APSIPA Transactions on Signal and Information Processing, 2024, 13, e3
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, for non-commercial use, provided the original work is properly cited.

Overview Paper

Sound Event Detection: A Journey Through DCASE Challenge Series

Tanmay Khandelwal[1,2*], Rohan Kumar Das[1] and Eng Siong Chng[3]

[1] *Fortemedia Singapore, Singapore*
[2] *New York University, USA*
[3] *Nanyang Technological University, Singapore*

ABSTRACT

The sense of hearing is fundamental to human beings, as it allows them to perceive their surroundings. However, this simple task of recognizing different sounds in complex environments poses a challenge for machines. Sound event detection (SED) is a field that aims to automate the human auditory system's detection and recognition of sound events with their onset and offset points. Training an SED system typically requires a large labeled set, but is associated with high annotation costs and is dependent on the subjective judgments of annotators. Therefore, significant efforts have been made in this area, including the major DCASE challenge series, which brings researchers together annually to address this issue. The DCASE challenge was started in the year 2013, and it has evolved over the years to witness some significant breakthroughs in the field of SED. In this study, we delve into the methods proposed by various authors in the DCASE challenge series, providing a thorough discussion of feature extraction, machine learning techniques, and post-processing methods. We also study the results from top teams in each edition of the DCASE challenge to bring out the highlights of the best-performing SED systems and explore potential future research directions.

*Corresponding author: Tanmay Khandelwal, f20170106p@alumni.bits-pilani.ac.in.

24 July 2023; Revised 06 December 2023
ISSN 2048-7703; DOI 10.1561/116.00000051
© 2024 T. Khandelwal, R. K. Das and E. S. Chng

Keywords: sound event detection, DCASE challenge, feature extraction, post-processing, machine learning

1 Introduction

In recent years, there has been a reduction in the size of computing devices, which has led to rapid development in electronic devices to make every device fit into our lives. This has increased the interaction between humans and electronic devices. The main objective of this development is to find a way to make the device understand and recognize the human environment, and then use the environmental information to ease human lives. This ability of a system to gather information about its environment at any given time and adapt behavior accordingly is called context awareness [134]. It is further desirable to identify context implicitly rather than through the user explicitly providing contextual details.

Sound is one of the major components in understanding the physical context. The sound events act as good descriptors for an auditory scene [9]. The procedure of a machine using the ambient sound signal to construct a symbolic description of the auditory scene is known as computational auditory scene analysis (CASA) [147]. A sound event is a label that individuals may use to describe a recognizable event in the realm of sound. This type of label typically aids people in comprehending the notion and relating this event to other well-known events. Humans have an inbuilt system that aids in perceiving environmental changes and comprehending the auditory scene. Machines, on the other hand, are still incapable of providing consistent precision for this task.

CASA entails several related tasks, including source separation [66, 89], sound event detection (SED), and acoustic scene recognition [64, 132]. The task of SED deals with the automation of this built-in system to identify the temporal onset and offset of sound events and to categorize specific sound event types in a wide range of environments. The subtask in SED that categorizes different sound events without onset and offset is known as audio tagging. These automated, sound-based systems have few advantages over vision-based systems [88, 164]. First, the accuracy of the sound-based systems is not affected in dark environments. Second, sound has the quality to penetrate through obstacles. Third, some events can only be detected using sound, like an alarm sound.

There are two main subcategories of SED systems, as represented in Figure 1: monophonic and polyphonic [115]. Monophonic SED systems can only detect one sound event at a time, which is frequently the most noticeable one, regardless of how many sound events are actually occurring at any given time. Due to the fact that simultaneous sound events frequently occur in real

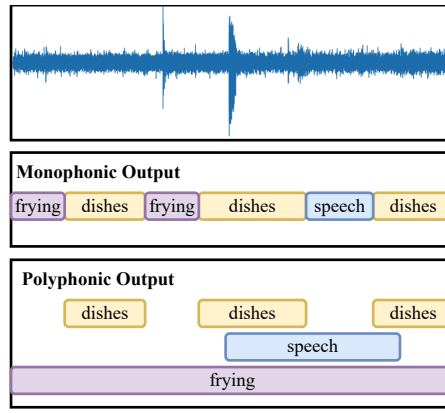

Figure 1: SED system setup, describing the monophonic and polyphonic system.

life [112], the practical application of such systems is limited by the number of audible events that can be detected. However, polyphonic SED focuses on detecting multiple simultaneous sound events that are present at any given time instance, which is more applicable to real-world applications [11, 106, 61]. However, this complicates the polyphonic system since features retrieved from the mixture may not match any of the features extracted from isolated sounds. Furthermore, it is unknown how many occurrences can be contained in a recording.

The detection and classification of acoustic scenes and events (DCASE) challenge is an annual event that focuses on the detection of polyphonic sound events. It aims to spearhead the field of SED by encouraging participants to develop innovative approaches and solutions to benchmark on a common dataset. Over the years, the challenge has sparked several advancements in this domain. It has undergone significant evolution, incorporating changes in datasets, evaluation metrics, baselines, and training methods. These modifications ensure that the challenge remains dynamic and reflects the current state-of-the-art in SED. One of the key strengths of the DCASE challenge is its ability to foster collaboration among participants. It serves as a platform where researchers and practitioners can come together to exchange ideas, share their findings, and make their systems accessible to the wider community. By encouraging collaboration, the DCASE challenge promotes knowledge sharing and accelerates progress in the field of acoustic scene and event detection.

While there are numerous works available that provide comprehensive reviews of SED [115, 16, 122] in general, none of them specifically focus on SED from the perspective of the DCASE challenge series. This work aims to address this gap by providing a dedicated exploration of the journey through

the various DCASE challenges over the past years. By focusing on the DCASE challenges, this work serves as an entry point for analyzing the systems that have been developed and the evolution of the dataset used in these challenges. It offers a unique perspective on how the DCASE challenge has progressed and transformed over time, shedding light on the advancements made in SED techniques within the context of this specific task.

These developed automatic SED systems face additional obstacles when operating in real-world conditions, including intra-class variations, pertaining to the nature of the sound, how it manifests in natural environments, and those related to data collection and annotation procedures. These additional challenges are listed below and also summarized in Figure 2.

Figure 2: Obstacles faced by automatic SED systems.

- Intra-class variability: The acoustic properties of each sound class can vary substantially, so a system must be robust to all possible examples that may be encountered for a given class. For instance, a system must be able to recognize a variety of car horns to be able to detect the class 'car horn'.
- Occurrence in nature: The sounds of the natural environment are polyphonic by their very nature; as a result, there is the possibility that multiple sounds will be active all at once. For instance, a recording from a cafeteria may include the sounds of people conversing, a coffee machine, a blender, and non-destructive sounds of dishes. As a result, the system

must be able to recognize the acoustic characteristics of each unique sound event in the mixture.

