $p = 0.1$.

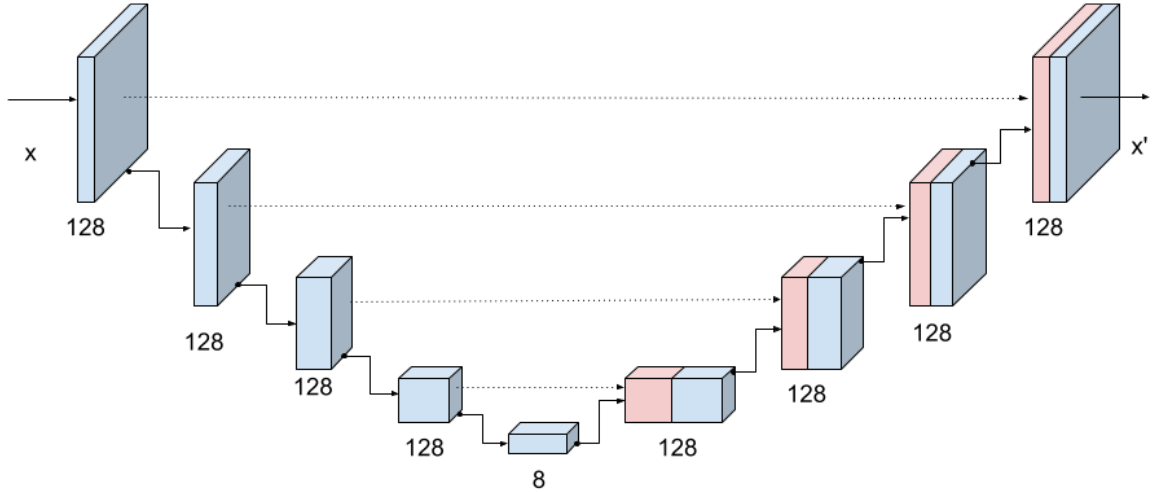


Figure 5. Skip-GANomaly Adaption in Generator Network.

5 Experiments

Python, in conjunction with the TensorFlow and PyTorch frameworks, served as the primary tools for training and evaluating the proposed models [Van Rossum and Drake, 2009; Abadi *et al.*, 2015; Paszke *et al.*, 2019]. Implementations were executed on a computational system equipped with an Intel® Core™ i7-8700 CPU running at a clock speed of 3.2 GHz, supported by 32 GB of primary memory, 2 TB of secondary memory and 2 NVIDIA GTX 1080 Ti with 11 GB VRAM each to promote hardware speedup when training the models.

5.1 Architectures and Dataset

We adapted the architectures of the semi-supervised learning models EGBAD [Zenati *et al.*, 2018], GANomaly [Akçay *et al.*, 2019a], and SKIP-GANomaly [Akçay *et al.*, 2019b] as depicted in the previous section. To validate such adaptations to the audio domain, we utilized real-world audio samples from ToyADMOS and MIMII datasets [Koizumi *et al.*, 2019; Purohit *et al.*, 2019]. The audio data comprises six types of industrial machinery: (i) Fan; (ii) Pump; (iii) Slide Rail; (iv) ToyCar, (v) ToyConveyor; and (vi) Valve. The first two classes mentioned belong to ToyADMOS, while the remaining classes are from the MIMII dataset. Audio instances were divided into typical and anomalous categories, each lasting approximately 10 s. However, it is important to note that during the training phase, only audio data representing typical operation was utilized. The data is structured based on identifiers (ID), labels, and the intended purpose of use.

Table 2 presents the distribution of labels between the training and test sets, where each label represents a specific type of object or event to be detected. The quantities of training and test examples differ for each class. For instance, the class ToyCar comprises of 4000 examples in the training set and 2459 examples in the test set, while the label ToyConveyor has 3000 training examples and 3509 test examples. Similar variations exist for the remaining classes between the training and test sets. These variations were taken into

account when obtaining the weighted average metric used in the models evaluation.

Table 2. Dataset size stratified by train and test.

