Multimedia Tools and Applications (2022) 81:5537–5586
https://doi.org/10.1007/s11042-021-11817-9

Anomalous sound event detection: A survey of machine learning based methods and applications

Zied Mnasri[1,2] · **Stefano Rovetta**[2] · **Francesco Masulli**[2]

Received: 26 April 2021 / Revised: 11 August 2021 / Accepted: 14 December 2021 /
Published online: 27 December 2021

Abstract
With the development of multi-modal man-machine interaction, audio signal analysis is gaining importance in a field traditionally dominated by video. In particular, anomalous sound event detection offers novel options to improve audio-based man-machine interaction, in many useful applications such as surveillance systems, industrial fault detection and especially safety monitoring, either indoor or outdoor. Event detection from audio can fruitfully integrate visual information and can outperform it in some respects, thus representing a complementary perceptual modality. However, it also presents specific issues and challenges. In this paper, a comprehensive survey of anomalous sound event detection is presented, covering various aspects of the topic, i.e.feature extraction methods, datasets, evaluation metrics, methods, applications, and some open challenges and improvement ideas that have been recently raised in the literature.

Keywords Anomalous sound event detection · Feature extraction · Supervised learning · Unsupervised learning · Evaluation metrics

1 Introduction

Presentation of the topic *Anomalous* sound event detection (*anomalous* SED) is a relatively novel topic in audio and speech processing. It lies at the intersection of digital signal processing, in particular audio and speech processing, anomaly detection and machine learning. The number of applications is growing fast, as it started as an alternative/complementary method to video analysis to encompass a large range of applications from industrial monitoring to home assistants. This topic covers two main fields, (a) anomaly/outlier/

✉ Zied Mnasri
zied.mnasri@enit.utm.tn

Stefano Rovetta
stefano.rovetta@unige.it

Francesco Masulli
stefano.rovetta@unige.it

[1] Electrical Eng. Dept., ENIT, University of Tunis El Manar, Tunis El Manar, Tunisia

[2] University of Genoa, Genoa, Italy

novelty detection, and (b) sound/ audio/acoustic event detection. Each topic belongs to separate realm, as anomaly detection is a general problem, whereas SED is a specific application within signal processing and understanding.

Interest of the survey Automated surveillance applications are dominated by video. A count of some Google Scholar search results confirms this fact (Fig. 1), but at the same time shows that the interest in audio is relevant.

Compared to video, the audio modality offers some unique advantages: (i) In addition to the installation cost, audio stream acquisition is much less expensive in terms of bandwidth, memory storage and computational requirements; (ii) omnidirectional microphones and/or microphone arrays can cover a 360° perception field, and are insensitive to luminosity and many weather conditions; (iii) unlike video, most human-hearing-range sounds can be detected in presence of physical obstacles, even if this may be a problem for some tasks, e.g., localization; (iv) some relevant events for audio surveillance, like gunshots and screams, are more perceptible through audio than video; (v) audio data have more chance to be classified into categorical and strictly separable events than video scenes.

Another key remark is that, in principle, anomalous audio event detection has a wider scope than just surveillance. This survey covers examples of works showing that anomalous SED is able to provide solutions for a large range of other applications, such as (i) industrial equipment monitoring, including fault detection and machine condition monitoring, (ii) audio scene segmentation for automatic summarization and language acquisition, (iii) healthcare monitoring, using biological audio signals such as the phonocardiogram (PCG) and respiration or cough sounds, for early heart or respiratory disease diagnosis.

Present survey Several available surveys and reviews focus on anomaly detection. An inventory of these works has already been presented in [20], where previous anomaly detection reviews were categorized according to the described techniques or to the target applications. Technique based reviews include those relying on classification, clustering, nearest neighbour, statistical, information theoretic and spectral techniques. Application-based reviews are mostly interested in cyber-intrusion, fraud, medical anomaly, industrial damage, image processing, textual anomaly ad sensor networks.

On the other side, several reviews have recently had SED applications as a target topic such as [10, 19, 21, 142]. Also, a systematic review about anomalous SED [88] has been recently made available on arxiv.org, but still not published in any field-related journal.

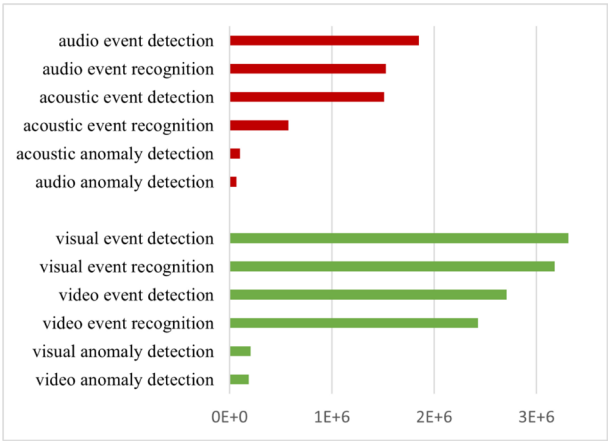

Fig. 1 Topic interest: Number of results of some Google Scholar queries

However, to the best of our knowledge, no extensive survey specifically focusing on anomalous SED has been recently published in indexed journals or proceedings.

2 Survey methodology

2.1 Anomaly definitions and assumptions

This survey is focused on *anomaly* detection, so it covers specifically those SED applications where the challenge is to recognize anomalous events rather than to segment an acoustic scene into known, typical event categories (event recognition/classification). Therefore, it is first necessary to define the notion of anomaly/novelty /outlier. In [124], this is characterized by means of (a) its scarcity, as anomalous/novel/outlier events occur less frequently than "normal" events; (b) its characteristics, as anomalous/novel/outlier events should have different characteristics than "normal" events; (c) its meaning, as such events should carry a specific and a different meaning than normal events. However, such a definition may be questionable, as: (i) scarcity as a criterion for anomaly definition may lead to a high rate of false positives, as many normal events could also occur less frequently; (ii) difference in characteristics, as can be measured by a high distance in the feature space, may not be enough to decide about outlierness; (iii) a-priori definitions of anomalous events refer to a classification where the qualification of outlierness depends above all on the limited number of classes. To overcome this narrow domain-related way of defining anomaly/novelty/outlier events, Xiang & Gong [143] suggest to define anomaly as an event that occurs infrequently; however, the concept of outlierness depends on the context and may change over time. In the same line, Dee & Hogg [31] define anomalies those events that cannot be explained by a "normal" model, as reported in [124].

2.2 Challenges of anomalous SED

2.2.1 Time-structured data

In addition to specifying the type of the acoustic event, SED aims at labeling its onset and offset instants. This reminds similar applications like speaker diarization [130] and automatic music transcription [14], where the interest is focused on turn changes rather than individual events [126].

2.2.2 Polyphony vs. monophony

Naturally, real life audio data should not be necessarily monophonic, however salient events are more likely to be detected in monophonic sounds, or at the presence of a reduced level of background noise.

In a recent review, [19] argued that SED should be more *challenging* for polyphonic sounds than for monophonic ones, for the following reasons: (a) Since polyphonic sounds contain a mixture of signals, single sound events can easily coincide, hence become less likely to be identified as advanced by [43, 44, 72]; (b) For the same reasons, the extracted features within a frame for example do not necessarily correspond to each separate audio signal, but rather to the mixture of sounds [47]; (c) Because of the mixture effect in polyphonic sounds, the number of events to be detected is not known

a priori. We also believe that these difficulties are confirmed by the absence of a model able to represent a mixture of sounds. In fact, whereas speech/music signal generation can be achieved using a source-filter model or a generative vocoder, like WaveNet [92] or World [81], the reverse operation, i.e. identifying the components in a polyphonic sound is still problematic.

Some proposed solutions in case of multi-source audio data have been proposed. For instance, source separation can be achieved using a multi-source probabilistic model [127], whereas a noise reduction technique can be applied in case of a dominant source in presence of background noise [136]. Also approaches used for SED vary according to polyphony or monophony: probabilistic methods, such as PLSA (probabilistic latent semantic analysis) [78] are applied to detect overlapping audio events, whereas analytical methods like non-negative matrix factorization (NMF) are more adapted to detect non-overlapping audio events [28].

2.2.3 Anomalous data scarcity

Another major problem is related to the inherent scarcity of data, intrinsic in the definitions of anomaly. Some methods are better suited to this problem, in particular semi-supervised and unsupervised methods [132].

2.3 Sources

The aim of this survey is to present a comprehensive overview about anomalous SED and to provide a structured view of the topic. To this end, 128 papers were selected according to the following criteria:

- Papers were collected within the interest areas of audio/ signal processing and machine learning.
- Papers were identified primarily by accessing authoritative repositories of works in these areas, particularly the DCASE challenges, the ICASSP, EUSIPCO, InterSpeech conferences, and relevant IEEE Transactions (IEEE T. Neural Networks and Learning Systems, IEEE T. Audio and Speech Processing, IEEE T. Multimedia). Google Scholar searches provided links to additional works.
- The topic of papers was specifically anomaly detection.
- Priority was given to more recent works, dating back at most from 2013, although some works using more traditional approaches were older.
- Since the surveyed topic is relatively recent, most of developed methods are not shared by different authors, especially when the method or the application are too specific. Therefore, we found that for each particular method or application, there are usually a few contributions (generally equal or less than 2).
- However, the number of contributions is growing very fast since 2017. A comparison of the publication dates on Google scholar shows that for the sequence of key words {*anomalous, sound event detection, machine learnnig*}, the total number of publications is 27700, among which 62.5% date only from 2017, and more particularly 28.7% from 2020.

2.4 Scope and organization of the survey

The organization of the remainder of the paper is as follows: The next Section 3 presents the main datasets, with a focus on public ones, and especially those used for developing and benchmarking in the state of the art. In Section 4, features are thoroughly described, either the hand-crafted ones, or those relying on feature extraction techniques, particularly applied for anomalous SED. We also opted to present first the standard evaluation metrics in Section 5; then, in Section 6 the modeling techniques are thoroughly detailed. For every type of method, we opted to present first an overview of the general framework, then to expose the models developed under each framework. Following this organization logic, we present in Section 7 an inventory of the main applications of anomalous SED. Finally, Section 8 presents the open challenges and the problems that are still awaiting more satisfactory solutions, in addition to some ideas that were proposed in the literature to improve the overall or some particular issues in the surveyed topic.

3 Datasets

anomalous SED is most often approached as a data-driven problem. Therefore a variety of datasets has been elaborated to allow training and developing models, for each specific domain of application (cf. Table 1).

Table 1 Most used datasets for training and benchmarking of anomalous SED models

Field	Application	Datasets	Authors and references
Industrial monitoring	Motor sound monitoring	ToyADMOS	(Koizumi et al., 2019) [62]
	Industrial machine inspection	MIMII	(Purohit et al., 2019) [107]
Traffic monitoring	Road audio surveillance	MIVIA	(Foggia et al., 2015) [37]
	Car crash monitoring	AXA	(Sammarco et al. 2018) [118]
	Environmental noise monitoring	WASN	(Alsina-Pages et al., 2019) [6]
General purpose SED	Office sounds	Office-live (OL)	(Stowell et al., 2015) [126] [80]
		Office-synthetic (OS)	(Stowell et al., 2015) [126]
	Real-life SED	TUT dataset	(Mesaros et al., 2016) [80]
		Google's Audio Set	(Gemmeke et al., 2017)[40]
		Freesound	(Fonseca et al.,2017)[38]
		Urbansound 8K	(Salomon et al.,2014)[117]
		SINS	(Dekkers et al., 2017)[32]
Human healthcare	Phonocardiogram anomaly	MITHSDB	(Syed et al., 2007) [129]
		AADHSDB	(Shmidt et al.,2010) [119]
		AUTHHSDB	(Papadaniil et al. 2013) [94]
		Other PCG datasets	(Liu et al., 2016) [71]
	Respiration sound anomaly	ICBHI	(Rocha et al., 2017) [112]

3.1 Industrial equipment monitoring

3.1.1 Motor sound monitoring

ToyADMOS [62] is a dataset dedicated for motor sound monitoring, developed for the DCASE 2019 challenge. It consists of recorded sounds of three toy motors: a toy car designed for product inspection task, a toy conveyor designed for fault diagnosis of a fixed machine, and a toy train designed for fault diagnosis of a moving machine. The database is divided into three subsets, individual sounds (IND), continuous sounds (CONT) and environmental sounds (ENV).

3.1.2 Machine condition monitoring

MIMII [107] is a sound dataset for Malfunctioning Industrial Machine Investigation and Inspection. It includes various normal and anomalous sounds, recorded in real-life conditions. It has been recently proposed as supporting material for the industrial machine inspection task in DCASE 2020 challenge. Normal and anomalous sounds come from different sources, namely valve, pump, fan and slide rail.

3.2 Traffic monitoring

3.2.1 Road events

MIVIA dataset [37] was designed for an audio-based road surveillance system. Recordings were realized in a real road environment at 23 locations in the province of Salerno, Italy, covering city center, highways and country roads. Two audio events, car crash and tire skidding, are considered, whereas all other events are considered as background noise. The total duration of the database is approximately one hour, divided into 57 audio clips.

3.2.2 Road crash test

AXA dataset [118] was collected at 2016 for the crash test campaign in Switzerland, organized by AXA insurance company. It contains 6.2 GB of audio data, exclusively recorded inside car cabins. 46 audio clips of car crashes are included, annotated with the car speed and the impact angle at the crash time.

3.2.3 Road noise characterization

WASN dataset [6] was recorded in the framework of DYNAMAP European Life+ project. It consists of 156 hours and 20 minutes of audio clips recorded at 24 acoustic nodes distributed on the A90 highway surrounding the suburban area of Rome. It was recorded with the main concern of collecting environmental noise samples, necessary for the study of different anomalous sound events.

3.3 Generic sound event detection datasets

Notwithstanding anomalous SED is a particular case of SED, several works have relied on generic SED databases to develop models for anomalous SED. This has particularly

been the case of DCASE challenges [74, 104, 138, 139], where some databases have been utilized both for general-purpose and anomalous SED tasks.

3.3.1 Office sound events

For the purpose of sound event detection in the framework of DCASE 2013 challenge, two databases were prepared: event detection office live (OL) and event detection office synthetic (OS) [126]. Both were designed for detecting predominant events in the presence of background noise.

The OL dataset consists of 24 recordings of individual sounds per class for training, 3 recordings of scripted sequences for validation, and 11 recordings for test.

The OS dataset consists of 12 synthetic sequences created from clips from the OL dataset with varying duration. These were equally divided into 3 subsets, each having a different level of event density: low (1.11), medium (1.27) and high (1.81).

3.3.2 Acoustic scene analysis

The TUT acoustic scenes 2016 is a database for environmental sound research. A subset, named TUT sound event dataset, was used for the DCASE 2016 challenge [80]. The TUT acoustic scenes database was recorded in 15 acoustic scenes, varying from outdoor and indoor environments. The audio events were labeled and inventoried, as detailed in [80].

3.4 Urban sounds

Different datasets describing different sound sources are gathered in the framework of Freesound [38], which is a large repository containing more than 160,000 audio recordings provided by several contributors under a creative commons (CC) license. In particular, UrbanSound 8K provides different types of urban sounds, such as human, natural, mechanical and music sounds, distributed on 1302 audio recordings of different duration, varying from 1-2 sec for gunshot to 30 sec for jackhammer or idling machine [117].

3.4.1 Multiple events

Google's Audio Set is a corpus of audio segments extracted from YouTube, containing YouTube identifiers, start time, end time and one or more labels for each segment. Each audio clip has a duration of 10 seconds, except those extracted from shorter video clips. In total, Google's Audio Set contains 1789,621 segments covering 4971 hours, including more than 100 instances for 485 audio event categories [40].