- Data collection: The performance of the SED system may be hampered by parameters in the data collection process. Consider factors such as the separation between the recording equipment and the sound source, the kind of recording device used, and the setting in which the recording was made, as they could all affect the background noise.

- Labeling ambiguity: The situation is made more difficult by the absence of a well-established ontology for universally describing different types of sounds. Moreover, the labels of annotations are highly variable because they are based on the subjective judgment of the annotator. A human annotator could, for instance, label a baby's continuous cries as a single sound event or label each cry separately.

- Lack of structure: Some audio signals such as speech and music contain a certain amount of structure that can be used to draw out useful sound representations from the signal. For example, speech can be broken down into phonemes, and music into notes. However, finding a consistent definition of subdivision for sound occurrences is challenging for SED, making the SED task tough.

The rest of the paper is organized as follows: Section 2, outlines the formulation of the SED problem and presents an overview of the main applications of SED. Moving forward in Section 3, we embark on the DCASE journey, covering the dataset, feature extraction methods, machine learning techniques, post-processing approaches, and evaluation metrics. In Section 4, we provide a comprehensive summary and analysis of the various methods proposed over the years. This is followed by a discussion on future directions in Section 5. Finally, Section 6 offers the concluding remarks for this work.

2 SED Problem Formulation and Applications

The task of determining the start and end of the sound events can be divided into two distinct stages: the training stage and the testing stage, depicted in Figure 3. The algorithm learns how the features taken from the audio input and the annotation that shows the activity of each class correspond during the training phase. The annotations are displayed as a binary matrix, where each element denotes a class that is either active (1) or inactive (0) for brief periods of time. Two additional components make up the training stage: the feature representation part and the classification part. In the feature representation part, acoustic features are extracted for each short time frame t in the audio

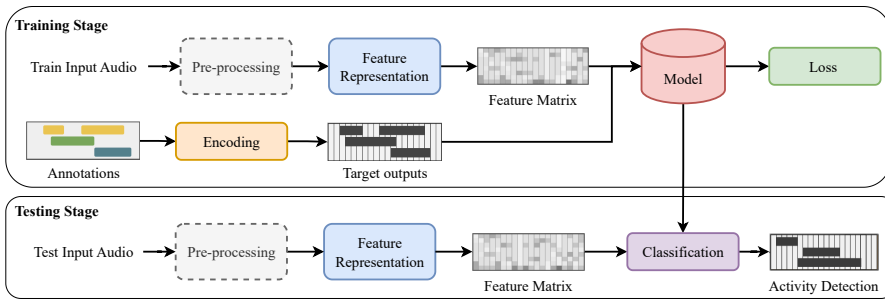

Figure 3: Overview of SED system divided into training stage and testing stage.

signal to produce the feature vector $x_t \in \mathbb{R}^M$, where M is the number of features per frame. The objective of the classification part is to learn an acoustic model that would calculate the event presence probabilities for each predefined sound event class. Further, during the testing stage, the system receives the features extracted from a test audio recording and provides the event presence probabilities using constant thresholding to obtain the matrix indicating binary activity for each sound class in consecutive time segments. The start and times of the sound event classes are calculated by combining the presence predictions for succeeding time frames.

As automatic SED systems can specify the acoustic properties of an environment, they can be used to develop context-aware devices [25, 158]. The following are some SED applications summarized in Figure 4 that could come from this feature:

- Speech comprehension and accessibility: In general, people who have a hearing aid or cochlear implant still have difficulty understanding speech in noisy environments. It is possible to use SED-based devices to classify and detect sound events, which may enhance speech comprehension in challenging listening environments. SED systems can also be used to automatically subtitle TV shows and films for viewers who have hearing impairments, which enhances the experience by assisting them in following the storyline.

- Biodiversity conservation: These devices can be utilized extensively for the conservation of biodiversity [140]. The advantages of using sound over vision make it superior for detecting and studying biodiversity movement, with a longer detection range and independent dependability regardless of the time of day.

- Safeguarding homes and cities: Based on the same benefits over optical sensors, SED systems have been integrated into smart-home devices [92,

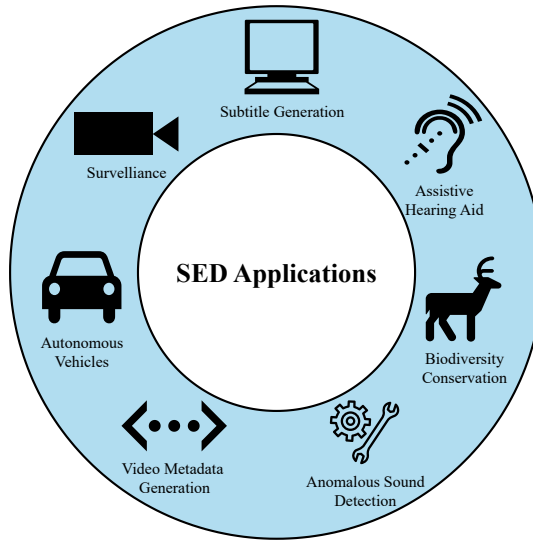

Figure 4: Applications of SED.

148] for surveillance. These are used to detect sound events such as glass shattering, the noise of a person falling, gunshots, and smoke alarms, among others [131]. Moreover, SED has been proposed for urban sound analysis in smart cities [5, 6, 27], for tasks such as monitoring noise pollution.

- Video metadata generation: With the increase in content creation through various platforms, SED systems can be used to generate metadata for the videos uploaded by the users [35]. Even though users provide specific descriptions for each video, the generated metadata can be used to efficiently search through millions of hours of multimedia data.

- Enhancing autonomous vehicles: Modern self-driving cars currently make most of their assessments and decisions using visual, ultrasonic, and radar sensors. The SED system can be integrated further into the autonomous vehicle system to recognize unusual sound events [121, 128] like car horns, railroad crossing bells, tire screeches, and ambulance sirens.

- Anomalous sound detection: It plays a crucial role in monitoring machine conditions [75, 36], particularly in the context of factory automation driven by artificial intelligence. It is an indispensable technology in the fourth industrial revolution. This advanced automation is vital for reducing the likelihood of machine malfunctions and ensuring the long-term effectiveness of the application.