Labels	Train		Test	
	Samples	%	Samples	%
Fan	3675	66.21 %	1875	33.79 %
Pump	3349	79.65 %	856	20.35 %
Slide Rail	2804	68.50 %	1290	31.50 %
ToyCar	4000	61.92 %	2459	38.08 %
ToyConveyor	3000	46.09 %	3509	53.91 %
Valve	3291	78.93 %	879	21.07 %

5.2 Baselines

DCASE[Koizumi *et al.*, 2020a] is one of the main competitions involving machine learning in audio tasks. Particularly, the anomaly detection task was started in 2020, and among the solutions presented, it was observed that the competition itself presented a base model (DCASE baseline) and the winning solution was GMADE. Besides the fact that any DCASE solution is based on GAN models, we selected these models to use as metric baselines to conduct a comparative analysis. They are described as follows:

- **DCASE baseline.** This model is the official baseline of the DCASE 2020 Challenge that involved unsupervised learning. As the initial baseline, its results helps validating the adaptations to GAN-based networks establishing inferior performance limits.
- **GMADE.** By incorporating metadata from the training set to classify audios, this model employs a semi-supervised learning method [Giri *et al.*, 2020]. Thus, in the context of the comparative evaluation, it represents the upper limit, assuming the availability of additional information not typically present in anomaly detection problems.

5.3 Experimental Results

All models underwent training on each audio class, and assessments were exclusively carried out on the test data partition. This guarantees that a model comprehensively learns the typical operations associated with each industrial machinery type. Consequently, the performance of a single model may differ across various audio types. It is crucial to emphasize that model training is conducted solely on audio samples, without the inclusion of any metadata. This is intended to demonstrate the effectiveness of these adaptations on real-world datasets.

The entire set of evaluated models were applied to data using the optimal parameters identified through a grid search methodology. The unique characteristics of the audio classes require different hyperparameters to achieve improved results from customized models for each case. The search space for hyperparameters are detailed in Table 3.

Table 3. Hyperparameters grid search space.

Parameters	Range
Epochs	{10, 20, 30, 90}
Learning Rate	{0.0002, 0.001, 0.005}
Latent space dimension	{8, 16, 100}
Dropout regularization	{0, 0.25, 0.5}
Weight Initialization	{Glorot Uniform, Random Normal}
Exponential Decay Rate	{0.3, 0.5}
Hidden Layers (HL)	{3, 4}
λ_{adv}	{1, 50}
λ_{con}	{1, 50}
λ_{enc}	{1, 50}
Batch Size (Batch)	{8, 16, 32, 128, 512, 1024}

Several optimizers were evaluated, considering their impact on training convergence, convergence speed, the ability to handle different gradient scales, and computational efficiency [Sutskever *et al.*, 2013]. However, the Adam optimizer demonstrated superior performance across all cases evaluated [Kingma and Ba, 2015]. Some grid search results for hyperparameters were consistent across different architectures. These common parameters include LR (0.0002), latent space dimension (8), λ value (1.0), weight initializer (Glorot Uniform), Dropout regularization (0), and exponential decay rate (0.5).

The evaluation outcomes for all audio categories of the modified EGBAD architecture [Zenati *et al.*, 2018] are depicted in the bar chart shown in Figure 6a. The model exhibited satisfactory performance solely for the Valve class, achieving AUC and pAUC metrics of 69% and 57%, respectively. However, its performance on the ToyConveyor, Fan, and Pump classes proved inadequate, with results falling below 50%. Additionally, it demonstrated AUC percentage deteriorations of 118.48%, 64.83%, and 125.70%, respectively, when compared to the DCASE baseline. These findings suggest that the model faced challenges in distinguishing between typical and anomalous audio across the majority of the dataset.

The previously mentioned experimental result emphasizes the crucial role of the latent vector in anomaly detection, as elaborated in Section 4.1. The EGBAD architecture relies on latent representations, either as a generator input or from the encoder, to gauge the discriminator’s confidence in deter-

mining if a sample originates from the typical data distribution. Nevertheless, this direct comparison of audio instances leads to an increased error score during the training phase, impeding the EGBAD architecture from attaining high metrics. Given its subpar performance compared to the baseline, this architecture was not chosen for the hyperparameter grid search.