3.5 Healthcare

Heart anomaly diagnosis through sound has also been an interesting subject of anomalous SED. In 2016, the PhysioNet/CinC challenge tried to collect heart sound databases

from cardiology departments from worldwide universities. Nine teams were admitted for participation, each providing its own database. In this survey, only the most relevant ones are cited (cf. Table 1) . Further details about the challenge and the participating teams and their databases are in [71]. In such databases, the PCG (Phonocardiogram) has been recorded for different categories of diseases, such as mitral valve prolapse (MVP), aortic disease (AD) and miscellaneous pathological conditions (MPC) in [129], coronary artery disease (CAD) in [119], aortic stenosis (AS) and mitral regurgitation (MR) in [94].

4 Features

Initially, anomalous SED was based on standard features usually used for audio signal characterizations in the time and frequency domains, using different types of hand-crafted audio features. However, with the subsequent development of end-to-end learning methods, data-driven feature extraction, or *representation learning*, is currently a popular alternative.

It is worth noting that since anomalous SED spans a large set of applications, no particular standard sets of audio features have been designed for the particular purpose of anomaly detection in audio signals. In fact, most of models developed for anomalous SED rely on standard low-level features, commonly used for SED, or on tailored techniques of feature extraction, through either feature learning/embedding or basic signal or spectrogram-image processing methods, as will be detailed hereafter.

4.1 Hand-crafted audio features

Different sets of hand-crafted audio features have been proposed in the literature for anomalous SED. However, most of them are based on the same concept, i.e., statistical descriptors of low-level quantities computed in the time, frequency and multi-resolution domains.

4.1.1 Low-level descriptors (LLD)

Ntalampiras et al. 2011 [87] presented an inventory of the main LLD's used as input features for novelty detection in SED. The main rationale behind merging different-domain types of features is the general thought that this may improve robustness and performance of the SED system [87]:

MPEG-7 audio protocol features namely spectrum flatness, waveform min, waveform max and fundamental frequency (F0). The advantage of using these features is their compact representation of the waveform shape, periodicity and flatness of the spectrum in different frequency bands. The extraction of these features is standardized by the MPEG-7 audio protocol [57].

Mel-frequency cepstral coefficients (MFCC) The Mel-frequency cepstrum is a representation based on the cosine transform of a log-power spectrum, computed on a biologically-motivated nonlinear scale of frequency, i.e. the *Mel* scale [125]. The computed coefficients, i.e. the MFCC, have long been used for speech and speaker recognition for their considerable efficiency and their ability to capture the gross spectral characteristics of an audio event. Usually, 13 MFCC coefficients, in addition to their first and second

derivatives (Δ−MFCC and Δ−Δ−MFCC), are extracted from the Mel-log spectrum, including MFCC(0) that represents the log-energy [108].

Other cepstral coefficients In addition to MFCC, other coefficients are extracted from the cepstrum, such as linear prediction cepstral coefficients (LPCC), derived from LPC (Linear predictive coding) and Gammatone frequency cepstral coefficients (GFCC) using a Gammatone filter bank, instead of the Mel-scale filter bank. Both types of coefficients have been used in general-purpose audio event detection, either as low-level descriptors [8], or as high-level ones, i.e.as a bag of features (super-features) [103].

Intonation and Teager energy operator (TEO) based features These features describe the change of speech and intonation in case of anomalous vocal events, such as stress in speech. Associated to other speech-related features, such as F0, Δ−F0 and harmonic-to-noise ratio (HNR), they are useful to recognize speech signal produced under anomalous conditions.

Perceptual wavelet packet (PWP) integration analysis Using multi-resolution-based parameters is thought to reflect the degree of variability of wavelet coefficients within a particular frequency band.

4.1.2 DCASE 2013 challenge standard feature set

To the best of our knowledge, there has not been a dedicated LLD feature set especially proposed for anomalous SED, neither in literature nor in any DCASE challenge. In fact, LLD features are mainly focused on extracting some particular acoustic cues from the signal, that could characterize the phenomenon searched, independently from the frequency of its occurring.

For instance, a standard set of features was proposed in the first challenge for detection and classification of acoustic scenes and events (DCASE 2013) [126]. These LLDs can be divided into temporal (energy, zero-crossing rate), spectral (spectral roll-off, flux, entropy, variance, aperiodicity bands energy, etc.) and MFCC, in addition to time-frequency

Table 2 DCASE 2013 challenge standard feature set for acoustic scene classification (ASC) and sound event detection (SED) tasks [126]

Type	Low-level descriptor	# of features
Energy	MFCC0	4
Harmonic	Fundamental frequency (F0)	4
	Audio harmonicity	4
Perceptual	TL-Sone (Total loudness in Bark scale)	32
Temporal	Autocorrelation coefficient	13
	Zero-crossing rate	4
	log-attack time	1
	Temporal centroid	1
Spectral	Audio spectrum roll-off	4
	Audio spectrum spread	4
	Audio spectrum centroid	4
	Audio spectrum flatness	16
	MFCC1-12	96

features, extracted from the wavelet analysis, such as PWP coefficients. Table 2 shows the complete list of these standard features. Each LLD is represented by a set of statistical functionals, e.g., mean, variance, skewness and kurtosis.

4.2 Feature extraction methods

Another way to compute features consists in using data-driven feature extraction methods. Different types of input are used, such as raw audio, spectrogram images or low-level features like MFCC. The output is a latent representation of the signal. A popular approach is to train a specialized type of neural network, e.g., an autoencoder (cf. Table 3).

4.3 Rationale for feature extraction

In this paper, we focus on feature extraction as a mapping that allows transforming the input, i.e.the audio signal/frame, into a vector that can be the input to a machine learning-based anomaly detection model, such as DNN, CNN, one-class SVM, etc. Feature extraction is an alternative way to hand-crafted/engineered feature computing, since it allows discovering latent knowledge. In [25, 149], feature extraction methods are thoroughly explained using either clustering-based feature subset selection or feature evaluation and selection, respectively. There are different ways to derive features from a signal, namely hand-craft feature computing (either as low-level or high-levels descriptors (cf. Table 2)) and feature extraction, using either simple/deep/ variational autoencoders, or from spectrogram image-based CNN (cf. Fig. 2). The main difference between the hand-crafted features and the extracted ones lies in two main aspects, as follows:

Hand-crafted features: They are calculated using explicit formulas, as they correspond to some properties of the audio signal, such as fundamental frequency, energy, noise level, etc. Therefore, they can be easily interpreted, and/or analyzed to provide a clear view about the relationship of each property and outlierness. For instance, a peak of energy in a narrow frequency band may indicate the presence of an abrupt event. In audio and speech processing, such a type of features has long been used in event recognition, mainly as input to machine learning-based classifiers or anomaly detection models.

Hand-crafted features thoroughly exploit experts' knowledge, and are generally very specific and well fitted to the specific application for which they have been designed. On the other hand, and for the same reasons, known features may not perform as well on different tasks, and new features are very difficult, time-consuming and overall expensive to obtain.

Feature extraction: It is relatively new, and intends to replace hand-crafted features, by embedding the data into an appropriate feature space learnt from the data themselves. Actually, the fast development of deep learning made it possible to reduce the raw audio or the spectrogram image into a vector of values that are collected at a hidden layer (e.g., the code layer for autoencoders) or at the output layers (e.g., for a CNN), respectively.

It has been demonstrated in several works that using such a vector as input to the anomaly detection system provides outstanding results, much better than those obtained by classical features. However, this increased performance is obtained at the expense of interpretability, as we do not have an a-priori knowledge about their physical meaning. In fact, all what we know about such collected features is that they are obtained during the training process, and therefore we believe they may represent well the signal. This is confirmed by the fact that using these features in an inverse transform, e.g., the decoder network, allows

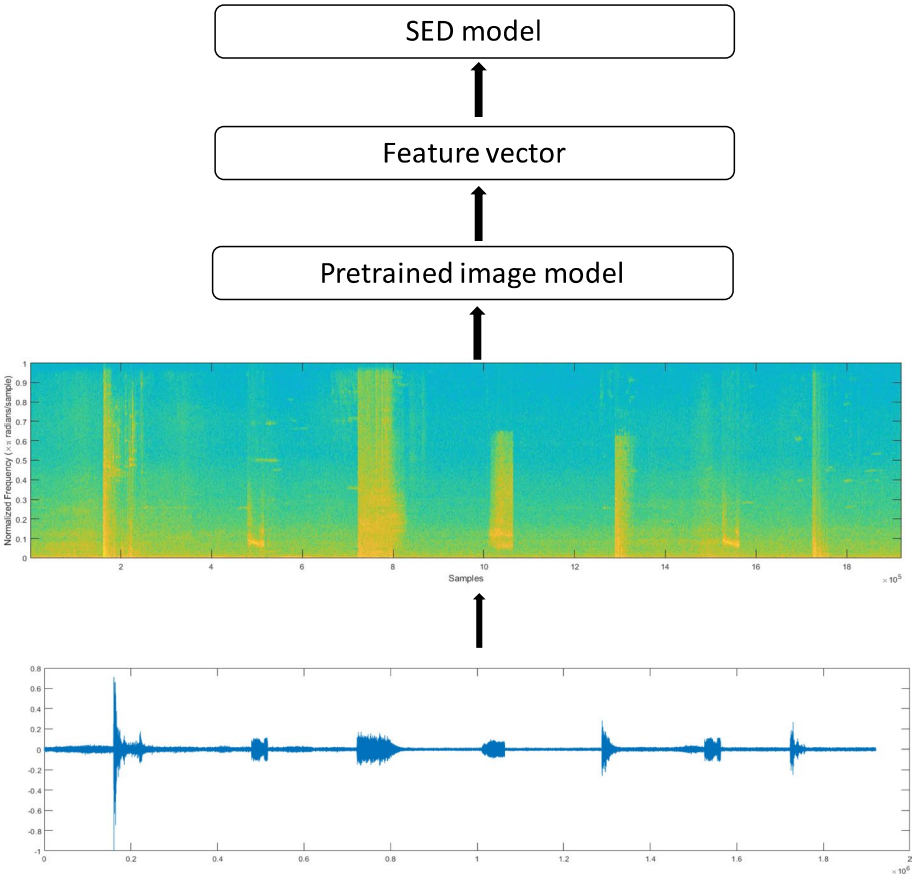

Fig. 2 Feature extraction by spectrogram-image transfer learning for anomalous SED [83]

reconstructing the signal, with some error. If this error is minimum, then we can assume that the features collected at the code layer provide a good representation of the signal, even if we do not really know to what they correspond. Therefore, these features can be used for other tasks such as classification or anomaly detection.

4.3.1 Feature extraction by autoencoders

The autoencoder is a neural network whose objective function approximates the identity map. It is an unsupervised learning technique, commonly used to extract features from unlabeled data. The mean square difference between the given input and the obtained output is minimized; then, the value of a hidden layer is used as an encoded representation of the input.

Simple and deep autoencoders A simple autoencoder has only one hidden layer. It is therefore parameterized by weights ($w \in \mathbb{R}^{mxn}$, $\tilde{w} \in \mathbb{R}^{nxm}$) and biases ($b, \tilde{b} \in \mathbb{R}^{m}$), as follows:

Table 3 Main feature extraction methods and techniques for anomalous SED

Method	Technique	Features	Example references
Feature embedding	Deep autoencoders	Output of the code layer	(Perez-Castanos et al., 2020) [97]
	Variational autoencoders	Mean and variance at the code layer	(Koizumi et al., 2017) [61]
Spectrogram image processing	Region-based CNN	log-Mel energy vector	(Muller et al., 2020) [83]
Raw signal processing	Single and multi-channel signal processing	log-Mel energy	(Adavanne & Virtanen, 2017) [4]
	Feature reconstruction from sub-sampled signal using LSTM		(Kawaguchi et al., 2018) [56]

$$\begin{cases} h = f(wx + b), \\ \tilde{x} = \tilde{f}(h\tilde{w} + \tilde{b}), \end{cases} \tag{1}$$

where $x = (x_1, x_2, \dots, x_m) \in \mathbb{R}^m$, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m) \in \mathbb{R}^m$ and $h = (h_1, h_2, \dots, h_n) \in \mathbb{R}^n$ are respectively the inputs, the outputs and the hidden layer code, and f, $\tilde{f}$ are non linear activation functions, such as the sigmoid function, $f(z) = \frac{1}{1+e^{-z}}$ [85].

It can be shown that the encoding obtained from a simple *linear* autoencoder, i.e. with $f(z) = \tilde{f}(z) = z$, spans the n principal components of the data space, recovering therefore the same embedding as PCA of order n. In this sense we may state that an autoencoder is a nonlinear generalization of PCA. Regularisation can also be added to encourage sparsity or reduce noise sensitivity [137]

A deep autoencoder can be split into two parts: (a) the encoder, from the input layer to the middle layer, and (b) the decoder, from the middle layer to the output layer. The encoded features are obtained at the output of the encoder layer, whereas the reconstructed input is recuperated at the output layer. Hence, to reduce the dimension of the input space, the encoder layer should have a lower dimension than the input layer. The encoder layer provides a useful transformation of input features, that allows, first discovering hidden structure in the input features, and secondly generating new features through the non-linear transformation of the input features by the activation functions of the hidden layers.

Perez-Castanos et al. [97] have utilized autoencoders to extract features from a Gammatone audio representation in an unsupervised way. This approach has been recently presented in DCASE 2020 challenge Task 2 for industrial monitoring.

Variational autoencoder (VAE) This is also a reconstruction network, as it learns a compressed representation of the input to reconstruct the output. However, the encoder layer of VAE stores the parameters of a probability distribution, e.g., mean and variance, representing the input in a latent space. Then, the decoder uses the probability distribution to generate an approximated reconstruction of the input data. The main issue in VAE is how to choose the parametric probability distribution. Given a feature vector X, VAE aims to find the probability of X with respect to its representation Z:

$$P(\mathrm{X}) = \int P(X|Z)P(Z)dZ. \tag{2}$$

To find $P(X|Z)$ and $P(Z)$, the VAE tries to infer $P(Z)$ using the a-priori distribution $P(Z|X)$, which is determined by variational inference by minimizing the loss given by

$$\log P(X) = -\{||X - \hat{X}||_2 + \mathrm{KL}(Q(Z|X)||P(Z))\}, \tag{3}$$

where $||.||_2$ is the L^2 norm and KL is the Kullback-Leibler divergence, given by:

$$KL(A||B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx. \tag{4}$$

Hence, the goal of VAE is to train the encoder output $Q(Z|X)$ such that the divergence between $Q(Z|X)$ and $P(Z|X)$ is minimized. For instance, if $P(Z)$ is a Gaussian distribution, the encoder generates the mean and the variance, that will be used to generate $P(Z|X)$. Then, the decoder layer reconstructs the approximation of X using (2), [66].

In the work of Koizumi et al. [61], the optimization of an acoustic extractor for anomalous sound detection based on Neyman-Pearson lemma is proposed. The acoustic feature

extractor is optimized to extract a set of acoustic features using a variational autoencoder maximizing the true positive rate (TPR) under a given false positive rate (FPR).

4.3.2 Feature extraction based on spectrogram image processing

One-dimensional convolutional neural networks are applied by Lim et al. [69] at each input time-frequency frame to extract spectral features. A more elaborated approach is proposed by Kao et al. [54], where a region-based CNN (RCNN) is developed for SED. The overall approach will be described in Section 6. Every 30 seconds, an audio clip is admitted as input to extract high level features. For each 46-ms frame (with 50% overlap), 64-dimensional log filterbank energies are calculated and aggregated to generate the input spectrogram. The process is exemplified in Fig. 3.