3 DCASE Challenge on SED

The DCASE community annually organizes challenges to advance research in SED. These challenges provide participants with a standardized dataset and evaluation protocol to assess the performance of their algorithms. The datasets typically consist of real-world audio recordings, and the objective is to automatically detect and classify different sound events present in the audio. These challenges foster innovation and collaboration among researchers and practitioners in the field of SED. Over the years, the DCASE challenge has gained significant attention and has become a central platform for evaluating and pushing the boundaries of this domain.

We will examine the evolution of the DCASE challenge over the years as summarized in Figure 5, highlighting the key developments that have contributed to its growth. The challenge was initiated in 2013 [51] and comprised two subtasks: office live (OL) and office synthetic (OS), both conducted in an office environment. The main evaluation metric employed was the acoustic event error rate (AEER), which was computed for frame-based, event-based, and class-wise event-based evaluations. The OL subtask involves strongly sequential processing, where only one sound event is active at a time (monophonic detection), while the OS subtask deals with different degrees of overlapping events (polyphonic detection). Subsequently, in 2016, the challenge continued with a specific emphasis on polyphonic real-life audio, resembling everyday environments. The primary evaluation metric used was the total error rate (ER) [114]. In the following year, the challenge continued with the same theme but introduced a different dataset and utilized segment-based ER [111] as the primary metric. In 2018, [136], the challenge aimed to investigate the use of a large amount of imbalanced and unlabeled training data alongside a small, weakly annotated training batch to improve system performance. In subsequent years, the challenge introduced various changes and additions. The dataset underwent significant transformations, and the event-based F1-score became the primary metric used from 2018-2020. In 2019, despite the focus on semi-supervised learning, a newly generated synthetic set was introduced for training. Later in 2021, the organizers incorporated sound separation techniques in conjunction with SED. The dataset was expanded, and a new metric called polyphonic sound event detection scores (PSDS) [7] was introduced, focusing on two different scenarios. In 2022, the challenge introduced the use of embeddings from pretrained models and allowed the incorporation of external datasets. Moreover, starting this year, there was an increased emphasis on environmental impact, with the introduction of an energy consumption metric. Finally, in 2023, the challenge continued with the threshold-independent version of the PSDS metric [42] and introduced a complementary metric, multiply–accumulate operations (MACs), to measure the computational complexity. Furthermore, in the same year, the challenge included an additional

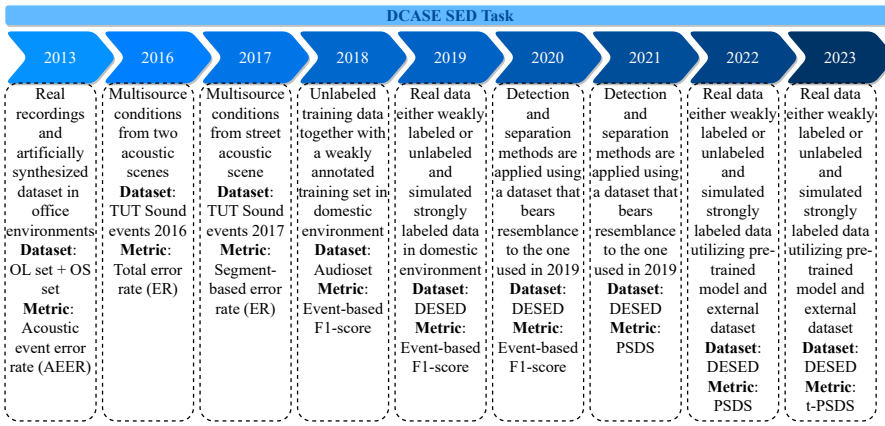

Figure 5: Summary of DCASE challenges series in SED.

subtask that involved training with soft labels. These soft labels assigned a numerical value between 0 and 1 to each label, representing the certainty level of human annotators. Participants were also given the flexibility to incorporate external datasets and utilize embeddings extracted from pretrained models, allowing them to train their systems using any combination of these resources. These continuous advancements and adaptations in the DCASE challenge have shaped the field of SED, promoting the development of more effective algorithms and methodologies while considering factors such as environmental impact and computational efficiency.

3.1 Dataset

Over the years, the dataset used in the DCASE challenge for SED has undergone significant changes, reflecting the advancements and evolving needs in the field of audio analysis. This expansion aims to capture the complexity of real-world scenarios, enabling researchers and practitioners to develop robust SED systems that can accurately recognize and classify a broader spectrum of audio events. The changes in the DCASE challenge dataset reflect the ongoing efforts to push the boundaries of SED and facilitate the development of innovative techniques that can handle real-world audio data more effectively. Throughout the years, the DCASE challenge's dataset for SED has experienced notable transformations, and Table 1 contains detailed descriptions of these changes. The training dataset contains 1-minute long clips for the year 2013, while for the years 2016 and 2017, the clip duration was extended to 3–5 minutes. Subsequently, from 2018 to 2023, the clip duration remained consistent at 10 seconds.

Table 1: This table provides a summary of different training datasets used in various editions of the DCASE challenge, including the number of classes (C), the count of strongly labeled (SL) clips, weakly labeled (WL) clips, and unlabeled (UL) clips.

Year	Subset	C	SL	WL	UL	Remark
2013	OL set + OS set [33]	16	320	-	-	Real recordings and artificially generated sounds simulating an office environment
2016	TUT sound events 2016 [114]	18	22	-	-	Recordings from two acoustic scenes: home and residential area
2017	TUT sound events 2017 [114]	6	24	-	-	Recordings from acoustic street scenes with various levels of traffic and other activity
2018	Audioset [50]	10	-	1,578	14,412	Contains an additional 39,999 unlabeled out-of-domain clips
2019	Audioset + Freesound dataset (FSD) + SINS [28]	10	2,045	1,578	14,412	Synthetic clips with FSD foreground events and SINS dataset background texture
2020	DESED [143] + SINS + TUT acoustic scenes 2017	10	2,584	1,578	14,412	Synthetic clips generated with DESED foreground and SINS dataset background
2021	DESED + SINS + TUT acoustic scenes 2017 + FSDK50K [46] + FUSS	10	10,000	1,578	14,412	Synthetic clips with DESED foreground and SINS dataset + TUT acoustic scenes 2017 background
2022-2023	DESED + SINS + TUT acoustic scenes 2017 + FSDK50K + FUSS	10	10,000	1,578	14,412	Additionally, it includes 3,470 labeled clips from Audioset (External Set) along with same set of synthetic clips from 2021.