After carrying out the grid search, the adapted GANomaly architecture [Akçay *et al.*, 2019a] demonstrated promising results, as illustrated in Figure 6a. The bar chart shows the performance of this model across all labels. AUC exceeded 90% for both ToyCar and fan classes, while all remaining labels have AUC metrics above 75%. These results indicate the model’s capacity of identifying and distinguishing anomalous data from typical data. The pAUC metric showed results above 75%, for most classes, with lower results noted for the pump and valve classes. Although these last results are higher than 50%, they suggest that the neural network faced challenges in generalizing cases where anomalous noise are close to typical frequencies in sound. Table 4 presents the best parameters for each audio label, exhibiting minimal variation compared to other evaluated models.

Table 4. Hyperparameters for the best performing GANomaly adapted architecture.

Labels	Epochs	HL	λ_{adv}	λ_{con}	λ_{lat}	Batch
Fan	10	3	1	50	1	16
Pump	10	3	50	1	50	1024
Slide Rail	10	3	1	50	1	16
ToyCar	90	4	1	50	1	512
ToyConveyor	30	4	1	50	1	512
Valve	10	3	1	50	1	16

Figure 6b illustrates the evaluation of the adapted SKIP-GANomaly architecture [Akçay *et al.*, 2019b] using a bar chart detailing results for all labels. The model presented both AUC and pAUC metrics exceeding 75% and 64%, respectively, for all audio types. The slider class achieved the highest AUC metric, and the fan label achieved the highest pAUC metric. Table 5 outlines the optimal parameters selected for each audio class. This model showed a higher parameter variation among the classes, requiring specific adjustments in batch size and weights assignment to the optimization function. This can be due to the residual structure of the model.

Table 5. Hyperparameters for the best performing SKIP-GANomaly adapted architecture.

Labels	Epochs	HL	λ_{adv}	λ_{con}	λ_{lat}	Batch
Fan	10	3	1	50	1	16
Pump	10	3	50	1	50	1024
Slide Rail	10	3	50	50	50	512
ToyCar	20	3	1	1	50	128
ToyConveyor	30	4	1	50	1	512
Valve	30	4	1	50	1	512

5.3.1 Overall Performance

In order to present an overall performance comparison among the models, we present the weighted average and stan-

dard deviations computed for all evaluated models in Figure 6d and Table 6. Once the test datasets present different amount of data, this results hold significance as higher results in extensively populated datasets have more statistical relevance. The evaluation reveals that the adapted GANomaly architecture has a slightly better performance than GMADE model in the AUC metric, and SKIP-GANomaly lags slightly behind. Additionally, the standard deviation of the AUC metric in the GANomaly and SKIP-GANomaly models are lower than that in the GMADE model, with values of 4.19% and 3.73%, respectively. Similar behavior can be observed with pAUC metric, but now GANomaly is behind GMADE with a little higher standard deviation. We can conclude that the GANomaly model has the best overall results by not using metadata as GMADE does, and this can be attributed to its characteristic of comparing latent vectors to analyse the context of an audio.

Table 6. Weighted Average Performance.

	AUC	pAUC
DCASE	73.81 % \pm 6.31	60.47 % \pm 14.76
EGBAD	46.23 % \pm 14.12	52.55 % \pm 15.47
GANomaly	88.16 % \pm 4.19	78.05 % \pm 11.77
SKIP-GANomaly	86.11 % \pm 3.73	72.64 % \pm 10.86
GMADE	88.02 % \pm 7.18	79.63 % \pm 11.05

Table 7 summarizes the model results for every audio dataset. It shows that GMADE has the best results for many audio classes on both AUC and pAUC metrics. As it uses metadata for anomaly detection, which is not a feasible approach for real applications based only on audio samples, it can be seen as a superior limit in our comparative evaluation. The adapted GANomaly and SKIP-GANomaly architectures present the best performance in the ToyConveyor, fan, and pump classes for both AUC and pAUC metrics, except by pAUC in pump class. Besides, they have superior performance than the other models used as baseline.

Regarding the results presented in previous work[Neto and Figueiredo, 2023], we can see that customising model training through grid search was important to achieve significant better results. Particularly, GANomaly increased overall AUC from 72 to 88%, and pAUC from 69 to 78%. SKIP-GANomaly increased overall AUC from 66 to 86%, and pAUC from 54 to 72%. These results are import to show that the GAN architectures must be trained on the specific audio class to better detect anomalies.