More recently, Muller et al. [83], substituted the classical solution of using autoencoders by utilizing image transfer learning, to extract features from the Mel-spectrogram (cf. Fig. 2). Hence, a d-dimensional feature vector is computed using a feature extractor $f : \mathbb{R}^{TxF} \rightarrow \mathbb{R}^d$ for each audio sample x_i; T, F being the time dimension and the number of frequency bins, respectively. First, the Mel-spectrogram is computed for each audio signal in the training set, using 64 Mel-bands and a Hann window of length 1024 with 256 hop size. Then 64x64 Mel-spectrogram patches ($\approx$ 1 sec) are extracted in a sliding window and converted to RGB images. Afterwards, the feature vector is extracted for each patch using neural network models pretrained on ImageNet [33], such as AlexNet [65], ResNet [45], and SqueezeNet [50].

4.3.3 Feature extraction based on signal processing methods

Two types of features are extracted for SED, i.e.single-channel and binaural features. First, single-channel features consist in log Mel-band energy (*mbe*), that have already been used for SED in [2, 3, 17, 95]. *mbe* features are extracted in a 40-ms Hamming window. Then a 40-channel Mel-log filterbank is applied in the frequency range of [0, 22.5 KHz], so that a single 40-coefficient vector is extracted for each frame. Secondly, binaural features are also *mbe* features, however extracted from multi-channel audio. Hence, for an N-channel audio signal, Nx40 outputs are extracted for each frame. Another type of such features are the multi-resolution binaural features, where these features are extracted for each channel using different window sizes. For instance, in [128], three different window lengths, i.e.1024, 4096 and 16384, are utilized to extract (40x3x2) features from each stereo audio frame.

Adavanne & Virtanen [4] achieved low-level spatial feature extraction from multi-channel audio for SED. Convolutional RNN are extended to handle many types of these multi-channel features by learning for each type separately. The main finding is that the network learns sound events in multichannel audio better from separate layers of features than from a stacked vector of concatenated features. The proposed spatial features outperform monaural features when used by the same network in terms of F1-score, when tested on TUT dataset [2].

In [56], the main purpose was extracting features for different anomalous SED from sub-sampled audio signal. To achieve that, a feature reconstruction model based on LSTM network is proposed. The main advantage consists in reconstructing an approximation of the feature vector of the original signal from the sub-sampled signal.

4.4 Feature selection for *anomalous* SED

In addition to feature extraction, feature selection is useful in *anomalous* SED for: (i) reducing the computation time, (ii) improving prediction performance, and (iii) better understanding the role of hand-crafted/extracted features. Therefore, its application to anomaly detection, in particular for audio data is quite necessary.

4.4.1 Feature selection methods

There are several methods of feature selection methods, that can be split into three main families: Filter methods They use variable ranking techniques to select features by ordering. In fact, filtering means that features are selected before any classification is undertaken. Filtering methods can be based on thresholding or ranking. Basic filter methods used for evaluating feature relevance are, e.g. Pearson correlation coefficient, mutual information and Kullback-Leibler (KL) divergence. Also, more elaborated filter methods are based on information gain (IG) ratio and Chi-square. IG ratio is a measure that weighs a feature from a high-dimension feature space; whereas the Chi-square measure assesses two types of statistical measures: a test of independence and a test of goodness of fit. The test of independence allows estimating how much a class label is independent of a feature, whereas the test of goodness of fit describes how well the model, based on the selected features, fits the set of observations [109]. Wrapper methods In these methods, the predictor is used as a black box and the predictor performance as the objective function to evaluate the variable subset. An optimal subset of features is searched heuristically by using a search algorithm, that aims to maximise the objective function, i.e. the classification performance. Amongst these search algorithms, classification trees (CT) are used even though they may lead to an exponential number of searched subsets. Other computationally lighter search algorithms are e.g. genetic algorithms (GA) and particle swarm optimization (PSO). Embedded methods The aim of this type of methods is to further reduce the computation time required for reclassifying the different subsets of features, as done by wrapper methods. To do so, the feature selection is incorporated/embedded as part of the training process. For instance, a greedy search algorithm is proposed in [22] to evaluate the features subsets initially selected by MI. A further improvement is given in [22] where the MI is estimated using Parzen window method.

4.4.2 Feature selection techniques for *anomalous* SED

The main motivation for using feature selection in *anomalous* SED can be summarised in the following rationale [16]: (a) The problem of misclassification of a decision rule does not increase as the number of features increases, as long as the class-conditional densities are completely known; (b) In general no non-exhaustive sequential feature selection procedure can be generated to produce the optimal subset. Following both principles, a set of feature selection techniques has been particularly applied to the problem of *anomalous* SED:

Sequential feature selection (SFS) It is a wrapper feature selection technique, that supports both forward or backward approaches. In the forward, respectively the backward, strategy, the number of features starts from 0, respectively from N (total number of features), to increase, respectively decrease, in function of the performance of the objective function, which is the same as the performance of the classifier.

mRMR feature selection This is a filter method that selects the features that maximise the MI between the selected features and the class labels. Each feature subset is evaluated using the same objective function used in wrapper methods. The features subsets are nested, so that $S_1 \subseteq S_2 \subseteq \dots S_{M-1} \subseteq S_M$ where M is the number of feature subsets.

PCA feature selection: Principal component analysis (PCA) can be considered as a feature extraction and a feature selection method at the same time. In fact, PCA is a linear transformation defined as $Y = X \times H$, where X and Y are the matrix of original features and of extracted features, respectively, and H is the matrix of eigenvalues. Thus, PCA returns the eigen-vectors that have the highest eigen-values, so that it selects the features which linear transformation is the largest. Therefore it can be considered as a feature extraction method as features are transformed, and a feature selection method as only the highest ones are kept.

The aforementioned three methods were applied to *anomalous* SED with valuable performance for different sound types and different tasks (cf. Table 4).

5 Evaluation metrics

The evaluation metrics for SED are somehow standardized in the framework of DCASE challenge events. In fact, since the first DCASE challenge, in 2013, a set of metrics was proposed. These metrics can be categorized as frame-based, event-based and class-wise event-based [79, 126]. It should be emphasized that along this survey work, no metrics specifically tailored for anomalous SED have been encountered in the literature. For instance, all DCASE challenge tasks about anomalous SED use the same metrics proposed for the other general purpose SED tasks (cf. Table 5).

5.1 Evaluation metrics for supervised sound event detection

5.1.1 Frame-based metrics

These metrics are calculated at each frame, of a fixed duration, and then averaged on the total number of frames in the audio signal. The classical frame-based metrics are precision (P), recall (R) and F1-score, defined by (5)

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\ \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\ \mathrm{F1} = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}, \tag{5}$$

where TP, TN, FP and FN are respectively: The number of true positives, i.e. anomalous events detected as anomalous, true negatives, i.e. normal events detected as normal, false positives, i.e. normal events detected as anomalous, and false negatives, i.e. anomalous events detected as normal. Hence, P gives the rate of correctly estimated samples among the predicted ones, whereas R yields their rate among the ground truth ones. F1 is the geometric mean of P and R. F1 is more significant than overall accuracy (Acc), defined by (6), as it shows whether a high accuracy hides a low value of P or R

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{6}$$

Table 4 Main feature selection techniques and applications for anomalous SED

Application	Technique	Features	Example references
Gunshot, glass break, traffic noise detection	(Joint) Mutual information, mRMR	MPEG-7 features	(Kiktova et al. 2014) [97]
Meeting room noise, music, speech, speech over music, speech over noise	SFS (backward), mRMR, PCA	Frequency-filtered log-filterbank energy vector	(Butko et al., 2011) [16]
Speech activity detection	PCA with Bayesian accuracy evaluation	MFCC and log-Mel energy	(Zhuang et al. , 2008) [152]
Acoustic event detection	Adaboost-based feature selection	MFCC	(Zhuang et al. 2010) [151]
Audio tagging	Feature selection at the output of each gated linear unit block	Region-based CRNN learned from spectrogram images	(Yan et al. 2010) [146]

Table 5 Main evaluation metrics used for anomalous SED with the references that recommended their use for each type of learning

Type of learning	Evaluation metric	Formula	Recommending references
Supervised learning	Precision (P), Recall (R), F1-score (F1)	(5)	(Mesaros et al., 2016) [79]
	Accuracy (Acc)	(6)	(Mesaros et al., 2016) [79]
	Audio event error rate (AEER)	(7)	(Stowell et al., 2015) [126]
Unsupervised learning	Area-under-ROC curve (AUC)	(8)	(Koizumi et al., 2020) [59]
	Partial area under-the-ROC curve (p-AUC)	(9)	(Koizumi et al., 2020) [59]

In addition, the audio event error rate (AEER) is calculated for SED in an analog way to speech recognition as in (7)

$$\text{AEER} = \frac{\text{D} + \text{I} + \text{S}}{\text{N}}, \tag{7}$$

where N is the number of all events to detect, D is the number of deletions (missing events), I is the number of insertions (added events) and S is the number of substitutions, defined as $\text{S} = min\{\text{D}, \text{I}\}$ [126].

5.1.2 Event-based metrics

For this type of evaluation, the onset and onset-offset times are taken into consideration. In onset-based evaluation, an event is considered as correctly detected if its onset time tolerance is less than 100 ms. For onset-offset-based evaluation, the onset tolerance is also set to 100ms, whereas the offset tolerance is calculated as 50% of the event duration. Hence a duplicated event is counted as a false alarm. Then, P, R, F1 and AEER are calculated for event-based evaluation types.

5.1.3 Class-wise event-based evaluation

This type of evaluation is useful to ensure that the metrics are not biased by repetitive events. P, R, F1 and AEER are calculated for each class of events and then averaged on the number of event classes.

It is worth noting that the aforementioned metrics have been adopted in DCASE 2013, 2016 and 2017 challenges as standard metrics [79, 126].

5.2 Evaluation metrics for unsupervised sound event detection

In DCASE 2020 challenge, another type of evaluation metrics was added to evaluate Task 2, i.e.unsupervised anomalous SED for machine condition monitoring [59]. It consists in AUC (Area Under the ROC (Receiver Operating Characteristic) Curve) and p-AUC (partial-AUC), defined by (8) and (9), respectively:

$$\text{AUC} = \frac{1}{N_- N_+} \sum_{i=1}^{N_-} \sum_{i=j}^{N_+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \tag{8}$$

$$p\text{-AUC} = \frac{1}{\lfloor pN_- \rfloor N_+} \sum_{i=1}^{\lfloor pN_- \rfloor} \sum_{i=j}^{N_+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \tag{9}$$

where $\mathcal{H}(x) = 1$ if $x > 0$ and 0 otherwise, $\{x_i^-\}_{i=1}^{N_-}$ and $\{x_j^+\}_{j=1}^{N_+}$ are the normal and anomalous test samples, respectively, sorted in descending order of anomaly scores, N_- and N_+ are the number of total normal and anomalous samples respectively. p-AUC is calculated as an AUC for a low false positive rate (p) set in the range of [0,1]. The introduction of p-AUC is useful to check whether anomalous SED is frequently signaling false alarms. Therefore, it is important to fix low false positive rate while trying to increase the true positive rate [59].

6 Methods and models

Classically, SED has been achieved using generative methods, such as Gaussian mixture models (GMM) and Hidden Markov models (HMM), where the contextual and temporal cues of the signal are taken into account. However, to improve over these methods, other approa- ches such as discriminative one-class support vector machines (OC-SVM) and deep neural networks (DNN) have been employed. In the following, for each type of modeling, i.e.generative, discriminative or learning-based, an overview of the general method is presented, before detailing its application to anomalous SED.

6.1 Generative modeling

6.1.1 Overview

Gaussian mixture model (GMM) is a linear combination of parametric (Gaussian) distributions. GMM clustering identifies q single models that represent the largest possible set of audio events. First, a GMM with diagonal covariance matrix is constructed for audio samples labeled as normal. Then, for each pair of the normal set, the distance between their Gaussian distributions is calculated using Kullback-Leibler (KL) divergence. Finally, the model with the minimum distance is selected as the one representing the normal class. The KL divergence is theoretically calculated using (4); however, due to the lack of a closed-form solution for GMM, it is approximated as

$$KL(A||B) \simeq \frac{1}{N} \sum_{n=1}^{N} \log \frac{p_B(x_n)}{p_A(x_n)}, \tag{10}$$

if the set of samples $\{x_n\}_{n=1,2,\ldots,N}$ is large enough [87].

6.1.2 Generative models for anomalous SED

Probabilistic anomaly detection for audio surveillance Ntalampiras et al. [87] utilized three generative methods for probabilistic novelty detection for acoustic surveillance under

real-world conditions, namely a universal GMM model, a universal HMM model and a GMM clustering model. In addition, a maximum a posteriori adaptation model (MAP) is used to update the parameters of the Gaussian components [110].

Context-dependent GMM-HMM Heittola et al. [48] proposed a context-dependent SED. This approach comprises two stages, an automatic context recognition stage and a SED stage. Contexts are modeled by GMM, whereas sound events are modeled using a 3-state left-to-right HMM.

6.2 Discriminative modeling

6.2.1 Overview

One-class support vector machine (OC-SVM) is a variant of the SVM algorithms, which finds a linear separation between two classes in feature space [121]. Generally speaking, OC-SVM is a SVM, typically using the Gaussian kernel, which divides the input space into normal data and outliers. The training is performed on normal data. For each sample, if the decision function is positive, then the sample is called normal, otherwise it is an outlier. A detailed description of the OC-SVM problem formulation and algorithm is presented in [120].

6.2.2 Discriminative models for anomalous SED

In the work of Aurino et al. [9], an OC-SVM model is developed to detect burst-like anomalous sound events, such as gunshots, broken glasses and screams. The features are extracted from time and frequency representation of the audio signal and then fed into the OC-SVM classifier.

The problem of high-dimensionality and large scale anomaly detection is addressed by Erfani et al. [36] using OC-SVM and deep learning. High dimensionality is usually a problem in audio, due to the typical high dimensional representations. The proposed solution relies on robustness in anomaly detection for high dimensional spaces using an unsupervised feature extractor and a robust anomaly detector. In fact, the classical OC-SVM anomaly detector is effective at producing decision surfaces from well-behaved feature vectors, but it is proved to be less efficient at modelling variations in large and high-dimensional datasets. Therefore, unsupervised deep belief networks (DBN) are used to learn robust features to be used by OC-SVM for anomaly detection. Two variants of the OC-SVM are proposed, including support vector data description (SVDD) and plane-based OC-SVM (PSVM). The main difference between both variants is that SVDD essentially finds the smallest possible hypersphere around the majority of the training samples, excluding the points defined as anomalies, whereas PSVM tries to find a hyperplane separating best the data from the origin.

OC-SVM are also used in ensemble-architecture to model anomalous SED. For instance, Foggia et al. [37] proposed a two-layer approach based on low-level audio feature extraction, and high level bag-of-words approach to classify events into short and sustained ones. Finally an ensemble SVM is used for event classification. Also, an ensemble OC-SVM parallel to an MLP network is used by Rovetta et al. [114] to calculate the resulting anomaly score for audio events. In their approach, The OC-SVM yields a primary anomaly score, whereas the MLP

probability output indicates the event class score. The multiplication result of both scores is thresholded to indicate whether the event classified by the MLP is actually an outlier.

6.3 Supervised learning methods and models

anomalous SED can be approached as a classification problem if the training set is fully labeled. Therefore, different methods and models have been developed using labeled datasets (cf. Table 1). In particular deep learning techniques have been extensively investigated, such as recurrent and convolutional neural networks and multitask learning. Several supervised learning methods and models have been developed for anomalous SED, in particular in the following DCASE challenges: 2016 Task 3 and 2017 Task 2, i.e. Rare SED [138, 139] , 2017 Task 3 (real-life SED) [138] and 2019 Task 4 (SED in domestic environments) [74].