The DCASE 2013 Task 2 OL validation set is composed of 1 minute clips capturing everyday audio events in office environments. On the other hand, the DCASE 2013 Task 2 OS validation set contains 9 clips that were generated using artificial scenes built by sequencing recordings. In DCASE 2016 edition, there was no distinct validation set; instead, the validation data was integrated into the training set and utilized for cross-validation. The same approach was followed in DCASE 2017, with no explicit separation of a validation set. For DCASE 2018 edition, the validation set was established, comprising 288 clips, which accounts for approximately 20% of the training set for that specific year. In DCASE 2019, the validation set was created by combining the validation

and evaluation sets from 2018, resulting in a total of 1168 clips. This validation set setup persisted throughout the years, remaining unchanged until 2023.

The evaluation data for DCASE 2013 Task 2 OL comprises 11 stereo recordings lasting 1–3 minutes each, featuring non-overlapping acoustic events recorded in an office environment. Similarly, DCASE 2013 Task 2 OS includes 12 recordings in which overlapping acoustic events in an office environment were artificially concatenated. The evaluation set for DCASE 2016 includes 10 audio recordings, with an equal distribution of 5 recordings from home environments and 5 from residential areas. In DCASE 2017, the evaluation set comprises 8 audio recordings, all from a single acoustic scene. The evaluation set for DCASE 2018 was a subset of Audioset, comprising 880 clips, each lasting 10 seconds and having strong labels. During DCASE 2019, the evaluation set, comprising 10-second clips, was divided into two subsets. The first subset, extracted from YouTube and Vimeo, served for ranking purposes, while the second subset comprised synthetic clips used for analysis. Subsequently, from DCASE 2020 to DCASE 2023, the evaluation set followed a similar division to DCASE 2019 but with a distribution of 10 seconds and 5 minute long clips additionally including the public evaluation set within the first subset.

3.2 Feature Extraction

Audio signals are commonly recorded in real-life environments or studios, and their raw time representation is considered redundant for sound event classification. [96] created a system that utilizes the original waveforms as a means to extract features. This system used three steps: feature learning via multiple convolutional neural networks (CNNs), feature aggregation, and final classification. In this work, acoustic features were extracted from the audio signals, predominantly focusing on the frequency domain. In general, the feature extraction process consists of three main stages: frame blocking, windowing, and frequency spectrum calculation, represented in Figure 6. During frame blocking, the audio signal is divided into short time frames, allowing for the calculation of the frequency spectrum. The duration of these frames determines the trade-off between frequency and time resolution. Typically, frame lengths range from 20 to 50 milliseconds, with an overlap of 25% to 50% to ensure a smoother representation. After that, a window function is multiplied by each individual time frame signal to reduce discontinuities at the frame boundaries that can affect the precision of the frequency spectrum estimate. This stage is known as windowing, and popular window functions employed in SED include Hamming, Hann, and Blackman. Finally, using the discrete Fourier transform (DFT), the frequency domain representation of each short time frame signal is obtained. This procedure transfers the signal from the time domain to the frequency domain, exposing the spectral component distribution inside each frame. This section delves deeper into commonly used feature extraction techniques.

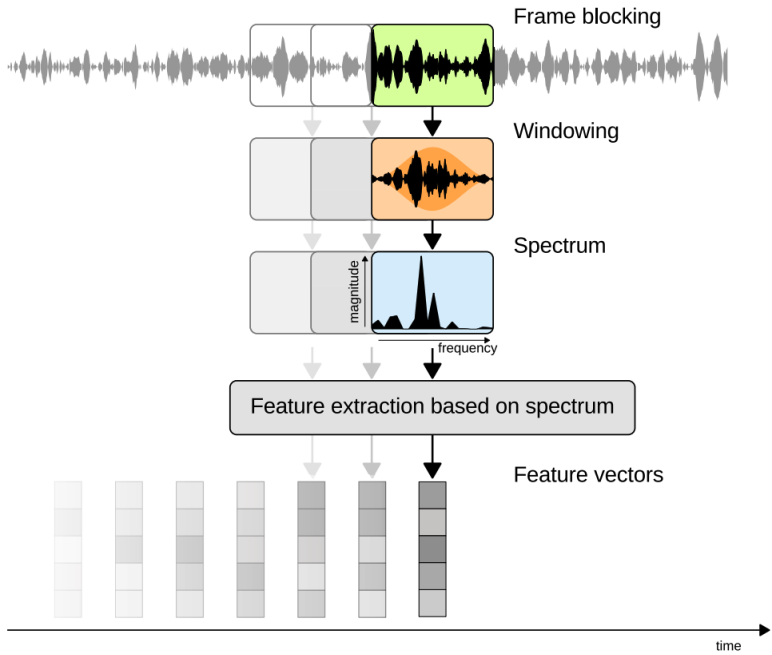

Figure 6: Stages of feature extraction given by [13].

3.2.1 Spectrogram

The spectrogram is a matrix of time-frequency features, [31]. The feature vectors from successive time frames of a recording are combined to create it. The phase information is ignored, and only the magnitude is taken into account when creating a spectrogram. This approach yields a condensed yet informative representation of sound events, leveraging the relative energy distribution in the frequency domain [63]. Due to its multidimensional nature [125], the spectrogram enables the application of extensive research on machine learning methods developed for image classification tasks to SED.

3.2.2 Mel-spectrogram

The mel-spectrogram is a representation that takes into account human auditory perception. Unlike a linear frequency scale, which humans don't perceive sound through [3], the mel-spectrogram employs a non-linear mel-scale that adjusts pitches to align with the human listener's sensitivity. This mel-scale is also utilized in mel frequency cepstral coefficients (MFCCs) [57], further emphasizing the importance of capturing human perception in audio analysis.

The mel spectrogram is a matrix of energy features obtained by applying the mel filter bank to consecutive time frames of the magnitude spectrogram. The mel filter bank employs triangle filters that become wider as the core frequencies rise, offering improved frequency resolution in the lower range. The log mel spectrogram [32] is created when the mel spectrogram undergoes a common transformation into the logarithmic scale in order to compress the dynamic range, as represented in Figure 7. In their study [123], the authors incorporated generalized cross-correlation phase transform (GCC-PHAT) in conjunction with a multi-channel log-mel spectrogram to improve the performance of SED of static sound sources.

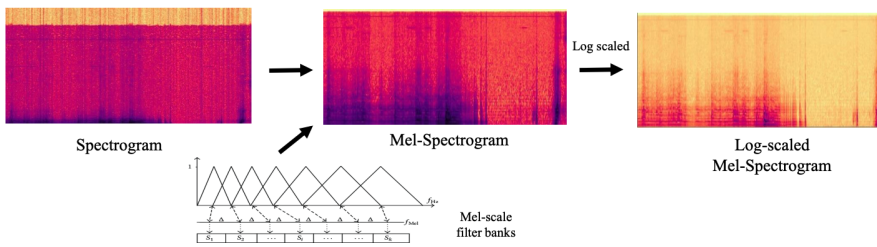

Figure 7: Deriving a mel-spectrogram from the spectrogram using mel-scale filter banks followed by log of the mel-spectrogram, adopted from [122].