These results indicate that approaches relying on GAN networks have high potential for addressing anomaly detection in audio. The use of architectures, mainly GANomaly and SKIP-GANomaly, proved the possibility to perform anomaly detection in various industrial machinery contexts. It is worth mentioning that selecting the appropriate hyperparameters, i.e. learning rate, batch size, and the number of epochs, through the experimentation process was relevant to achieve satisfactory results. On the other hand, results also emphasize that challenges persist in anomaly detection in acoustic events. For instance, we can observe different model performances for different audio datasets, which indicates that some specific model characteristics are better for specific au-

dio characteristics. Particularly, DCASE dataset comprises different but repetitive industrial sound, and this is a characteristic that may differ in different scenarios such as urban sounds. This is a limitation of the present work that can be evaluated in future work. Besides, there is an open road to new research, such as building ensemble models and adopting other data pre-processing.

6 Final Remarks

This study demonstrates the effectiveness of GAN-based models for audio anomaly detection. Three GAN-based anomaly detection models, originally designed for images, were successfully adapted to the audio domain. The evaluation utilized the well-established DCASE Challenge Dataset. The adapted GANomaly and SKIP-GANomaly models exhibited superior performance compared to the baselines from the literature.

The research presented herein offers performance benchmarks for GAN models in the audio domain, distinguishing them from their original applications and other works in the literature. By adapting these models, we have established metric benchmarks that will serve as a foundation for future advancements. It worth to mention that the observations are limited for the DCASE challenge dataset, which was used to compare proposals in literature, but it comprises a particular type of repetitive industrial audio. So, there is an open road for new research to obtain experimental results in other scenarios.

Future work may encompass: (i) assessing the outcomes of employing data augmentation techniques on original audio. Many classification and regression works use data augmentation techniques in detecting anomalies, such as [Bakir *et al.*, 2024; Wang *et al.*, 2023; Nanni *et al.*, 2020]. However, careful data augmentation is crucial, given the high sensitivity of the task in the latent space of typical occurrences. Sound activities, in particular, are sensitive to distortions, such as frequency patterns and temporal variations, requiring a balance between the amount of data and the preservation of relevant characteristics [Wei *et al.*, 2020]; (ii) exploring variations or combinations of adversarial models in ensembles. The current objective of this work was to evaluate several GANs models. However, ensembles may enhance performance by combining multiple models, each with its unique approach and strengths, to mitigate model biases and errors. In the context of audio analysis, ensembles can capture diverse aspects of sound patterns and spectral features; (iii) investigating diverse preprocessing techniques for audio features, including the use of varying numbers of Mel coefficients and assessing their impact.

Authors' Contributions

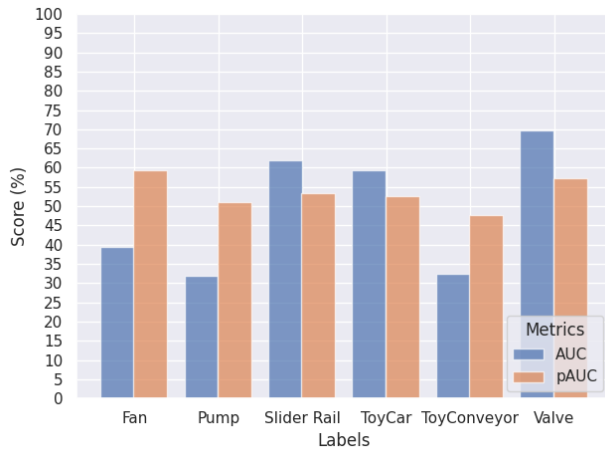
All authors contributed equally to this work and approved the final version of the manuscript.