6.3.1 Convolutional recurrent neural networks modeling

Overview The particularity of this method is that, while most previous SED works generate predictions at frame level first, and then use post-processing to predict the onset/offset timestamps of events from a probability sequence, the proposed method generates predictions at event level directly, and can be trained end-to-end with a multitask loss, which optimizes the classification and localization of audio events simultaneously. The end-to-end region-based convolutional recurrent neural network (R-CRNN) method for anomalous SED proceeds as follows: First a R-CRNN is applied to extract frame-level features from a 64-dimensional log filterbank energy spectrogram, as described in Section 4.3.2. Then a region proposal network (RPN) takes anchor intervals with fixed sizes to refine them at each frame, and then outputs k interval proposal. The cost function of the RPN is [54]

$$\mathcal{L}(\{p_i\},\{t_i\}) = \sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \lambda \sum_i p_i^* \mathcal{L}_{\text{reg}}(t_i, t_i^*), \tag{11}$$

where i is the index of anchor intervals, p_i and p_i^* are the predicted and the ground-truth probabilities of containing target events for anchor i respectively, $\mathcal{L}_{\text{cls}}$ is the cross-entropy cost function for binary classification. For the regression term, $\mathcal{L}_{\text{reg}}$ is the regression cost function, t_i and t_i^* are the predicted and the ground-truth coordinate vectors of event intervals, and λ is a tradeoff coefficient to balance the classification error and the regression

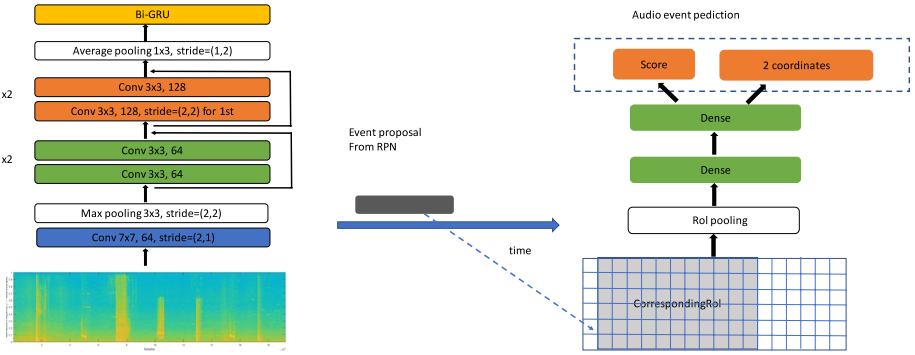

Fig. 3 End-to-end CRNN model for anomalous SED [54]

error, so that the multitask cost function optimizes binary classification and temporal localization simultaneously. Finally the SED classifier takes the event proposals generated by the RPN as input to generate audio event predictions, as shown in Fig. 3. Different variants of this methods were developed, including 1D-CRNN and R-CRNN, yielding better results than baseline methods, i.e.DNN and CNN, when tested on DCASE 2017 challenge dataset for SED.

Recurrent and convolutional neural networks models for anomalous SED A hierarchic and multi-scaled approach based on MLP-CNN for rare SED was proposed by Vesperini et al. [135] in the framework of DCASE 2017 challenge Task 2 (rare SED). This hierarchic approach comprises two stages: First, an MLP network to classify audio frames, then a CNN network whose role is to refine classification by operating at multiple resolutions and discarding blocks containing background events that have been misclassified by the MLP as rare events.

Lim et al. [69] developed a 1D-ConvNet architecture that is applied at each input time-frequency frame to extract spectral features. Then an RNN-LSTM network is used for classification thanks to its ability to incorporate the dependencies of the extracted features.

In the proposal of Dang et al. [30] at DCASE 2017 challenge Task 2, different architectures are tried out for rare SED. The first model is a CNN applied to log Mel-filterbank spectrogram, the second one is also a CNN applied to a feature set composed of MFCC and log Mel-filterbank spectrogram-extracted features, whereas the third model is a convolutional RNN (CRNN) applied to multiscale MFCC.

To cope with anomalous event scarcity in the training set, He et al. [46] developed a dilated-gated CNN (DG-CNN) to improve the detection accuracy and computational efficiency. The loss function includes a discriminative penalty term to reduce insertion errors:

$$\begin{aligned}\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[\omega_p y^{(i)}\log(\hat{y}^{(i)}) + \\ (1-y^{(i)})(y^{(i)})^2\log(1-\hat{y}^{(i)})],\end{aligned} \tag{12}$$

where $y^{(i)}$ and $\hat{y}^{(i)}$ are the actual and the predicted labels, respectively, and ω_p is a weight for positive samples only.

6.3.2 Weighted and multitask learning

Weighted loss functions are also used to counterbalance the anomalous data scarcity. multitask learning is an implicit regularization method that is expected to improve the generalization ability of a network [69].

Overview Phan et al. [102] proposed a combined weighted and multitask loss function. The weighted loss tackles the common issue of imbalanced data in background vs. foreground classification, whereas the multitask loss enables the network to simultaneously model the class distribution and the temporal structures of the target events.

i) Weighted loss for foreground/background classification: A general observation in SED shows that frames labeled as background noise are much more abundant than foreground frames. This makes the classifier biased towards background samples. The typical cross-entropy loss used for audio event classification is given by (13):

$$E(\theta) = -\frac{1}{N}\sum_{1}^{N} y_n \log(\hat{y}_n(x_n,\theta)) + \frac{\lambda}{2}||\theta||_2^2, \tag{13}$$

where θ denotes the network's parameters (weights and biases), λ the weight of the regularization term and $\hat{y}_n(x_n, \theta)$ the probability obtained for a feature vector x. To balance this loss function towards the foreground samples, a weighted loss is proposed as in (14):

$$E_w(\theta) = -\frac{1}{N}(\lambda_{fg}\sum_{n=1}^{N}\mathbb{I}_{fg}(x_n)y_n\log(\hat{y}_n(x_n,\theta)) + \lambda_{bg}\sum_{n=1}^{N}\mathbb{I}_{bg}(x_n)y_n\log(\hat{y}_n(x_n,\theta)) + \frac{\lambda}{2}||\theta||_2^2, \tag{14}$$

where $\mathbb{I}_{fg}$ and $\mathbb{I}_{bg}$ are foreground and background flag functions returning 1 if x_n is foreground and 0 if background, respectively, λ_{fg} and λ_{bg} are penalization weights for false negative errors for foreground and background samples, respectively.

ii) Multitask loss for event classification: This process aims to jointly model event classification and the temporal onset and offset distance from the center frame. Therefore a multitask loss function is tailored so that the output layer provides both the probability of the event class $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_C)$ where C is the number of event classes, and the estimated vector $\hat{d} = (\hat{d}_{\text{on}}, \hat{d}_{\text{off}})$ of the distances from the onset and the offset frames to the center frame, respectively. It is important pointing out that the class probability is returned by a softmax activation layer, whereas the distance vector is given by a sigmoid activation. Hence, the multitask loss is calculated by (15)

$$E_{mt} = \lambda_{\text{class}}E_{\text{class}}(\theta) + \lambda_{\text{dist}}E_{\text{dist}}(\theta) + \lambda_{\text{conf}}E_{\text{conf}}(\theta) + \frac{\lambda}{2}||\theta||_2^2, \tag{15}$$

where

$$E_{\text{class}}(\theta) = -\frac{1}{N}\sum_{n=1}^{N} y_n\log(\hat{y}_n(x_n,\theta)),$$

$$E_{\text{dist}}(\theta) = -\frac{1}{N}\sum_{n=1}^{N} y_n||d - \hat{d}_n(x_n,\theta)||_2^2,$$

and

$$E_{\text{conf}}(\theta) = -\frac{1}{N}\sum_{n=1}^{N} y_n||y_n - \hat{y}_n\frac{I(d_n,\hat{d}(x_n,\theta))}{U(d_n,\hat{d}(x_n,\theta))}||_2^2.$$

$E_{\text{class}}(\theta)$, $E_{\text{dist}}(\theta)$ and $E_{\text{conf}}(\theta)$ are the class, the distance and the confidence losses, respectively; whereas $I(d,\hat{d}) = min(d_{\text{on}},\hat{d}_{\text{on}}) + min(d_{\text{off}},\hat{d}_{\text{off}})$ and $U(d,\hat{d}) = max(d_{\text{on}},\hat{d}_{\text{on}}) + max(d_{\text{off}},\hat{d}_{\text{off}})$ return the intersection and the union of the target and the predicted event boundaries, respectively. Hence, the confidence loss penalizes both classification and distance estimation errors.

This method was tested on DCASE 2017 challenge Task 2 (rare SED) using two different DNN and CNN architectures. The yielding results were much better than the baseline system in terms of AEER and F1-score.

Multitask learning models for anomalous SED Xia et al. [141] proposed a multitask learning classification scheme for SED to cope with the problem of ignoring the frame position within the audio events. Therefore, a joint learning based multitask learning

system is built, where the first task is to detect the acoustic event type and the second task is to predict the frame position information.

In the work of Phan et al. [100], a multitask and multilabel framework based on convolutional RNN (CRNN) is proposed to unify the detection of isolated and overlapping audio events. The network jointly determines first whether and secondly when an event of a certain category occurs, by estimating the onset and the offset positions at recurrent time step.

Another model developed by Phan et al. [101] is based on a CNN-DNN architecture coupled with a novel weighted and multitask loss function and phase-aware signal enhancement. The proposed approach is characterized by the following aspects: (i) the loss functions are tailored for event detection in audio streams, (ii) the weighted loss is designed to tackle the common issue of imbalanced data in background/foreground classification, (iii) the multitask loss enables the network to simultaneously model the class distribution and the temporal structure of the target events.

The innovative idea presented by Imoto et al. [51] is to leverage multitask learning for SED and ASC in order to improve the performance of SED. In fact, both tasks are related since most of anomalous sound events occur in particular acoustic scenes. Therefore, exploiting the knowledge about ASC may be helpful to identify anomalous events.

6.4 Semi-supervised anomalous SED method and models

Generally, anomaly detection in raw audio methods suffer from the lack of anomalous samples in the training set. In most cases, training is made using only normal data. In particular for time series, another level of complexity is added by the contextual nature of anomalies. Therefore, some semi-supervised learning methods and models have been recently proposed. In fact, semi-supervised learning has been proved to be quite efficient in related problems, such as speech recognition , as reported in [111]. However, it has been noted that with such a type of learning, the quantity of unlabeled data should be at least 10 times that of labeled samples to obtain the same level of performance as supervised learning, as reported in [150].

6.4.1 Random forests semi-supervised model

One of the first semi-supervised models for audio event classification was proposed by [150]. It leverages low-level descriptors, such as those listed in Table 2, as input features to train random forests on labeled and unlabeled data. This choice is motivated by the ability of random forests to provide good generalization, especially for a high-dimensional feature space. In fact, each tree is modeled to fulfil a feature selection based on feature ranking through implicit information gain. Besides, feature sub-spaces are assigned randomly to the trees. A thorough description of random forests and their use in semi-supervised learning can be found in [29]. Training has been achieved by re-sampling the labeled samples and allocating them a higher weight, than unlabeled data, and by iterating the semi-supervised learning process. Hence, such a strategy succeeded to improve the performance of

semi-supervised learning in terms of F1-score, in comparison to the baseline supervised classifier [150].

6.4.2 Semi-supervised teacher-student model

In [70], a teacher-student model is implemented for weakly-labeled semi-supervised SED. The purpose of such a guided-learning model is to leverage the teacher model, initially tailored for audio tagging, to predict the time boundaries of target events. The learning process is of type end-to-end, where both labeled and unlabeled data are presented as input, to update the parameter sets, θ, θ', of both student and teacher models, respectively, through the following loss function:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_T + \mathcal{L}_{unsupervised} + a\mathcal{L}'_{unsupervised} \tag{16}$$

where $\mathcal{L}_T$ and $\mathcal{L}_S$ are the loss function of the teacher and the student models, respectively, $\mathcal{L}_{unsupervised}$ and $\mathcal{L}'_{unsupervised}$ are the loss functions calculated on the unlabeled data, respectively as

$$\mathcal{L}_{unsupervised} = J(\Phi(T'_\theta(x)), S_\theta(x)) \tag{17}$$

$$\mathcal{L}'_{unsupervised} = J(T'_\theta(x), \Phi(S_\theta(x)) \tag{18}$$

where $S_\theta(x)$ and $T'_\theta(x)$ are the frame-level predicted probabilities of the student and the teacher models, respectively, Φ is the clip-level prediction probability, J is the cross-entropy loss function, x are the input data and a is a regularisation term.

The application of such a guided-learning model in DCASE'2018-Task4 (weakly-labeled semi-supervised SED) using labeled and unlabeled data gave a better performance, in terms of event-based F1-score, than both the baseline and the teacher models [70].

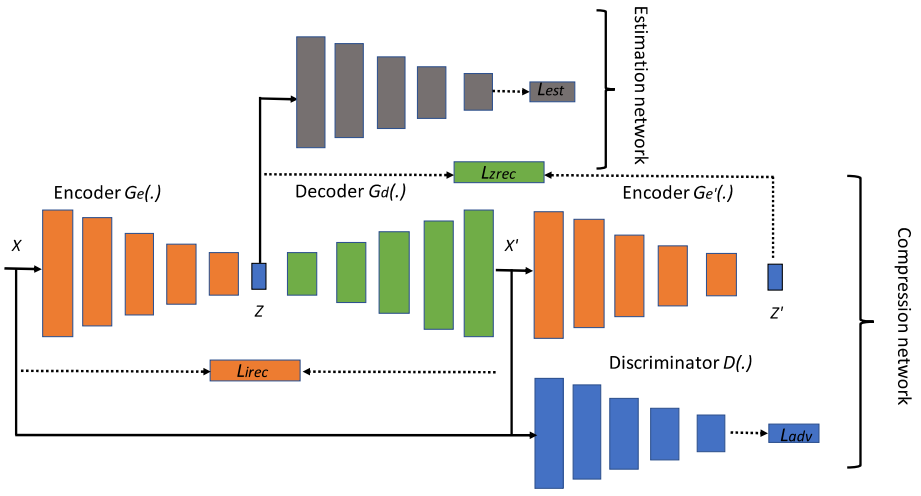

Fig. 4 GAN-based model architecture for anomalous SED [23]

6.4.3 Generative adversial network (GAN) modeling

Overview The anomaly detection GAN-based network architecture proposed by Chen et al. [23] is composed of a compression GAN and a GMM-parameter estimation network (cf. Fig. 4).

The GAN-based compression network includes a discriminator network and a generative autoencoder with an auxiliary encoder. The generator G aims to reconstruct the input spectrogram images, whereas the discriminator D tries to discard the "fake" images from the original ones. Both networks are competing, therefore the adversial loss $\mathcal{L}_{avd}$ is calculated by (19):

$$\mathcal{L}_{adv} = \min_{G} \max_{D} (E_{x\sim p(x)}[\log(D(x)] + E_{x\sim p(x))}[\log(1 - D(G(x))]). \tag{19}$$

The image reconstruction loss $\mathcal{L}_{irec}$ is calculated as the distance between the pixel-based representation of the original and the reconstructed images (cf. (20))

$$\mathcal{L}_{irec} = E_{x\sim p(x)} ||x - G(x)||_1. \tag{20}$$

For the auxiliary encoder, a latent representation loss $\mathcal{L}_{zrec}$ is calculated as the distance between the latent features of the input image $G_e(x)$ from the generator, and the encoded latent features of the image generated from the auxiliary encoder $G_{e'}(x')$ (cf. (21))

$$\mathcal{L}_{zrec} = \mathbb{E}_{x\sim p(x)} ||G_e(x) - G_{e'}(x')||_2. \tag{21}$$

The estimation network is used to estimate the GMM density parameters instead of using the classical expectation maximization (EM) parameter re-estimation approach. The estimation network is implemented as a multi-layer network with a softmax output function, so that the mixture-component membership is predicted as a K-dimensional vector $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_K)$, where K is the number of Gaussian distributions and $\hat{\gamma}_k$ is the probability that the input sample belongs to the k^{th} distribution. The estimation loss $\mathcal{L}_{est}$ is given by (22)

$$\mathcal{L}_{est} = \lambda_1 \sum_{i=1}^{N} E(z_i) + \lambda_2 \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{1}{\hat{\Sigma}_{jj}^k}, \tag{22}$$

where $E(z_i)$ is the sum of the energy function of a sample input defined by (23)

$$E(z) = -\log\left(\sum_{k=1}^{K} \hat{\Phi}_k \frac{exp(-\frac{1}{2}(z - \hat{\mu}_k)^T \hat{\Sigma}_{k^{-1}}(z - \hat{\mu}_k))}{\sqrt{2\pi\hat{\Sigma}_k}}\right), \tag{23}$$

where $\hat{\phi}_k$, $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are the weight, the mean and the covariance matrix of the k^{th} mixture component, respectively. The second term in (22) is for regularisation, used to avoid the singularity problem in GMM. λ_1 and λ_2 are meta-parameters in the estimation network.