3.2.3 Short Term Features

MFCCs [26] are generated through the application of a discrete cosine transform to the logarithm of the mel spectrogram, effectively minimizing the correlation between adjacent filter bank outputs. Alongside mel-scale representations, there exist alternative feature extraction techniques rooted in magnitude spectrograms. An instance is a gammatone spectrogram, which computes central frequencies based on the equivalent rectangular bandwidth (ERB) scale [52]. In their study [135], the authors incorporated a noise reduction signal enhancement process followed by Gabor filterbank (GBFB) feature extraction. The adoption of Gabor filters was motivated by their resemblance to the spectro-temporal patterns observed in the auditory cortex of mammals. [44] then first showed the effectiveness of MFCCs when comparing the GBFB, the separable Gabor filter bank (SGBFB), and Scatnet features. This was feasible because MFCCs are better built for speech and focus on lower frequencies rather than a wider frequency range. [57] showed that 15 MFCCs produce about the same performance as 20 MFCCs, indicating that the relevant information is in the general shape of the spectral envelope rather than its tiny details. In studies such as the one conducted by [1], the authors suggested utilizing three sets of

features, namely log mel-band energies, pitch frequency, its periodicity, and time difference of arrival (TDOA). These features were employed to identify overlapping sound events in a mixture and to leverage pitch cues and the stereo (multi-channel) audio signal for spatially localizing these events. Furthermore, the authors also compared three distinct binaural features and demonstrated that these features yielded comparable or improved error rates compared to single-channel features.

3.3 Machine Learning Methods

The SED systems can be classified into conventional machine learning methods and neural network learning methods based on their training process, as outlined in Figure 8.

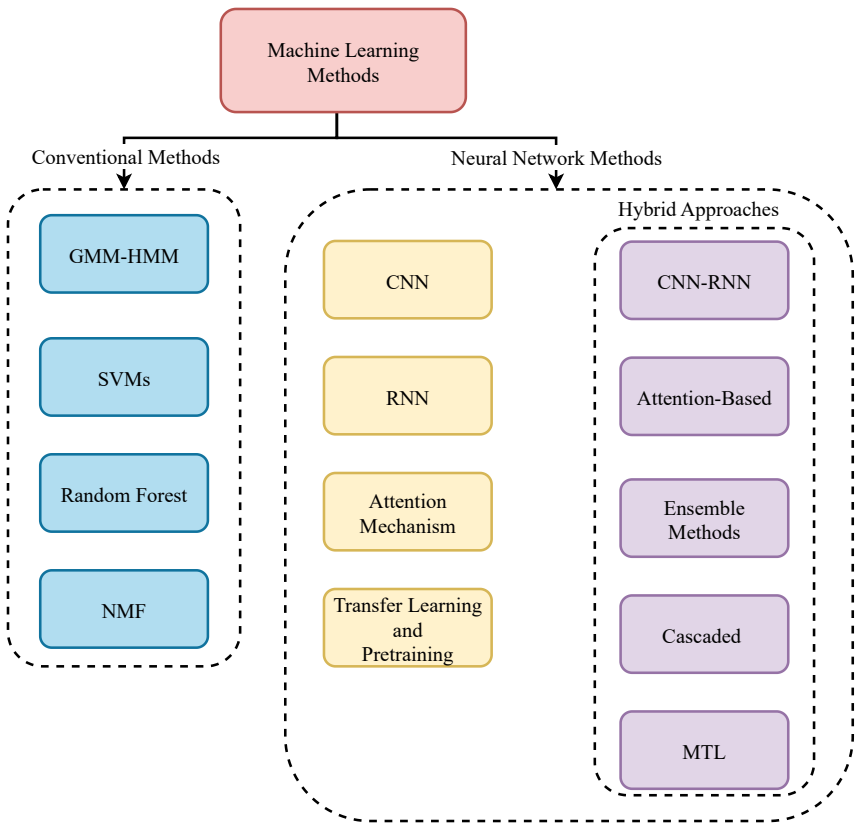

Figure 8: Split of machine learning methods.

3.3.1 Conventional Machine Learning Methods

In this field, conventional machine learning techniques include Gaussian mixture models (GMMs), hidden Markov models (HMMs), support vector machines (SVMs), random forests (RFs), and negative matrix factorization (NMF). Here, we briefly discuss each of these methods and their application to the detection and classification of sound events.

GMM-HMM

Prior research on machine learning for SED mostly focused on adapting techniques that had been previously proposed for other machine hearing applications like music information retrieval and automatic speech recognition. GMMs [110, 67, 133, 146] are probabilistic models that presumptively generate all the data points from a combination of a limited number of Gaussian distributions. The expectation-maximization (EM) method is a common and effective technique for determining the probability distribution of each component. The EM method is a two-step iterative algorithm that alternately performs an expectation step and a maximization step. Each sound class serves as a component in the recognition of sound events, and the model is trained to determine the parameters linked to the distribution of each sound class. Because each audio signal segment is handled separately by GMM, the temporal relationships in the signal are not captured. To collect the contextual data and identify the sound events using temporal dependencies, HMMs [130, 4, 33] are employed. These are a class of probabilistic graphical models that enable the prediction of a sequence of unknown (hidden) variables based on a set of observed variables. In order to categorize the current audio segment, an HMM takes into account both previous audio segments and the one that is being processed. Using this functionality as the foundation, [33] employed Hidden Markov Models (HMMs) with Viterbi decoding to discover the most probable sequence. Likewise, [126] presented a hierarchical HMM comprising a two-layer structure, which aided in making the annotated data significantly less involved. The top layer of the model represented the sound events, while the bottom layer represented the sub-events of each class. Furthermore, [60], combined HMM with bi-directional long short-term memory (BLSTM) to extend HMM to multilabel classification problems. The obtained results from the study demonstrated that the BLSTM-HMM approach exhibited superior performance compared to the baseline method based on NMF and the standard BLSTM-recurrent neural network (RNN) method.