Competing interests

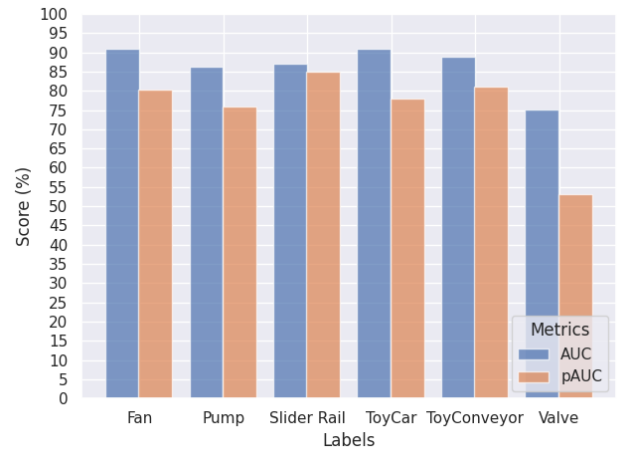
The authors declare that they have no conflict of interest.

Table 7. Baseline comparison.

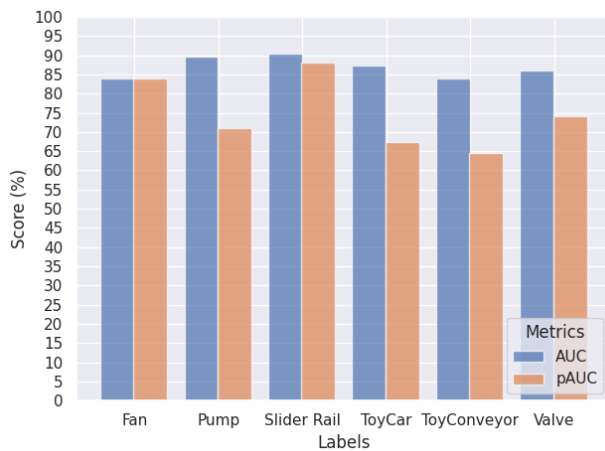
	Fan		Pump		Slide Rail		ToyCar		ToyConveyor		Valve	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
DCASE	65.15 %	52.59 %	72.00 %	60.00 %	84.00 %	66.00 %	80.09 %	67.22 %	72.68 %	60.65 %	66.00 %	50.00 %
EGBAD	39.50 %	59.40 %	31.90 %	51.10 %	62.00 %	53.40 %	59.40 %	52.60 %	32.40 %	47.70 %	69.80 %	57.30 %
GANomaly	91.00 %	80.00 %	86.19 %	75.81 %	87.16 %	85.11 %	91.00 %	78.10 %	88.80 %	81.20 %	75.00 %	53.00 %
SKIP-GANomaly	84.00 %	84.00 %	89.52 %	71.04 %	90.49 %	88.09 %	87.41 %	67.42 %	83.90 %	64.60 %	86.00 %	74.00 %
GMADE	82.33 %	78.97 %	86.94 %	79.60 %	97.28 %	89.54 %	95.04 %	90.39 %	80.67 %	65.90 %	97.38 %	91.21 %



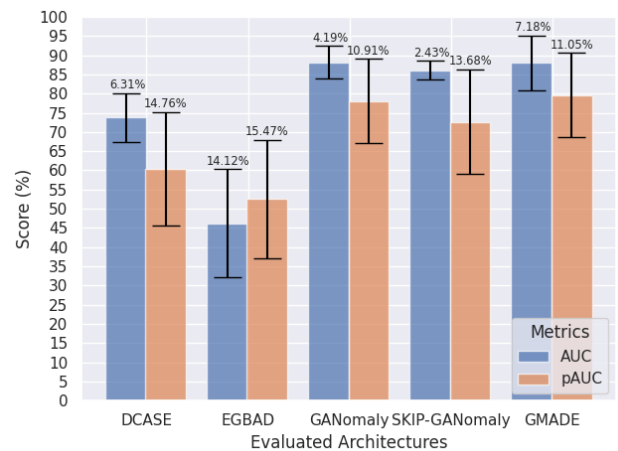
(a) EGBAD



(b) GANomaly



(c) SKIP-GANomaly



(d) Overall evaluation, showing weighted averages and standard deviations for each class.

Figure 6. Evaluation results for adapted models.