Finally, the overall loss is calculated as the weighted sum of all losses (cf. (24))

$$\mathcal{L} = w_{\text{irec}}\mathcal{L}_{\text{irec}} + w_{\text{adv}}\mathcal{L}_{\text{adv}} + w_{\text{zrec}}\mathcal{L}_{\text{zrec}} + w_{\text{est}}\mathcal{L}_{\text{est}}, \tag{24}$$

where w_{irec}, w_{adv}, w_{zrec} and w_{est} are the weights corresponding to each network, respectively.

Generative adversial network models for anomalous SED In the work of Chen et al. [23], a novel Gaussian mixture generative adversial network (GM-GAN) is proposed under semi-supervised framework, where the underlying structure of training data is not only captured in spectrogram reconstruction space, but can also be further restricted in the space of latent representation in a discriminant manner. This method was applied to detect anomalous events using DCASE 2017 challenge dataset for Task 2 (rare SED). The benchmarking with other generative methods such as convolutional autoencoders (CAE) and WaveNet [92] yielded better results in terms of AUC parameter.

6.4.4 Few shot learning

The method proposed by Koizumi et al.(a) [60], called SNIPER (few-Shot learNIng with ensured true-PositivE Rate), aims to reconstruct normal and overlooked audio events by training the model only using few shots of anomalous events. Therefore, a cascaded anomaly score is defined as the aggregation of anomaly scores of unknown anomlaous sounds and the similarity of K recorded anomalous sounds calculated by a specific anomaly detector. Performance of anomaly detection in sounds can be measured by TPR and FPR, i.e.true positive rate and false positive rate, respectively, so that the training algorithm is optimized to maximize TPR and to minimize FPR. TPR and FPR are defined by (25) and (26), respectively

$$\mathrm{TPR}(\theta_k, \phi) = \int \mathcal{H}(x, \phi) p(x|y \neq 0) dx, \tag{25}$$

$$\mathrm{FPR}(\theta_k, \phi) = \int \mathcal{H}(x, \phi) p(x|y = 0) dx, \tag{26}$$

where

$$\mathcal{H}(x, \phi) = \begin{cases} 0 \text{ if } \mathcal{A}(x_t, \theta_A) < \phi (\text{normal}) \\ 1 \text{ if } \mathcal{A}(x_t, \theta_A) \geq \phi (\text{anomalous}) \end{cases} \tag{27}$$

and

$$\mathcal{A}(x_t, \theta_A) = -\ln p(x_t|y = 0, \theta_A) \tag{28}$$

where θ_A is the set of parameters of the normal model, ϕ is a predefined threshold;

Few-shot with metric learning In the proposal of Shimada et al. [122], the problem of few-shot learning for sound event recognition is revisited. The challenge is how to perform few-shot learning using not only chunks of sounds for training, but real audio containing background noise and other events. The proposed solution consists in a metric learning with background noise for the few-shot detection. For so doing, the main recommendations are : (i) Introducing background noise as an independent detection class, (ii) implementing a suitable loss function that emphasizes this class, (iii) choosing a corresponding sampling strategy that assists training, (iv) providing a feature space where the event classes and the background noise class are sufficiently separated.

Attention network for one-shot learning In continuation to their work presented in [60], Koizumi et al. have recently proposed a similarity function for one-shot anomaly detection called SPIDERNet: SPecific anomaly IDentifiER Network

[63]. The goal of this novel method is to update anomalous SED by training often one overlooked anomalous sample. A previous solution consists in using memory-based one-shot learning. However, this method detects only short anomalous sounds such as collision sounds because its similarity function is based on a naive MSE error between the input and the memorized spectrogram. The proposed approach proceeds by detecting various anomalous sounds using only one shot samples, a VAE-based feature extractor for measuring similarity in embedded space, and an attention mechanism for absorbing time-frequency stretching. To train the SPIDERNet, J normal samples are selected from the normal sounds, whereas only one sample is available for each type of anomalous sounds. To increase the number of anomalous sounds, data augmentation is achieved by a random circular shift in the wave-domain. Then training is performed to minimize a cost function based on a similarity score between the input and the memorized spectrograms. Benchmarking with other similarity score methods, such as autoencoders and naive MSE show a better performance of the proposed method, when tested on machine condition monitoring audio datasets, like ToyADMOS [62] and MIMII [107], in terms of AUC score.

6.5 Unsupervised learning models for anomalous SED

Unsupervised learning provides a large choice of methods to deal with unlabeled data. For instance, clustering has been applied for anomaly detection in different domains [1]. In particular, for anomalous SED, reconstruction-based and metric learning-based methods have been an alternative to supervised/ semisupervised learning, to get around the issue of annotation.

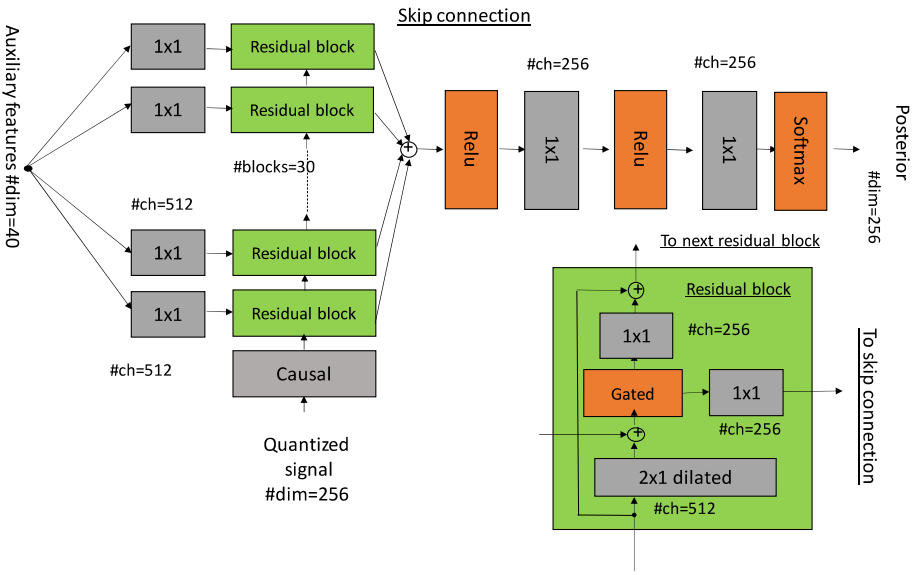

Fig. 5 The WaveNet network architecture for anomalous SED [42]

6.5.1 Autoregressive neural networks models (Wavenet) for anomalous SED

Overview In the work of Komatsu et al. [42], anomalous SED is achieved using an autoregressive neural model, namely WaveNet [92], to model and reconstruct the waveform. The use of WaveNet is motivated by its ability to model detailed structures, such as phase information, so that it is expected to detect anomalous sound events with more accuracy than other conventional reconstruction-based anomaly detection techniques. WaveNet is an end-to-end acoustic modeling tool based on convolutional neural networks (cf. Fig. 5). WaveNet approximates the conditional probability of a waveform given its auxiliary features by canceling the effect of past samples of a finite length (cf. (29) and (30)):

$$p(\mathbf{x}|\mathbf{h}) = \prod_{n=1}^{N} p(x_n|x_1, x_2, \dots, x_{n-1}, \mathbf{h}), \tag{29}$$

$$\text{WaveNet}(\mathbf{x}|\mathbf{h}) \simeq \prod_{n=1}^{N} p(x_n|x_{n-R-1}, x_{n-R}, \dots, x_{n-1}, \mathbf{h}), \tag{30}$$

where R is the number of past samples, h is the vector of auxiliary features and x_1, x_2, ⋯, x_{n-1} are the $(N-1)$ past samples of the waveform ($\mathbf{x}$). WaveNet has been successfully used to model acoustic waveforms, in particular for raw audio reconstruction and speech synthesis [93]. It is optimized through backpropagation using a cross-entropy objective function given by (31)

$$E(\Theta) = -\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c}\log(\hat{y}_{t,c}), \tag{31}$$

where Θ, is the network's parameter set (weights and biases), $y = (y_{t,1}, \dots, y_{t,C})$ is the one-hot vector of the target quantized signal and $\hat{y}_t = (\hat{y}_{t,1}, \dots, \hat{y}_{t,C})$ is the posterior of the amplitude class, t and c are the indices of the waveform samples and their amplitude class, respectively; T and C represent the number of waveform samples and the number of amplitude classes, respectively. anomalous SED using WaveNet is evaluated through the uncertainly of the prediction, quantified as an entropy of the posterior and calculated by (32):

$$e_\theta = -\sum_{c=1}^{C} \hat{y}_{t,c}\log_2 \hat{y}_{t,c}. \tag{32}$$

Finally, the posterior entropy is compared to a dynamic threshold calculated by (33)

$$\theta = \mu + \beta\sigma, \tag{33}$$

where θ is the threshold value, μ and σ are the mean and the standard deviation of the entropy sequence, respectively, and β is a heuristic hyperparameter.

Also, Rushe & Mac Namee [116] built a WaveNet model with two stacks of 10 layers of causal dilated convolutions. In each stack, residual and skip connections are used along with an exponentially growing dilation rate. Data is normalized to have zero-mean and unit-variance. Each sample is generated using a softmax distribution of a quantized integer range of 256 values. In each layer, 512 filters are used for skip connections and 256 filters for residual ones. Training is performed by minimizing the cross-entropy loss, whereas the reconstruction error is measured using MSE error. The developed WaveNet model, along

with a baseline convolutional autoencoder model were applied on the dataset of rare SED proposed at DCASE 2017 challenge Task 2. The results evaluated with AUC score show a clear advantage of using WaveNet for all types of indoor and outdoor sounds.

6.5.2 Autoencoder and metric learning models

In the proposal of Wei et al. [140] at DCASE 2020 challenge Task 2, a reconstruction autoencoder is used to calculate the anomaly score through metric learning. Therefore, different types of autoencoders are tested, such as deep autoencoders, variational autoencoders, etc.

In the system proposed by Giri et al. [41], the addressed problem is how to detect previously unseen anomalous sound events when the training set contains only normal data. The classical approaches like GMM and DNN seem unable to seize all the aspects of the problem. Therefore, a new approach is proposed, using a novel neural density estimation technique based on the group-masked autoencoder, that estimates the density of an audio time series by taking into account the intra frame statistics of the signal. In comparison to the baseline autoencoder approach, this method has shown better results.

Deep autoencoder models Deep autoencoder-based reconstruction is used by Marchi et al. [75] for acoustic novelty detection. In this work, the auditory spectral features of the next short-term frame are predicted from the previous one by means of LSTM denoising autoencoders. The error between the input and the reconstructed frame is used as activation signal to detect anomalous events.

More recently, Purohit et al. [106] trained a deep autoencoder on GMM distributions with hyperparameter optimization to detect anomaly in acoustic signals. The method baptized DAGMM-HO (deep autoencoder based on GMM with hyperparameter optimization) applies the conventional autoencoder-GMM to the audio domain. Optmization is obtained by reducing dimensions, and statistical modeling is thought to improve anomaly detection performance. In addition, the hyperparameter sensitivity problem of the conventional DAGMM is solved by performing hyperparameters optimization based on gap statistics and the cumulative Eigenvalues.

Complementary set variational autoencoder The problem of unseen anomalies is addressed by Kawachi et al. [55]. Actually, a drawback of conventional supervised learning consists in its unsuitability to detect unseen anomalies, ı.e. anomalous examples that have not been encountered in the learning phase. Therefore, unsupervised learning, and in particular deep autoencoders are chosen to resolve this issue. However, unsupervised autoencoders suffer from the reverse problem, as they are able to detect unseen anomalies, but not capable to detect seen anomalies, even if some data are available. Therefore, this work [55] presents an ano- maly detector able to find both seen and unseen anomalies in acoustic data. The proposed approach consists of a novel probabilistic representation of anomalies to solve the raised problem. Hence, normal and anomaly distributions are defined using the analogy between a set and a complementary set. Then, these distributions are applied to an unsupervised VAE to turn it out to a supervised method.

6.5.3 Density estimation models for anomalous SED

Density estimation aims at learning the underlying probability density of an independent and identically distributed set of samples. Therefore, neural networks represent a classical approach to learn such a density, especially for high-dimensional data [84]. In particular,

for anomaly detection, density estimation provides an intuitive tool, since normal and anomalous samples should be clustered into high and low density regions, respectively [64].

Unsupervised DNN-based normalizing flows In a recent work of Koizumi et al. [144], the addressed problem is how to solve unsupervised anomaly in general, with a particular focus on anomalous SED. A previous solution consists in density estimation, whereas the proposed solution relies on DNN-based density estimators (normalizing flows). However the problem is how to adapt such density estimators to the change in normal data distribution. The proposed solution consists in designing a new DNN-based density estimator that can be easily adapted to the change of the distribution. Hence the system is a unified model of normalizing flows and adaptive batch normalizing (AdaFlow) that enables DNN to adapt to new distributions.

Temporal trajectory mixtures The problem addressed by Chakrabarty & Elhilali [18] is how to define the "normal" behaviour of a crowd in an environment, such as an airport, a train station or a sport field as normal or not. The problem stems from the difficulty to define "normal" behavior of a crowd. The proposed solution consists of successfully capturing the heterogeneous nature of audio events in an acoustic environment, to be used as a reference against which anomalous behavior can be detected in continuous audio recordings. The proposed approach is based on a methodology for representing sound classes with a hierarchical network of convolutional features and mixtures of temporal trajectories (MTT). The framework comprises unsupervised and supervised learning to provide a robust scheme for detection of abnormal sound events in a subway station. It uses as input the time-frequency representation of the audio signal to be processed by three main components: (a) An acoustic modeling block using restricted Boltzmann machine (RBM), (b) a dynamic modeling block using MTT, and (c) an abnormal sound event detection block using a likelihood measure to yield the decision about sound abnormality [18].

KL-divergence based models In [15], an unsupervised framework is proposed to resolve the problem of SED for unlabeled data. A previous work presented in [15] relies on GMM training on purely normal data and estimation of the KL divergence between the input and the output. This approach is improved in [15] by trimming the quarter of the most divergent Gaussian distributions from the mixture model, in order to enhance the KL divergence performance (Table 6).

7 Applications

Since a few years, the fast development of methods and models for anomalous SED has allowed its extension to a wide range of applications, spanning different fields. Besides, the inclusion of anomalous SED as a research topic in related events such as DCASE and PhysioNet challenges, has contributed to developing novel models for real life problems (cf. Table 7).

7.1 Audio surveillance

Surveillance systems are getting more and more multimodal. Therefore, a variety of anomalous SED models have been designed for several audio surveillance applications.