SVMs

Another popular conventional machine learning method for SED is SVM [48, 127, 8, 30, 142]. These are discriminative models that use hyperplanes in high-dimensional space to divide data samples to produce a classifier with

the maximum margin of separation between the two classes. An SVM can be linear or nonlinear (based on a kernel). The former is used for cases that are linearly separable, while the latter is used for cases that are linearly non-separable but non-linearly (better) separable. As an example, [127] employed the linear distance between examples to construct the gram matrix, enabling the imperfect separation of training examples and smoothness of the classification boundary. The feature distribution of audio data is so complex that different classes may have overlapping regions that cannot be linearly separated; thus, a kernel-based SVM is employed. Using a sliding window, the audio clip is divided into smaller segments, and each segment is classified independently. However, SVM is unable to effectively handle large amounts of data because it scales super linearly with dataset size.

RF

RF [153, 122] is a method for ensemble machine learning that, while being trained, makes use of many decision trees. Each decision tree is an RF model that has been trained on a subset of the training data and serves as a nonlinear mapping from complex input spaces to continuous output spaces. Each tree produces a label for the input sample during inference, and the final prediction is created using the majority voting method. While overfitting is likely to occur for a single standard decision tree, a collection of randomly trained trees has high generalization power. The classification and detection of audio events are performed using RF models [44], which are trained to recognize the event in each audio segment using a set of computed features for each audio segment. In their research [126], the authors combined discriminative RF and generative hierarchical HMM described in Section 3.3.1 to merge two entirely distinct models. Furthermore, in the study by [161], a decision tree ensemble approach was employed. The systems utilized a one-vs-the-rest (OvR) multiclass/multilabel strategy, where a separate deep random forest was fitted for each event class.

NMF

Another approach that [94] popularized is a successful way to divide a non-negative matrix (X) into two non-negative matrices of size $L \times N$ into two non-negative matrices W and H of size $L \times K$ and $K \times N$, respectively, where K denotes the number of components. Which can be represented as

$$\mathbf{X} \approx \mathbf{W} \cdot \mathbf{H}$$

Here, W represents the dictionary matrix, and H represents the activation matrix, both of which are randomly initialized and updated using the multiplicative rule. Commonly, isolated events are used to extract W to create a dictionary, and SED is carried out by applying a threshold to the activation

matrix created by decomposing the test data. In their study, [49] proposed an exemplar-based approach using NMF, alongside HMMs and the Viterbi algorithm to estimate event probabilities. Furthermore, in [163], a novel SED method was introduced, which relied on supervised source separation using NMF. This method involved estimating a noise dictionary from the input signal through an unsupervised approach known as sparse and low-rank non-negative matrix factorization (SLR-NMF).

Due to specific techniques that enable the modeling of basic speech or musical units, such as the state-tying of phonemes or left-to-right topologies for simulating the temporal evolution of musical notes and phonemes, such methodologies are much more beneficial in modeling speech and music. Since sound events generally do not have the same fundamental building blocks as speech, these conventional methods are less useful for SED. Moreover, despite the fact that the aforementioned techniques are easy to implement, they are not designed to identify multiple overlapping classes. In contrast, the recently proposed deep neural networks, which can easily perform multilabel classification, have taken the lead in audio event detection and classification. They can exhibit simultaneous activation of multiple output neurons, indicating the simultaneous activity of multiple sound classes. [44] considered 12 classifiers decision tree, RF, Xtreme gradient boosting, SVMs, K-neighbors, and logistic regression, and compared the performance for the different feature types. Furthermore, [57] showed the effectiveness of using a non-parametric discriminant approach based on the k-nearest-neighbors (kNN) rule.

3.3.2 Neural Network Methods

An artificial neural network (ANN) is a machine-learning technique inspired by the information-processing capabilities of the human brain. Just as neurons in the human brain specialize in processing specific signals and continually improve their abilities to create a mapping between input signals and their cognitive representations, ANNs operate on a similar principle. ANNs consist of interconnected blocks of artificial neurons, working towards the goal of finding a mapping between input signals and desired output signals.

CNN

The advent of deep learning in the past decade, particularly CNNs, has revolutionized SED. CNNs, with their multiple convolutional layers, excel at capturing low-level spectral patterns and progressively learning higher-level representations, facilitating the extraction of discriminative features crucial for SED. [23] initially showed the effectiveness of using a CNN-based classifier in SED. Additionally, CNNs exhibit translation invariance properties, enabling them to detect sound events regardless of their temporal position within the audio signal. This property enhances the robustness of SED systems by

accommodating temporal variations. Furthermore, CNNs leverage parameter sharing, which effectively reduces the number of learnable parameters compared to fully connected networks, streamlining training and inference processes.

The novel approach proposed by [73] depicted in Figure 9 involves splitting the input into short-term data and long-term data, each with different time lengths. These segments are then merged after passing through two convolution layers, utilizing various merging techniques that were experimented with. Likewise, [56] introduced a multi-scale approach in their model, which incorporates two separate CNNs operating at fine-scale and coarse-scale respectively. In their study, [105] introduced a capsule-based neural network to effectively handle sound events of different time duration. To address this challenge, they utilized three windows of varying sizes to partition the output CNN layers. Moreover, they also incorporated an event activity detection (EAD) technique that leveraged energy information to enhance the detection of weak labels. [87] in their study employed two models: AlexNet with 4 layers and VGG with 8 layers. They aimed to showcase the system's performance on multiple tasks in order to illustrate the varying levels of difficulty associated with each task. On the other hand, [101] introduced the concept of the Inception module, which addresses the challenge of handling multiple receptive fields simultaneously within each CNN layer. This module is inspired by the idea of identifying the optimal local sparse structure in a convolutional vision network. With an aim to integrate traditional machine learning with CNNs, [17] proposed the idea of integration of NMF with CNN, to use NMF to provide an approximate strong label to the weakly labeled data.

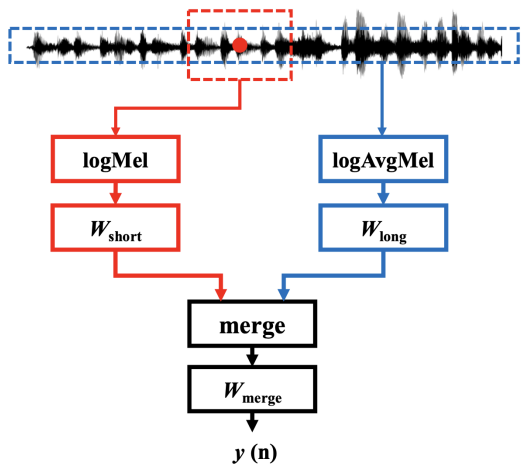

Figure 9: Overview of the system developed by [73] for detecting events from the red dot point using short- and long-term data.