Availability of data and materials

The DCASE 2020 challenge and MIMII datasets used in the current work are available at <https://dcase.community/challenge2020/index> and <https://zenodo.org/records/3384388>.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Is-

ard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Available at: <https://www.tensorflow.org/>.
 Akcay, S., Atapour-Abarghouei, A., and Breckon, T. (2019a). GANomaly: Semi-supervised anomaly detection via adversarial training. In *Lecture Notes in Computer Science*, page 622–637. Springer International Publishing. DOI:

- 10.1007/978-3-030-20893-6_39.
- Akçay, S., Atapour-Abarghouei, A., and Breckon, T. (2019b). Skip-GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. DOI: 10.1109/IJCNN.2019.8851808.
- An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18. Available at: <https://api.semanticscholar.org/CorpusID:36663713>.
- Bakır, H., Çayır, A. N., and Navruz, T. S. (2024). A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification. *Multimedia Tools and Applications*, 83(6):17601–17628. DOI: 10.1007/s11042-023-16200-4.
- Chachada, S. and Kuo, C.-C. J. (2014). Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, 3:e14. DOI: 10.1109/APSIPA.2013.6694338.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3). DOI: 10.1145/1541880.1541882.
- Chen, S., Wang, J., Wang, J., and Xu, Z. (2024). Mdam: Multi-dimensional attention module for anomalous sound detection. In Luo, B., Cheng, L., Wu, Z.-G., Li, H., and Li, C., editors, *Neural Information Processing*, pages 48–60, Singapore. Springer Nature Singapore. DOI: 10.1007/978-981-99-8178-6_4.
- Cheng, Z., Zhu, E., Wang, S., Zhang, P., and Li, W. (2021). Unsupervised outlier detection via transformation invariant autoencoder. *IEEE Access*, 9:43991–44002. DOI: 10.1109/ACCESS.2021.3065838.
- Dwivedi, D., Ganguly, A., and Haragopal, V. (2023). Contrast between simple and complex classification algorithms. In Goswami, T. and Sinha, G., editors, *Statistical Modeling in Machine Learning*, pages 93–110. Academic Press. DOI: 10.1016/B978-0-323-91776-6.00016-6.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874. DOI: 10.1016/j.patrec.2005.10.010.
- Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. DOI: 10.1109/ICASSP.2017.7952261.
- Giri, R., Tenneti, S. V., Helwani, K., Cheng, F., Isik, U., and Krishnaswamy, A. (2020). Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation. Technical report, DCASE2020 Challenge. Available at: <https://api.semanticscholar.org/CorpusID:246825245>.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, Barcelona. Curran Associates, Inc. Book.
- Kawaguchi, Y., Imoto, K., Koizumi, Y., Harada, N., Niizumi, D., Dohi, K., Tanabe, R., Purohit, H., and Endo, T. (2021). Dcase 2021 challenge task 2 development dataset. DOI: 10.5281/zenodo.4562016.
- Kingma, D. P. and Ba, J. L. (2015). Adam: a method for stochastic optimization. In *Proceedings of 3rd International Conference on Learning Representations*, San Diego, CA. ArXiv. DOI: 10.48550/arXiv.1412.6980.
- Kittler, J., Kaloskampis, I., Zor, C., Xu, Y., Hicks, Y., and Wang, W. (2018). Intelligent signal processing mechanisms for nuanced anomaly detection in action audio-visual data streams. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6563–6567. DOI: 10.1109/ICASSP.2018.8461595.
- Koizumi, Y., Kawaguchi, Y., and Imoto, K. (2020a). Description and discussion on dcase2020 challenge task2: unsupervised anomalous sound detection for machine condition monitoring. DOI: 10.48550/arXiv.2006.05822.
- Koizumi, Y., Kawaguchi, Y., Imoto, K., Nakamura, T., Nikaido, Y., Tanabe, R., Purohit, H., Suefusa, K., Endo, T., Yasuda, M., and Harada, N. (2020b). Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. DOI: 10.48550/arXiv.2006.05822.
- Koizumi, Y., Saito, S., Uematsu, H., Harada, N., and Imoto, K. (2019). ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 308–312. DOI: 10.1109/WASPAA.2019.8937164.
- Langr, J. and Bok, V. (2019). *Generative Adversarial Networks in Action – Deep Learning with Generative Adversarial Networks*. Manning Publications, Shelter Island. Book.
- Liu, G., Lan, S., Zhang, T., Huang, W., and Wang, W. (2021). Sagan: Skip-attention gan for anomaly detection. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2468–2472. DOI: 10.1109/ICIP42928.2021.9506332.
- Müller, R., Illium, S., and Linnhoff-Popien, C. (2021). Deep recurrent interpolation networks for anomalous sound detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. DOI: 10.1109/IJCNN52387.2021.9533560.
- Nanni, L., Maguolo, G., and Paci, M. (2020). Data augmentation approaches for improving animal audio classification. *Ecological Informatics*, 57:101084. DOI: 10.1016/j.ecoinf.2020.101084.
- Neto, W. O. and Figueiredo, C. (2023). Análise de redes GANs para detecção de anomalias em atividade sonoras. In *Anais do XV Simpósio Brasileiro de Computação Ubíqua e Pervasiva*, pages 11–20, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbcup.2023.230034.
- Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38. DOI: 10.1145/3439950.