Table 6 Methods and models for anomalous SED

Category	Method	Model	Authors and references
Generative methods	GMM	GMM-based probabilistic novelty detection	(Ntalampiras et al., 2011) [87]
		Context-dependent GMM	(Heittola et al., 2013) [48]
	HMM	HMM-based probabilistic novelty detection	(Ntalampiras et al., 2011) [87]
Discriminative methods	OC-SVM	OC-SVM for anomalous SED	(Aurino et al., 2014) [9]
		Ensemble-based OC-SVM for audio surveillance	(Foggia et al., 2015) [37]
		Support vector data description and plane-based OC-SVM for high dimensionality and large scale anomaly detection	(Erfani et al., 2016) [36]
		Support vector data description and plane-based OC-SVM	(Erfani et al., 2016) [36]
		Ensemble-based MLP-OC-SVM outlier detection model	(Rovetta et al., 2020) [114]
Supervised learning	RNN and CNN	RNN-BLSTM classifier	(Li & Li., 2017) [68]
		MLP-CNN hierarchic multi-scaled model	(Vesperini et al., 2017) [135]
		1D-ConvNets model	(Lim et al., 2017) [69]
		Spectrogram-image-based CNN classifier	(Dang et al., 2017) [30]
		Region-based convolutional RNN (R-CRNN) model	(Kao et al., 2018) [54]
		Dilated-gated R-CRNN with a discriminative penalty loss function (cf. (14))	(Phan et al., 2018) [102]
		RNN-LSTM classifier with different pooling functions	(Kao et al., 2020) [53]
	Weighted and/or multitask learning	Acoustic event type and frame position detection using multitask learning	(Phan et al., 2017) [101]
		DNN-CNN model with a weighted and multitask loss function (cf. (13)	(Phan et al., 2018) [102]
		Multitask multilabel model based on CRNN	(Phan et al., 2019) [100]
		Multitask learning for joint SED and ASC	(Imoto et al., 2020) [51]
		Multitask learning for joint event and frame position detection	(Xia et al., 2020) [141]
Semi-supervised learning	RF	Random forests (RF) for audio event classification	(Zhang & Schuller, 2012) [150]
	GAN	Gaussian-mixture GAN model	(Chen et al., 2020) [23]
	Few-shot learning	Few-shot learning model with ensured true positive rate	(Koizumi et al., 2017(a)) [60]

Table 6 (continued)

Category	Method	Model	Authors and references
		Few-shot model with metric learning	(Shimada et al., 2020) [122]
		Attention network for one-shot learning	(Koizumi et al., 2020) [63]
	TS-GL	Teacher-student guided learning model (TS-GL) for weakly-labeled SED	(Lin et al., 2020)) [70]
Unsupervised learning	Autoregressive neural networks	Denoising autoencoder	(Marchi et al., 2017) [75]
		Complementary set variational autoencoder	(Kawachi et al., 2018) [55]
		WaveNet-based anomalous SED	(Komatsu et al., 2018) [42]
		Group-masked autoencoder	(Giri et al., 2020) [41]
		Deep autoencoder based on GMM with hyperparameters	(Purohit et al., 2020) [106]
		Denoising autoencoder	(Marchi et al., 2020) [75]
	Density estimation	KL-divergence optimization for GMM-based density estimation	(Borges & Meyer, 2008) [15]
		Temporal trajectory mixtures	(Chakrabarty & Elhilali, 2016) [18]
		DNN-based normalizing flows	(Koizumi et al., 2017(b)) [61]

Table 7 Applications of anomalous SED

Field	Application	Specific application	Authors and references
Audio surveillance	Traffic Audio surveillance	Acoustic hazard detection for pedestrians	Lee & Rakotonirainy, 2011 [67]
Traffic Audio surveillance		UBM model for acoustic surveillance of urban traffic	(Ntalampiras, 2014) [86]
		anomalous SED on roads	(Foggia et al., 2015) [37]
		Sound-based car crash detection	Sammarco & Detyniecki, 2018 [118]
		Sound-based car crash detection	(Sammarco et al., 2018) [118]
	General-purpose audio surveillance	Scream and gun shot detection and localization	(Valenzise et al., 2007) [133]
		Probabilistic novelty detection for acoustic surveillance	(Ntalampiras et al., 2011) [87]
		DCASE 2013 challenge Task 2: SED	(Stowell et al., 2015) [126]
		Enhancing audio surveillance	(Colangelo et al., 2017) [26]
		Study of the effect of temporal dimension of the input signal	(Rossi et al., 2017) [113]
	Rare SED	DCASE 2016 challenge Task 3: Real-life SED	(Virtanen et al., 2016) [139]
		DCASE 2017 challenge Task 2: Rare SED & Task 3: Real-life SED	(Virtanen et al., 2017) [138]
		DCASE 2019 challenge Task 4: SED in domestic environments	(Mandel et al., 2019) [74]
Industrial equipment monitoring	Rare SED in IoT	Anomaly detection in 3D printers	(Uematsu et al., 2017) [132]
		Predictive maintenance IoT platform based on sound stream analysis in edges	(Yamato et al., 2017) [145]
		Server-client online acoustic anomaly detection system for industrial equipment	(Ahn et al., 2019) [5]
		Abnormality detection in SMD machine sound using Autoencoder residual error	(Dong et al., 2018) [89]
		Use of IOT audio stream for rare SED	(Janjua et al., 2019) [52]
		Anomaly detection in millig tools using acoustic signals and GAN	(Cooper et al., 2020) [27]
		Unsupervised anomalous SED for machine condition monitoring	(Koizumi et al., 2020) [59]

Table 7 (continued)

Field	Application	Specific application	Authors and references
		Acoustic anomaly detection in additive manufacturing (3D printers)	(Becker et al., 2020) [13]
	Machine condition monitoring	Anomaly detection of PC's cooling fans	(Ono et al., 2013) [91]
		Audio anomaly detection on rotating machine	(Prego et al., 2016) [105]
		Acoustic signal processing for anomaly detection in HVAC machine room environments	(Kao et al., 2018) [54]
		Acoustic anomaly detection for different datasets	(Duman et al., 2019) [35]
	Multimodal robotics	Multimodal (force, sound, kinematic signals) event detection in human environment by a robot	(Park et al., 2019) [96]
Speech and music processing	Speech	Change detection for audio segmentation	(Omar et al., 2005) [90]
		Nominal world model applied for speech activity change detection for audio segmentation	(Borges & Meyer, 2008) [15]
	Music	Change detection for music analysis	(Vallim & Mello, 2015) [134]
		Detection of corrupted or distorted samples in music datasets	(Lu et al., 2016) [73]
Healthcare	Sound-based diagnosis	Anomalous respiration sound detection	(Ye et al., 2012) [147]
			(Pham et al., 2021) [99]
		Anomalous PCG signal detection	(Zahibi et al., 2016) [148]
			(Dissanayacke et al., 2020) [34]

7.1.1 Urban traffic monitoring

For instance, anomalous SED has been used in urban traffic monitoring using several approaches. Foggia et al. [37] developed a two-layer model, using bag-of-words based feature extraction and ensemble OC-SVM classifier to detect hazardous sounds on the road, in particular car crash and tire skidding.

The concern of the work of Lee et al. [67] was pedestrian's safety by sending an alarm message online. The method is developed using a set of statistical techniques for feature mining and a three-component heuristic. Ntalampiras [86] proposed a non-intrusive, passive monitoring framework based on audio modality. Thus, a universal background model (UBM) is trained with the goal to recognize and detect a large number of audio events encountered in urban areas. Sammarco & Detyniecki [118] proposed a system named CrashZam for car cash detection using in-car installed microphone signal. Two models were developed, the first using spectral features extracted from the raw audio signal, whereas the second is based on learning features from spectrogram images.

7.1.2 Novelty detection in general-purpose audio surveillance

The approach proposed by Valenzise et al. [133] leverages audio data extracted from video surveillance systems to detect and localize alarming audio events such as screams and gunshots. Each event is identified using a GMM classifier trained on temporal, spectral and perceptual audio-extracted features.

In the work of Colangelo et al. [26], audio events (glass breaking, gunshots and screams) mixed with different types of background noise (car passing by, crowd, etc.) at different SNR levels, are classified by an anomalous SED model using hierarchical RNN trained on spectrogram image-extracted features. Probabilistic anomaly detection is used by Ntalampiras et al. [87] to detect abnormal and life threatening situations, where GMM and Kullbalck-Leibler (KL) divergence are utilized to train and detect anomalous events.

7.1.3 Anomalous sound event detection in DCASE challenge events

The goal of the first DCASE challenge, organized in 2013 [126], was to develop general sound recognition in any environment. The challenge comprised detection and classification tasks of acoustic scenes and events.

Among The DCASE 2016 challenge [77], only Task 3, SED in real-life audio, can be considered as rare SED, and thus belonging to the realm of anomaly detection. In real-life SED, the events of interest are arbitrarily rare and classes are often unbalanced. Besides, the main particularity of this task is temporal annotation, i.e.onset and offset timing.

In DCASE 2017 challenge [76], in addition to Task 3, real-life SED, also Task 2, rare SED, is an anomaly detection problem. The audio data were generated by mixing background acoustic scenes and rare target sound events.

In DCASE 2019 challenge, Task 4, SED in domestic environments [131], can be considered as anomalous SED, since the goal was detecting the type and the timing of any occurring event using weakly-/unlabeled data from different real-life SED datasets (cf. Table 1).

7.2 Industrial equipment monitoring

In the industrial realm, early anomaly detection is an important and cost effective maintenance tool. Therefore, a special attention has been paid to industrial equipment monitoring among the emergent applications of anomalous SED.

7.2.1 Rare sound event detection on IoT

In the work of Janjua et al. [52], IoT data streams are used for rare SED. The method is based on unsupervised learning, where data are first segmented into micro clusters, which are in their turn agglomerated in macro clusters. Uematsu et al. [132] utilized anomaly detection through IoT to collect information from diverse sensors such as microphones for machine condition monitoring. The approach relies on normality modeling based on DNN applied to acoustic features extracted from the spectrogram of the recorded sound. Another IoT-based anomalous SED application is proposed by Yamoto et al. [145]. In this work, a maintenance IoT-based platform able to analyze sound datastreams in edges is designed to analyze only anomaly data in cloud and to order maintenance online.

7.2.2 Machine condition monitoring

In DCASE 2020 challenge [59], Task 2 is launched to present models able to identify anomalous sounds issued from a target machine . The challenge consists in detecting unknown anomalous sounds under the condition that only normal sounds are provided in the training set. Therefore, unsupervised learning was proposed as a modeling technique.

Ahn et al. [5] proposed a system of acoustic anomaly detection for machine condition monitoring. The system was designed to capture acoustic signals and to classify them using machine learning. The system also includes a server for sound management and model training, a mobile client for sound capturing and real-time classification, and a workbench acting like a user interface.

Another method of anomalous SED for machine condition monitoring is proposed by Becker et al. [13], with application to additive manufacturing, in particular for 3D printing. Thus, a machine learning model is developed to detect flaws and errors of a 3D printer with varying difficulty using audio recordings from the printing machine. Acoustic features such as MFCC and Mel-filterbank energies are extracted to be trained by an LSTM-based multiclass classifier.

In the system designed by Cooper et al. [27], anomaly detection is searched in milling tools using acoustic signals and generative adversial networks (GAN). The proposed approach is based on training a GAN on only a single readily obtained class of acoustic data, and then inverting the generator to perform anomaly detection.

For the same goal, Dong et al. [89] proposed a method for SMD machine monitoring based on anomaly detection using the residual error of a reconstruction autoencoder. This unsupervised learning method tries to specify if the sound of a SMD machine is normal or anomalous based on the reconstruction of the spectrogram images performed by an autoencoder.

Another application of anomaly detection in motors with feature emphasis using only normal sounds is proposed by Ono et al. [91]. The goal of this work is to detect operating

motor anomalies from sounds without using abnormal data, so that training is made only on normal data. The proposed approach is based on calculating the distance between the feature vector and the model learned from the normal data only.

Also, the problem of audio anomaly detection in rotating machinery is addressed by Prego et al. [105]. The method is based on leveraging rotating machinery sound recordings to extract spectrogram-image-based features. Then a similarity measure is calculated between reference and degraded signals using either a 2D-cross correlation or KL divergence measure.

For acoustic anomaly detection in industrial processes such as painting, cutting and welding, Duman et al. [35] proposed an approach based on applying unsupervised convolutional autoencoders on log Mel-spectrograms. Data augmentation is used to compensate the lack of training data by superimposing industrial environment noise at different levels to the recorded audio clips. This work has been continued in [12] using sequential convolutional-LSTM autoencoders and an Euclidean distance-based reconstruction error to detect anomalous sounds.

Kao et al. [58] used anomalous sound detection to monitor equipment in commercial buildings, such as in machine rooms with HVAC system components. Then audio data is analyzed by an ensemble of machine learning algorithms to be judged as normal or abnormal. The audio data were recorded using mobile phone from machine rooms and an elevator shaft. The collected audio data is analyzed and processed by an ensemble of machine learning classifiers.

7.3 Speech and music processing

A classical domain of application of anomalous SED is speech and music processing, since speech/speaker turn change and music annotation have been among the first applications of anomalous SED. Besides, it can be used for developing further speech and music-related applications.

7.3.1 Speech analysis and recognition

One of the earliest works of anomaly/novelty detection in speech was presented by Omar et al. [90], where the problem of automatic segmentation of audio streams according to speaker identities, environmental and channel conditions, as a preprocessing step for speech/ speaker recognition and audio data mining is addressed. Therefore automatic segmentation is proposed based on a cumulative sum algorithm for automatic audio segmentation, that minimizes the missing probability of a given false alarm rate.

Later, Borges & Meyer [15] developed an unsupervised approach in the aim to perform anomaly detection for a self-diagnostic speech activity detector. The anomalous events are estimated by measuring the KL divergence (cf. (4)) between the Gaussian distribution of input features and a nominal world model.

7.3.2 Music processing

The method designed by Lu et al. [73] aims to systematically identify anomalies in music datasets. Therefore, an unsupervised model that integrates categorical regression and robust estimation techniques to infer anomalous scores in music clips is developed. The model was applied to detect corrupted, distorted or mislabeled audio samples on commonly used

features in music information retrieval. Also, Vallim et al. [134] developed a new change detection algorithm that ensures model modification corresponding to actual data changes. It is mainly intended to detect changes in music audio streams. The proposed approach is based on a new stability concept adapted for unsupervised change detection.

7.4 Healthcare

Another important field of application of anomalous SED is the study of biological sounds, such as respiration sounds and the phonocardiogram (PCG). In fact, several works have proved the efficiency of anomalous SED for early disease detection. Respiration anomaly is studied through anomalous sound detection by Ye et al. [147]. In their work, adaptive modeling of the mainstream of respiration is achieved, as well as detecting irregular patterns. For so doing, FLAC (local auto-correlation on complex Fourier values) features are analyzed through online learning, in order to adapt the respiration sound pattern CCI-PCA (candid covariance-free incremental-based PCA). More recently, in [99], respiration sound database ICBHI [112] has been analysed to classify anomalous respiration and to detect lung disease. To fulfill that, the authors used a very deep CNN network, such as VGG-7 [123] to classify log-Mel spectrogram images of the recorded respiration sounds. Furthermore, to ensure the tradeoff between model performance and complexity, a knowledge distillation model has been implemented. In such a scheme, the parameters of the best classification model (Teacher model) are leveraged to train another classifier with fewer parameters (Student model), yet having nearly the same performance.

In order to analyze heart anomalies through sound, the PCG (phonocardiogram) signal is used by Dissnayake et al. [34] and Zahibi et al. [148] in ambulatory monitoring to evaluate heart hemodynamic states and to detect a cardio-vascular disease. Both approaches are based on developing automatic classification method of anomaly detection, i.e.normal or abnormal, and quality, i.e.good or bad, of PCG recording with and without segmentation, respectively.

8 Open challenges and proposals for improvement

Despite the fast development of methods and models for anomalous SED, it still suffers from some problems, mainly related to the inherent issue of anomalous data scarcity. Besides, some criticism was expressed regarding the learning and evaluation methodologies. Therefore, some ideas and solutions were suggested (cf. Table 8).