Specifically, focussing on the localization ability, [162] suggested the utilization of selective kernel (SK) units, represented in Figure 10. These units allowed each neuron to dynamically adjust its receptive field, enabling adaptability for both short- and long-duration events. Building upon this concept, [21] further extended the application of SK units by integrating them with a VGG block. This integration involved incorporating four residual blocks, each equipped with SK units, resulting in the model known as VGGSK [104]. Similarly, [149] incorporated a multi-scale CNN block along with efficient channel attention to effectively capture more comprehensive features and combine features of various scales, prioritizing the crucial areas within the features. A separate team, concentrated on developing a lightweight design referred to as CDur [10], to have a real-world application-oriented approach for SED that utilized unsupervised data augmentation, and successfully reduced the model size to a mere 600k parameters using their proposed architecture.

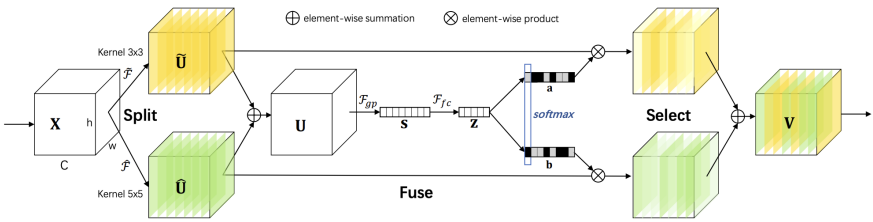

Figure 10: Representation of SK units utilized in [162].

Recurrent Neural Network (RNN)

The emergence of RNNs marked a significant breakthrough in capturing temporal dependencies and modeling sequential data. While vanilla RNNs and Elman networks [45] initially showcased potential in modeling sequential audio data, they encountered challenges with the vanishing gradient problem, hindering their ability to capture long-term dependencies. To address this issue, the introduction of long short-term memory (LSTM) networks proved pivotal. [1] used the RNN-LSTM network for multilabel SED using spatial and harmonic features. LSTMs incorporated memory cells and gating mechanisms, enabling effective modeling of long-term dependencies by retaining and selectively utilizing past information. Building upon this progress, bi-directional LSTMs (BiLSTMs) were introduced, processing input sequences in both forward and backward directions to capture contextual information from both past and future contexts. Additionally, gated recurrent units (GRUs) [108] were introduced as a simpler alternative to LSTMs, offering comparable capabilities in modeling temporal dependencies while streamlining the architecture.

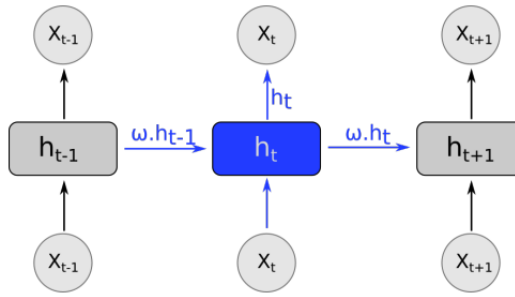

Figure 11: Structure of the recurrent links between RNN cells according to the weight w, as proposed by [15]. Where the hidden state h_t of an RNN cell depends on x_t, the incoming output of the previous layer at time t, and h_{t-1}, its hidden state at time $t-1$.

In a similar time period, [15] investigated the potential of incorporating a weighted-GRU model depicted in Figure 11, where $\omega = 1$ represents the standard GRU. The intention behind this proposal was to reduce the impact of hidden states to prevent consistent predictions of the same score throughout an entire recording. In their work, [38] advanced further by proposing a method that utilizes an RNN as a language model (LM) in the SED task. This approach involves incorporating the RNN before the final layer of the SED system and conditioning the RNN input with the previous time step's class activities. To bolster the decision layer's robustness, [160] proposed deep-RNN (DRNN). This architecture establishes dense connections between each pair of RNN layers, enabling multiple rounds of thinking and decision-making in the decision layer. While much of the attention was directed towards architectural aspects, [98] took a different approach by focusing on addressing overfitting in GRU. Their submission incorporated a regularization method into the GRU component, which aimed to enforce consistency in the output produced by different sequence modeling processes.

Attention Mechanisms

The attention mechanisms improved the performance by focusing on relevant parts of the audio signal when detecting sound events. Attention mechanisms allow the network to dynamically allocate weights to different temporal segments, emphasizing informative regions and suppressing irrelevant ones. Early attention mechanisms in SED were often inspired by visual attention models [37, 70, 129]. These mechanisms used techniques such as soft attention, where a weighted combination of features is computed at each time step based on their relevance to the sound event being detected. To capture long-term dependencies and improve temporal modeling, temporal attention [118, 151, 77]

mechanisms were introduced in SED. Temporal attention mechanisms enable the network to attend to past and future contexts, facilitating the detection of sound events with complex temporal patterns. Self-attention mechanisms, such as transformer-based architectures [19, 91, 54], have also gained prominence in SED by enabling models to capture relationships between different parts of the audio signal without sequential dependencies. To further enhance attention modeling, multi-head attention [116, 99] was introduced in SED. Multi-head attention mechanisms allow the network to attend to different parts of the audio signal simultaneously, capturing multiple aspects of sound events. Hierarchical attention mechanisms [19, 144] have also been employed, where attention is applied at different levels of granularity, such as attending to both local and global temporal contexts.

The model [81] included a convolutional block attention module (CBAM) in the convolutional layers to attend to relevant features, enhancing SED, as shown in Figure 12. The authors of [157] expanded on the concept of CBAM by emphasizing the similarity between the temporal dimension of time-frequency features and the channel dimension in computer vision. In light of this observation, they introduced CBAM-T, a variant where the input features are transposed along the time and channel dimensions. In their research, [165] introduced a split attention mechanism, consisting of two parts: group and attention. This approach allows for the independent learning of diverse sub-features and generates attention weights to assess the significance of each sub-feature. Delving deeper into attention mechanisms, [24] suggested the implementation of an axis-wise attention module (AWAM) that draws inspiration from the parallel temporal-spectral attention method. This involves calculating a sigmoid-based score for each axis and incorporating it into the input feature map. The author's [80] subsequent work involved integrating large kernel attention (LKA) into a frequency dynamic convolution-recurrent neural network (FDY-CRNN). This integration, combined with pretrained bi-directional encoder representation from audio transformers (BEATs) [20] embeddings, effectively captured time-frequency patterns, long-term dependencies, and high-level semantic information in audio signals.

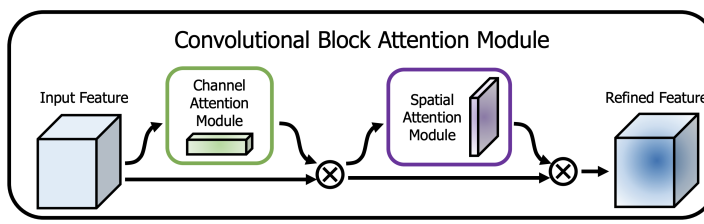

Figure 12: Overview of convolutional block attention module (CBAM) structure to improve attending to the relevant features incorporated by [81].