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc. Available at: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- Purohit, H., Tanabe, R., Ichige, T., Endo, T., Nikaido, Y., Suefusa, K., and Kawaguchi, Y. (2019). MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pages 209–213. DOI: 10.48550/arXiv.1909.09347.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597. DOI: 10.48550/arXiv.1505.04597.
- Rovetta, S., Mnasri, Z., and Masulli, F. (2020). Detection of hazardous road events from audio streams: An ensemble outlier detection approach. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pages 1–6. DOI: 10.1109/EAIS48028.2020.9122704.
- Sabui, M., Zhou, M., Bezemer, C.-P., and Musilek, P. (2021). Applications of generative adversarial networks in anomaly detection: A systematic literature review. *IEEE Access*, 9:161003–161029. DOI: 10.1109/ACCESS.2021.3131949.
- Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 1041–1044, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2647868.2655045.
- Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. (2019). f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30 – 44. DOI: 10.1016/j.media.2019.01.010.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117. DOI: 10.1016/j.neunet.2014.09.003.
- Suefusa, K., Nishida, T., Purohit, H., Tanabe, R., Endo, T., and Kawaguchi, Y. (2020). Anomalous sound detection based on interpolation deep neural network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 271–275. DOI: 10.1109/ICASSP40776.2020.9054344.
- Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA. PMLR. Available at: <https://proceedings.mlr.press/v28/sutskever13.html>.
- Tanabe, R., Purohit, H., Dohi, K., Endo, T., Nikaido, Y., Nakamura, T., and Kawaguchi, Y. (2021). MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions. In *arXiv e-prints: 2006.05822*, 1–4. DOI: 10.48550/arXiv.2105.02702.
- Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA. Book.
- Wang, Q., Du, J., Wu, H.-X., Pan, J., Ma, F., and Lee, C.-H. (2023). A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1251–1264. DOI: 10.1109/TASLP.2023.3256088.
- Wei, S., Zou, S., Liao, F., and weimin lang (2020). A comparison on data augmentation methods based on deep learning for audio classification. *Journal of Physics: Conference Series*, 1453(1):012085. DOI: 10.1088/1742-6596/1453/1/012085.
- Wilinghoff, K. (2023). Design choices for learning embeddings from auxiliary tasks for domain generalization in anomalous sound detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. DOI: 10.1109/ICASSP49357.2023.10097176.
- Xu, W., Jang-Jaccard, J., Singh, A., Wei, Y., and Sabrina, F. (2021). Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset. *IEEE Access*, 9:140136–140146. DOI: 10.1109/ACCESS.2021.3116612.
- Zenati, H., Foo, C. S., Lecouat, B., Manek, G., and Chandrasekhar, V. R. (2018). Efficient gan-based anomaly detection. DOI: 10.48550/arXiv.1802.06222.
- Zhao, Y., Zou, Z., Wu, L., and Li, Y. (2015). Frequency detection algorithm for frequency diversity signal based on STFT. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, pages 790–793. DOI: 10.1109/IMCCC.2015.173.
- Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 665–674, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3097983.3098052.
- Zhou, X., Xiong, J., Zhang, X., Liu, X., and Wei, J. (2021). A radio anomaly detection algorithm based on modified generative adversarial network. *IEEE Wireless Communications Letters*, 10(7):1552–1556. DOI: 10.1109/LWC.2021.3074135.