8.1 Rare and imbalanced data

Handling imbalanced datasets, where anomalous data are a minority is still an open challenge for the topic of this study. Therefore, some works have recently addressed this problem, either using data augmentation, or a tailored learning method, where a higher weight is attributed to the least represented class. Then, classification with neural networks and application to sound event detection is treated in [7].

Table 8 Open challenges and recent proposed solutions in anomalous SED

Problem	Challenges	Proposed solution	Authors and references
Rare and imbalanced data	Data augmentation	Virtual data training using GAN	(Xia et al., 2020) [141]
		Dynamic time warping for data augmentation	(Chen et al., 2019) [24]
	Weighted learning	Input data mapping embedding to clusters	(Arora et al., 2019) [7]
		Type-2 fuzzy sets based on an inverted-weight membership function	(Rovetta et al., 2021) [115]
Learning methodology	Necessity for more realistic training and test methodology for anomalous SED	Taking into account the bias introduced by prior/background knowledge while performing training	(Baumann et al., 2020) [11]
Effect of temporal evolution on RNN	Effect of temporal evolution of the input signal on the performance of RNN for anomalous SED	Variation of the length of the input sequence and size of temporal window for feature extraction	(Rossi et al., 2017) [113]
Performance measure	Risk of unbiased performance measures in high class imbalance conditions	Adaptation of performance metric criteria to imbalance conditions	(Forman & Sholtz, 2010) [39]
Computation efficiency	Necessity to extract relevant features in a big and noisy audio dataset	Parallel computing using MapReduce programming model in Hadoop	(Mulimani & Koolagudi, 2019) [82]

8.1.1 Data augmentation

Data augmentation, i.e. simulating virtual data, either by replication or simulation, can help increasing the amount of anomalous data. For instance, [141] proposes to generate virtual training data categorically using an auxiliary GAN classifier. Then soft labels of acoustic events are calculated to represent the acoustic event localization information. In [24], a novel data augmentation method, based on dynamic time warping (DTW) [98], is proposed. It is achieved in three main steps: (i) Randomly choosing multiple instances from a same event, (ii) rescaling each instance, (iii) randomly generating the weight vector and computing the weighted DTW average. Finally the weighted average is returned as a new sample.

8.1.2 Weighted learning

To cope with the issue of data imbalance, the common approach consists in using class weights in the objective function while training. In [7], a more elaborated approach proceeds by mapping the input to clusters in an embedding space in order to balance the learning by incorporating inter-cluster and inter-class margins. The proposed approach consists in learning the embedding using a novel objective function, qualified as triple-header cross-entropy. The experimental evaluation results show that this method is more effective for SED with imbalanced data.

Recently, [115] proposed a novel type-2 fuzzy set approach for hazardous events detection from traffic audio data. Since hazardous events, i.e.car accidents, tire skidding, harsh braking, etc. are a minority in the dataset, the weights of such classes have been manipulated in the type-2 fuzzy membership function, so that the weight of each class is inversely proportional to the class samples. Then a membership degree is calculated for each event using an upper/optimistic membership component and a lower/pessimistic one. Finally, interval comparison is performed to select the event for which the membership is highest.

8.2 Learning methodology

Recently, Baumann et al. [11] suggested a reflection about the necessity of performing SED training and test methodology in a more realistic way. The outcome is a novel approach based on (i) eliminating much prior knowledge on the test data, (ii) assuming additional unknown acoustic events both in training and test data, which in practice have to be identified as background, (iii) by taking this into account while training, the robustness in real-world scenarios can be significantly increased, (iv) evaluating the advantage of multi-event classifiers over single-event ones.

8.3 Effect of temporal dimension of the input signal

The scope of the study presented by Rossi et al. [113] is to study the effect of environmental temporal evolution of the input signal on the performance of audio surveillance RNN models. The proposed approach is based on varying the length of the input sequence and the size of the time window used for feature extraction, in order to compare the temporal correlations extracted at the feature level with the one learned by a representational

structure. The obtained results show that sequential models are not necessarily the best fitted to work with temporal data, and that optimization of the temporal dimension, i.e.input sequence or window size, remains an open issue.

8.4 Performance measures

Since 2010, Forman & Scholz [39] have demonstrated that F1-score, accuracy and AUC may be biased by the method by which they are calculated, especially when calculated for cross-validation-based learning and under high class imbalance conditions, that is indeed the case of anomalous SED. Therefore, some recommendations are provided to carry out these measures in imbalanced data, using some adapted variants of F1-score and AUC, and taking care of the criteria of reducing FPR and enhancing TPR [39].

8.5 Computational efficiency

In a recent work, Mulimani & Koolagudi [82] developed a novel parallel method for extracting significant information from spectrograms using MapReduce programming model [49] for audio-based surveillance systems, in order to recognize acoustic events in surrounding environment. The relevance of this work consists in how to extract features from spectrograms in a big noisy audio dataset, which is very demanding in terms of computational time. The proposed solution consists in parallel computing using MapReduce programming model in Hadoop. This method was applied on spectrograms of event data from MIVIA database [37], showing a better computational time efficiency and a high recognition rate of critical acoustic events in different noisy conditions.

9 Conclusion

A survey of anomalous SED based on machine learning has been presented in this paper. Despite the relative novelty of this topic, the state of the art has grown in an impressive way in the last decade. Actually, since the organization of the first DCASE challenge in 2013, a large number of methods and models in various domains of application has been developed and presented in the main related events and publications.

As a final note, we would like to conclude with some reflection remarks: (a) Through this survey we focused on deep learning techniques, as they are the state of the art in anomaly detection in general, and anomalous SED in particular. However, other machine learning techniques like discriminative methods, such as OC-SVM, and generative methods like GMM and HMM have also been quite successful, and it would be potentially interesting to pursue their development. (b) The use of hand-crafted features is decreasing in favor of feature extraction/embedding methods and end-to-end modeling. Nevertheless, low-level audio descriptors have the advantage to help understanding the physical meaning of the signal parameters and their role in the acoustic dynamics. Therefore, some interaction between the two types of feature computing should be investigated. (c) Anomalous SED is very useful, and perhaps will be one of the main future technologies in public safety, therefore more concern about improving the evaluation metrics should be taken, so that they can reflect all aspects of anomalous event detection. (d) For the same purpose, i.e.public safety, anomalous SED could be in a near future embedded in everyday life appliances, to provide

vital services such as alarm messaging by mobile phones or home assistants. Therefore, the computational efficiency has to be improved to make that happen.

Acknowledgements This work has been funded by the University of Genoa in the framework of the project *Xpert*.

References

1. Abdullatif A, Masulli F, Rovetta S (2018) Clustering of nonstationary data streams: A survey of fuzzy partitional methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1258
2. Adavanne S, Parascandolo G, Pertila P, Heittola T, Virtanen T (2016) Sound event detection in multichannel audio using spatial and harmonic features. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), pp 6–10
3. Adavanne S, Pertilä P, Virtanen T (2017) Sound event detection using spatial features and convolutional recurrent neural network. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 771–775
4. Adavanne S, Virtanen T (2020) A report on sound event detection with different binaural features. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017). Munich, Germany
5. Ahn JW, Grueneberg K, Ko BJ, Lee WH, Morales E, Wang S, Wang X, Wood D (2019) Acoustic anomaly detection system: demo abstract. In: Proceedings of the 17th Conference on Embedded Networked Sensor Systems, pp 378–379
6. Alsina-Pagès RM, Orga F, Alías F, Socoró JC (2019) A wasn-based suburban dataset for anomalous noise event detection on dynamic road-traffic noise mapping. Sensors 19(11):2480
7. Arora V, Sun M, Wang C (2019) Deep embeddings for rare audio event detection with imbalanced data. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 3297–3301
8. Atrey PK, Maddage NC, Kankanhalli MS (2006) Audio based event detection for multimedia surveillance. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol 5. IEEE, pp 813–816
9. Aurino F, Folla M, Gargiulo F, Moscato V, Picariello A, Sansone C (2014) One-class svm based approach for detecting anomalous audio events. In: 2014 International Conference on Intelligent Networking and Collaborative Systems. IEEE, pp 145–151
10. Babaee E, Anuar NB, Abdul Wahab AW, Shamshirband S, Chronopoulos AT (2017) An overview of audio event detection methods from feature extraction to classification. Applied Artificial Intelligence 31(9–10):661–714
11. Baumann J, Lohrenz T, Roy A, Fingscheidt T (2020) Beyond the dcase 2017 challenge on rare sound event detection: A proposal for a more realistic training and test framework. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 611–615
12. Bayram B, Duman TB, Ince G (2021) Real time detection of acoustic anomalies in industrial processes using sequential autoencoders. Expert Systems 38(1):e12564
13. Becker P, Roth C, Roennau A, Dillmann R (2020) Acoustic anomaly detection in additive manufacturing with long short-term memory neural networks. In: 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA). IEEE, pp 921–926
14. Benetos E, Dixon S (2013) Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. The Journal of the Acoustical Society of America 133(3):1727–1741
15. Borges N, Meyer GG (2008) Unsupervised distributional anomaly detection for a self-diagnostic speech activity detector. In: 2008 42nd Annual Conference on Information Sciences and Systems. IEEE, pp 950–955
16. Butko T (2011) Feature selection for multimodal: acoustic Event detection. Universitat Politècnica de Catalunya
17. Cakır E, Parascandolo G, Heittola T, Huttunen H, Virtanen T (2017) Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(6):1291–1303

18. Chakrabarty D, Elhilali M (2016) Abnormal sound event detection using temporal trajectories mixtures. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 216–220
19. Chan T, Chin CS (2020) A comprehensive review of polyphonic sound event detection. IEEE Access 8:103339–103373
20. Chandola V, Banerjee A, Kumar V (2007) Outlier detection: A survey. ACM Computing Surveys 14:15
21. Chandrakala S, Jayalakshmi S (2019) Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. ACM Computing Surveys (CSUR) 52(3):1–34
22. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Computers & Electrical Engineering 40(1):16–28
23. Chen C, Chen P, Yang L, Mo J, Song H, Xie Y, Ma L (2020) Acoustic anomaly detection via latent regularized gaussian mixture generative adversarial networks. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). Tokyo, Japan http://dcase.community/challenge2020/index. Preprint: arxiv: 2002.01107
24. Chen Y, Jin H (2019) Rare sound event detection using deep learning and data augmentation. In: INTERSPEECH, pp 619–623
25. Chen Z, Chen Q, Zhang Y, Zhou L, Jiang J, Wu C, Huang Z (2021) Clustering-based feature subset selection with analysis on the redundancy-complementarity dimension. Computer Communications 168:65–74. https://doi.org/10.1016/j.comcom.2021.01.005
26. Colangelo F, Battisti F, Carli M, Neri A, Calabró F (2017) Enhancing audio surveillance with hierarchical recurrent neural networks. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp 1–6
27. Cooper C, Zhang J, Gao RX, Wang P, Ragai I (2020) Anomaly detection in milling tools using acoustic signals and generative adversarial networks. Procedia Manufacturing 48:372–378
28. Cotton CV, Ellis DP (2011) Spectral vs. spectro-temporal features for acoustic event detection. In: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, pp 69–72
29. Criminisi A, Shotton J (2013) Semi-supervised classification forests. In: Decision Forests for Computer Vision and Medical Image Analysis. Springer, pp 95–107
30. Dang A, Vu TH, Wang JC (2017) Deep learning for dcase2017 challenge. In: Workshop on DCASE2017 Challenge, Tech. Rep
31. Dee HM, Hogg DC (2005) On the feasibility of using a cognitive model to filter surveillance data. In: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005. IEEE, pp 34–39
32. Dekkers G, Lauwereins S, Thoen B, Adhana MW, Brouckxon H, van Waterschoot T, Vanrumste B, Verhelst M, Karsmakers P (2017) The SINS database for detection of daily activities in a home environment using an acoustic sensor network. In: Proceedings of the Detection and Classication of Acoustic Scenes and Events 2017 Workshop (DCASE2017), pp 32–36
33. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee, pp 248–255
34. Dissanayake T, Fernando T, Denman S, Ghaemmaghami H, Sridharan S, Fookes C (2021) Domain generalization in biosignal classification. IEEE Transactions on Biomedical Engineering 68(6):1978–1989. https://doi.org/10.1109/TBME.2020.3045720
35. Duman TB, Bayram B, İnce G (2019) Acoustic anomaly detection using convolutional autoencoders in industrial processes. In: International Workshop on Soft Computing Models in Industrial and Environmental Applications. Springer, pp 432–442
36. Erfani SM, Rajasegarar S, Karunasekera S, Leckie C (2016) High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recognition 58:121–134
37. Foggia P, Petkov N, Saggese A, Strisciuglio N, Vento M (2015) Audio surveillance of roads: A system for detecting anomalous sounds. IEEE transactions on intelligent transportation systems 17(1):279–288
38. Fonseca E, Pons J, Favory X, Font F, Bogdanov D, Ferraro A, Oramas S, Porter A, Serra X (2017) Freesound datasets: a platform for the creation of open audio datasets. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017). Suzhou, China, pp 486–493
39. Forman G, Scholz M (2010) Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. Acm Sigkdd Explorations Newsletter 12(1):49–57

40. Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, 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). IEEE, pp 776–780
41. Giri R, Cheng F, Helwani K, Tenneti SV, Isik U, Krishnaswamy A (2020) Group masked autoencoder based density estimator for audio anomaly detection. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020).Tokyo, Japan, pp 51–55. http://dcase.community/documents/workshop2020/proceedings/DCASE2020Workshop_Giri_66.pdf
42. Hayashi T, Komatsu T, Kondo R, Toda T, Takeda K (2018) Anomalous sound event detection based on wavenet. In: 2018 26th European Signal Processing Conference (EUSIPCO). IEEE , pp 2494–2498
43. Hayashi T, Watanabe S, Toda T, Hori T, Le Roux J, Takeda K (2017) Blstm-hmm hybrid system combined with sound activity detection network for polyphonic sound event detection. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 766–770
44. Hayashi T, Watanabe S, Toda T, Hori T, Le Roux J, Takeda K (2017) Duration-controlled lstm for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(11):2059–2070
45. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
46. He KX, Zhang WQ, Liu J, Liu Y (2019) Dilated-gated convolutional neural network with a new loss function on sound event detection. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp 1491–1495
47. Heittola T, Mesaros A, Eronen A, Virtanen T (2013) Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing 2013(1):1–13
48. Heittola T, Mesaros A, Eronen A, Virtanen T (2013) Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing 2013(1):1
49. Holmes A (2012) Hadoop in practice. Manning Publications Co
50. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2017) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings
51. Imoto K, Tonami N, Koizumi Y, Yasuda M, Yamanishi R, Yamashita Y (2020) Sound event detection by multitask learning of sound events and scenes with soft scene labels. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 621–625
52. Janjua ZH, Vecchio M, Antonini M, Antonelli F (2019) Irese: An intelligent rare-event detection system using unsupervised learning on the iot edge. Engineering Applications of Artificial Intelligence 84:41–50
53. Kao CC, Sun M, Wang W, Wang C (2020) A comparison of pooling methods on lstm models for rare acoustic event classication. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 316–320
54. Kao CC, Wang W, Sun M, Wang C (2018) R-crnn: Region-based convolutional recurrent neural network for audio event detection. Proc. Interspeech 2018:1358–1362
55. Kawachi Y, Koizumi Y, Harada N (2018) Complementary set variational autoencoder for supervised anomaly detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 2366–2370
56. Kawaguchi Y (2018) Anomaly detection based on feature reconstruction from subsampled audio signals. In: 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, pp 2524–2528
57. Kim HG, Moreau N, Sikora T (2006) MPEG-7 audio and beyond: Audio content indexing and retrieval. John Wiley & Sons
58. Ko BJ, Ortiz J, Salonidis T, Touma M, Verma D, Wang S, Wang X, Wood D (2016) Demo abstract: acoustic signal processing for anomaly detection in machine room environments. In: Proc. of ACM BuildSys
59. Koizumi Y, Kawaguchi Y, Imoto K, Nakamura T, Nikaido Y, Tanabe R, Purohit H, Suefusa K, Endo T, Yasuda M, Harada N (2020) Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). Tokyo, Japan. http://dcase.community/challenge2020/index. Preprint: arxiv: 2006.05822
60. Koizumi Y, Murata S, Harada N, Saito S, Uematsu H (2019) Sniper: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 915–919

61. Koizumi Y, Saito S, Uematsu H, Harada N (2017) Optimizing acoustic feature extractor for anomalous sound detection based on neyman-pearson lemma. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, pp 698–702
62. Koizumi Y, Saito S, Uematsu H, Harada N, Imoto K (2019) Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, pp 313–317
63. Koizumi Y, Yasuda M, Murata S, Saito S, Uematsu H, Harada N (2020) Spidernet: Attention network for one-shot anomaly detection in sounds. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 281–285
64. Kriegel HP, Kröger P, Sander J, Zimek A (2011) Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(3):231–240
65. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6):84–90
66. Latif S, Rana R, Qadir J, Epps J (2018) Variational autoencoders for learning latent representations of speech emotion: a preliminary study. In: Interspeech 2018: Proceedings, pp. 3107–3111. International Speech Communication Association (ISCA)
67. Lee J, Rakotonirainy A (2011) Acoustic hazard detection for pedestrians with obscured hearing. IEEE Transactions on Intelligent Transportation Systems 12(4):1640–1649
68. Li Y, Li X (2017) The seie-scut systems for ieee aasp challenge on dcase 2017: Deep learning techniques for audio representation and classication. In: Proc. Detection Classication Acoustic Scenes Events 2018 Workshop
69. Lim H, Park J, Han Y (2017) Rare sound event detection using 1d convolutional recurrent neural networks. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, pp 80–84
70. Lin L, Wang X, Liu H, Qian Y (2020) Guided learning for weakly-labeled semi-supervised sound event detection. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 626–630
71. Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, Castells F, Roig JM, Silva I, Johnson AE et al (2016) An open access database for the evaluation of heart sound algorithms. Physiological Measurement 37(12):2181
72. Liu Y, Tang J, Song Y, Dai L (2018) A capsule based approach for polyphonic sound event detection. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp 1853–1857
73. Lu YC, Wu CW, Lu CT, Lerch A (2016) An unsupervised approach to anomaly detection in music datasets. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp 749–752
74. Mandel M, Salamon J, Ellis DPW (2019) Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019). New York University, NY, USA
75. Marchi E, Vesperini F, Squartini S, Schuller B (2017) Deep recurrent neural network-based autoencoders for acoustic novelty detection. Computational intelligence and neuroscience 2017
76. Mesaros A, Diment A, Elizalde B, Heittola T, Vincent E, Raj B, Virtanen T (2019) Sound event detection in the dcase 2017 challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27(6):992–1006
77. Mesaros A, Heittola T, Benetos E, Foster P, Lagrange M, Virtanen T, Plumbley MD (2017) Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(2):379–393
78. Mesaros A, Heittola T, Klapuri A (2011) Latent semantic analysis in sound event detection. In: 2011 19th European Signal Processing Conference. IEEE, pp 1307–1311
79. Mesaros A, Heittola T, Virtanen T (2016) Metrics for polyphonic sound event detection. Applied Sciences 6(6):162
80. Mesaros A, Heittola T, Virtanen T (2016) Tut database for acoustic scene classification and sound event detection. In: 2016 24th European Signal Processing Conference (EUSIPCO). IEEE, pp 1128–1132
81. Morise M, Yokomori F, Ozawa K (2016) World: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE TRANSACTIONS on Information and Systems 99(7):1877–1884
82. Mulimani M, Koolagudi SG (2019) Extraction of mapreduce-based features from spectrograms for audio-based surveillance. Digital Signal Processing 87:1–9
83. Müller R, Ritz F, Illium S, Linnhoff-Popien C (2020) Acoustic anomaly detection for machine sounds based on image transfer learning. arXiv:2006.03429

84. Nachman B, Shih D (2020) Anomaly detection with density estimation. Physical Review D 101(7):075042
85. Ng A, et al (2011) Sparse autoencoder. CS294A Lecture notes 72(2011), 1–19
86. Ntalampiras S (2014) Universal background modeling for acoustic surveillance of urban traffic. Digital Signal Processing 31:69–78
87. Ntalampiras S, Potamitis I, Fakotakis N (2011) Probabilistic novelty detection for acoustic surveillance under real-world conditions. IEEE Transactions on Multimedia 13(4):713–719
88. Nunes, E.C.: Anomalous sound detection with machine learning: A systematic review. arXiv preprint arXiv:2102.07820(2021)
89. Oh DY, Yun ID (2018) Residual error based anomaly detection using auto-encoder in smd machine sound. Sensors 18(5):1308
90. Omar MK, Chaudhari U, Ramaswamy G (2005) Blind change detection for audio segmentation. In: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 1. IEEE, pp I–501
91. Ono Y, Onishi Y, Koshinaka T, Takata S, Hoshuyama O (2013) Anomaly detection of motors with feature emphasis using only normal sounds. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp 2800–2804
92. Van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: A generative model for raw audio. In: 9th ISCA Speech Synthesis Workshop, pp 125–125
93. Oord A, Li Y, Babuschkin I, Simonyan K, Vinyals O, Kavukcuoglu K, Driessche G, Lockhart E, Cobo L, Stimberg F, et al (2018) Parallel wavenet: Fast high-fidelity speech synthesis. In: International conference on machine learning. PMLR, pp 3918–3926
94. Papadaniil CD, Hadjileontiadis LJ (2013) Efficient heart sound segmentation and extraction using ensemble empirical mode decomposition and kurtosis features. IEEE journal of biomedical and health informatics 18(4):1138–1152
95. Parascandolo G, Huttunen H, Virtanen T (2016) Recurrent neural networks for polyphonic sound event detection in real life recordings. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 6440–6444
96. Park D, Kim H, Kemp CC (2019) Multimodal anomaly detection for assistive robots. Autonomous Robots 43(3):611–629
97. Perez-Castanos S, Naranjo-Alcazar J, Zuccarello P, Cobos M (2020) Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). Tokyo, Japan http://dcase.community/challenge2020/index. Preprint: arxiv: 2006.15321
98. Petitjean F, Forestier G, Webb GI, Nicholson AE, Chen Y, Keogh E (2014) Dynamic time warping averaging of time series allows faster and more accurate classification. In: 2014 IEEE international conference on data mining. IEEE, pp 470–479
99. Pham LD, Phan H, Palaniappan R, Mertins A, McLoughlin I (2021) Cnn-moe based framework for classification of respiratory anomalies and lung disease detection. IEEE Journal of Biomedical and Health Informatics
100. Phan H, Chén OY, Koch P, Pham L, McLoughlin I, Mertins A, De Vos M (2019) Unifying isolated and overlapping audio event detection with multi-label multi-task convolutional recurrent neural networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE , pp 51–55
101. Phan H, Krawczyk-Becker M, Gerkmann T, Mertins A (2017) Dnn and cnn with weighted and multi-task loss functions for audio event detection. In: Proc. DCASE 2017-Workshop Detect. Classification Acoust. Scenes Events
102. Phan H.,Krawczyk-Becker M, Gerkmann T, Mertins A (2018) Weighted and multi-task loss for rare audio event detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 336–340
103. Plinge A, Grzeszick R, Fink GA (2014) A bag-of-features approach to acoustic event detection. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 3704–3708
104. Plumbley MD, Kroos C, Bello JP, Richard G, Ellis DP, Mesaros A (2018) Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018). Tampere University of Technology. Laboratory of Signal Processing

105. Prego TDM, de Lima AA, Netto SL, da Silva EA (2016) Audio anomaly detection on rotating machinery using image signal processing. In: 2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS). IEEE, pp 207–210
106. Purohit H, Tanabe R, Endo T, Suefusa K, Nikaido Y, Kawaguchi Y (2020) Deep autoencoding gmm-based unsupervised anomaly detection in acoustic signals and its hyper-parameter optimization. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). Tokyo, Japan. http://dcase.community/challenge2020/index. Preprint: arxiv: 2009.12042
107. Purohit H, Tanabe R, Ichige K, Endo T, Nikaido Y, Suefusa K, 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), pp 209–213. Tokyo, Japan. https://doi.org/10.33682/m76f-d61
108. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2):257–286. https://doi.org/10.1109/5.18626
109. Rachburee N, Punlumjeak W (2015) A comparison of feature selection approach between greedy, ig-ratio, chi-square, and mrmr in educational mining. In: 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE). IEEE , pp 420–424
110. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. Digital signal processing 10(1–3):19–41
111. Riccardi G, Hakkani-Tur D (2005) Active learning: Theory and applications to automatic speech recognition. IEEE transactions on speech and audio processing 13(4):504–511
112. Rocha B, Filos D, Mendes L, Vogiatzis I, Perantoni E, Kaimakamis E, Natsiavas P, Oliveira A, Jácome C, Marques A, et al (2017) A respiratory sound database for the development of automated classification. In: International Conference on Biomedical and Health Informatics. Springer, pp 33–37
113. Rossi A, Montefoschi F, Rizzo A, Diligenti M, Festucci C (2017) Auto-associative recurrent neural networks and long term dependencies in novelty detection for audio surveillance applications. In: IOP Conference Series: Materials Science and Engineering
114. Rovetta S, Mnasri Z, 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). IEEE, pp 1–6
115. Rovetta S, Mnasri Z, Masulli F, Cabri A (2021) Audio surveillance of road traffic: An approahc based on interval comparison and type 2 fuzzy sets. In: The 12th Conference of the European Society for Fuzzy Logic and Technology. EUSFLAT
116. Rushe E, Mac Namee B (2019) Anomaly detection in raw audio using deep autoregressive networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE , pp 3597–3601
117. Salamon J, Jacoby C, Bello JP (2014) A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia, pp 1041–1044
118. Sammarco M, Detyniecki M (2018) Crashzam: Sound-based car crash detection. In: VEHITS, pp 27–35
119. Schmidt SE, Holst-Hansen C, Graff C, Toft E, Struijk JJ (2010) Segmentation of heart sound recordings by a duration-dependent hidden markov model. Physiological measurement 31(4):513
120. Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J (1999) Support vector method for novelty detection. Advances in neural information processing systems 12:582–588
121. Shawe-Taylor J, Cristianini N (2004) Kernel Methods for Pattern Analysis. Cambridge University Press. https://doi.org/10.1017/CBO9780511809682
122. Shimada K, Koyama Y, Inoue A (2020) Metric learning with background noise class for few-shot detection of rare sound events. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE , pp 616–620
123. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings
124. Sodemann AA, Ross MP, Borghetti BJ (2012) A review of anomaly detection in automated surveillance. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(6):1257–1272
125. Stevens SS, Volkmann J, Newman EB (1937) A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America 8(3):185–190
126. Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD (2015) Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia 17(10):1733–1746

127. Stowell D, Plumbley MD (2013) Segregating event streams and noise with a markov renewal process model. The Journal of Machine Learning Research 14(1):2213–2238
128. Su TW, Liu JY, Yang YH (2017) Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE , pp 791–795
129. Syed Z, Leeds D, Curtis D, Nesta F, Levine RA, Guttag J (2007) A framework for the analysis of acoustical cardiac signals. IEEE Transactions on Biomedical Engineering 54(4):651–662
130. Tranter SE, Reynolds DA (2006) An overview of automatic speaker diarization systems. IEEE Transactions on audio, speech, and language processing 14(5):1557–1565
131. Turpault, N., Serizel, R., Parag Shah, A., Salamon, J.: Sound event detection in domestic environments with weakly labeled data and soundscape synthesis (2019). Preprint: https://hal.inria.fr/hal-02160855
132. Uematsu H, Koizumi Y, Saito S, Nakagawa A, Harada N (2017) Anomaly detection technique in sound to detect faulty equipment. NTT Technical Review 15(8)
133. Valenzise G, Gerosa L, Tagliasacchi M, Antonacci F, Sarti A (2007) Scream and gunshot detection and localization for audio-surveillance systems. In: 2007 IEEE Conference on Advanced Video and Signal Based Surveillance. IEEE, pp 21–26
134. Vallim RM, de Mello RF (2015) Unsupervised change detection in data streams: an application in music analysis. Progress in Artificial Intelligence 4(1–2):1–10
135. Vesperini F, Droghini D, Ferretti D, Principi E, Gabrielli L, Squartini S, Piazza F (2017) A hierarchic multi-scaled approach for rare sound event detection. In: Proc. DCASE 2017-Workshop Detect. Classification Acoust. Scenes Events
136. Vincent E, Barker J, Watanabe S, Le Roux J, Nesta F, Matassoni M (2013) The second 'chime'speech separation and recognition challenge: An overview of challenge systems and outcomes. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE , pp 162–167
137. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11:3371–3408
138. Virtanen T, Mesaros A, Heittola T, Diment A, Vincent E, Benetos E, Elizalde BM (2017)Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017). Tampere University of Technology. Laboratory of Signal Processing
139. Virtanen T, Mesaros A, Heittola T, Plumbley M, Foster P, Benetos E, Lagrange M (2016)Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016). Tampere University of Technology. Department of Signal Processing
140. WEI, Q., LIU, Y.: Auto-encoder and metric-learning for anomalous sound detection task(2020). http://dcase.community/challenge2020/index. Preprint: http://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Wei_49_t2.pdf
141. Xia X, Togneri R, Sohel F, Zhao Y, Huang D (2019) Multi-task learning for acoustic event detection using event and frame position information. IEEE Transactions on Multimedia 22(3):569–578
142. Xia X, Togneri R, Sohel F, Zhao Y, Huang D (2019) A survey: neural network-based deep learning for acoustic event detection. Circuits, Systems, and Signal Processing 38(8):3433–3453
143. Xiang T, Gong S (2008) Incremental and adaptive abnormal behaviour detection. Computer Vision and Image Understanding 111(1):59–73
144. Yamaguchi M, Koizumi Y, Harada N (2019) Adaflow: Domain-adaptive density estimator with application to anomaly detection and unpaired cross-domain translation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE , pp 3647–3651
145. Yamato Y, Fukumoto Y, Kumazaki H (2017) Predictive maintenance platform with sound stream analysis in edges. Journal of Information processing 25:317–320
146. Yan J, Song Y, Guo W, Dai LR, McLoughlin I, Chen L (2019) A region based attention method for weakly supervised sound event detection and classication. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 755–759
147. Ye J, Kobayashi T, Higuchi T (2012) Smart audio sensor on anomaly respiration detection using flac features. In: 2012 IEEE Sensors Applications Symposium Proceedings. IEEE, pp 1–5
148. Zabihi M, Rad AB, Kiranyaz S, Gabbouj M, Katsaggelos AK (2016) Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In: 2016 Computing in Cardiology Conference (CinC). IEEE , pp 613–616
149. Zhang Y, Zhu R, Chen Z, Gao J, Xia D (2021) Evaluating and selecting features via information theoretic lower bounds of feature inner correlations for high-dimensional data. European Journal of Operational Research 290(1):235–247. https://doi.org/10.1016/j.ejor.2020.09.028

150. Zhang Z, Schuller B (2012) Semi-supervised learning helps in sound event classification. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 333–336
151. Zhuang X, Zhou X, Hasegawa-Johnson MA, Huang TS (2010) Real-world acoustic event detection. Pattern Recognition Letters 31(12):1543–1551
152. Zhuang X, Zhou X, Huang TS, Hasegawa-Johnson M (2008) Feature analysis and selection for acoustic event detection. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp 17–20