Transfer Learning and Pretraining

Transfer learning and pretraining revolutionized this landscape by allowing models to benefit from pre-learned representations, enabling them to extract higher-level and more discriminative features. Early applications of transfer learning in SED involved utilizing pretrained models [86] trained on general audio tasks, such as music classification or speech recognition [124]. Researchers adapted these models by fine-tuning them to specific SED datasets, effectively transferring the learned representations to the target SED task. In recent years, there has been a shift towards end-to-end fine-tuning of pretrained models in SED. Instead of using pretrained models solely for feature extraction, the entire model is fine-tuned on the target SED dataset. The general approach of incorporating a pretrained model into the proposed CRNN architecture is depicted in Figure 13. This process involves concatenating the embeddings obtained from the pretrained model using an aggregation method. The resulting output is then fed into the proposed RNN component.

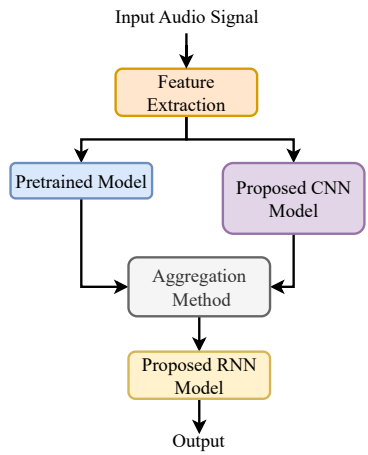

Figure 13: Overview of utilizing the pretrained model in combination with the proposed architecture.

For their submission, [62] integrated the pretrained audio neural networks (PANNs) [86] model and pretrained audio spectrogram transformer (AST) [53] model to extract embeddings. These embeddings were then combined with the SK-CRNN proposed by [162] and FDY-CRNN [156, 47] introduced by [119] which applies a kernel that adapts to frequency components of the input. During the same challenge year, [154] conducted experiments involving the self-supervised audio spectrogram transformer (SSAST) [54] model, alongside PANNs, to enhance the generalization and robustness of their models. In the recent study [39], the performance of various pretrained models, such as AST,

PANNS, patchout fast spectrogram transformer (PaSST) [91], hierarchical token-semantic audio transformer (HTSAT) [19], and BEATs, was compared alongside the FDY-CRNN model. The experiments revealed that the features extracted from BEATs outperformed those of other pretrained models. While keeping the BEATs features fixed, one of the other pretrained models was selected, and its features were combined with BEATs features and CNN features using a fusion approach. Task-aware fine-tuning (TAFT) and self-distillated mean teacher (SdMT) are two ways that the PaSST fine-tuning procedure proposed in [97]. While SdMT assisted in the training of a robust model through the distillation of soft knowledge, TAFT was used to make use of both local and semantic information from PaSST.

Hybrid Approaches

The advent of hybrid approaches in model architecture for SED stemmed from the recognition that single-model solutions often struggled to capture the full complexity of sound events. Researchers began combining different architectural elements, such as CNNs, RNNs, and attention mechanisms, to create more powerful and adaptable SED models. We further describe them as follows:

- CNN-RNN Hybrid Model: CRNNs combine CNNs and RNNs, as CNNs excel at capturing local spectral patterns, while RNNs are effective in modeling temporal dependencies. By integrating these architectures, hybrid models could capture both short-term and long-term contexts, leading to improved SED accuracy. [12] demonstrated the use of CRNN depicted in Figure 14 and showed improvement over feedforward neural networks (FNNs) and CNNs. In the study conducted by [100], they showcased the efficacy of a hybrid neural network combining 1D ConvNet and RNN with LSTM units. This hybrid model proved to be highly effective in accurately determining the onset time. Building upon the previous work, an extension was presented in [56] by introducing a multi-scale CRNN. The primary objective was to combine information from various time resolutions, enabling the model to capture both fine-grained and coarse-grained features of sound events. In [59], gated mechanisms were introduced into the CRNN network to selectively allow or block the flow of information based on the presence or absence of relevant audio events. In their submission for DCASE 2019 challenge, the authors of [159] presented the application of a residual CRNN framework to establish a relationship between local features and contextual features. This network was later extended by [82] in their work involving a self-training-based noisy student model for predicting strong labels for sound events. Furthermore, as a continuation of this idea, a self-mask module was integrated into ResNet [83] as a region proposal network. This

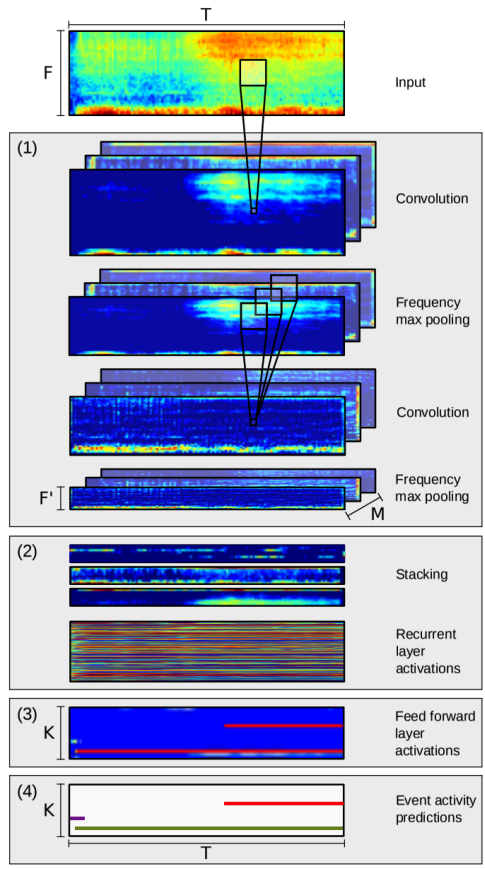

Figure 14: The information flow and learning process in the CRNN architecture, as suggested by [12].

addition allowed for the detection of event time boundaries, enabling the self-mask module to limit the duration of both silent and sound events. Next, [40] proposed the forward-backward convolutional recurrent neural network (FB-CRNN) with two RNN classifiers sharing a CNN for preprocessing. One RNN operates in the forward direction and the other in the backward direction. The goal is to promote early event tagging by training the RNNs to jointly predict audio tags at each time step, considering the collective processing of the entire recording by both RNNs. Subsequently, [43] extended this work by incorporating multiple iterations of self-training into the system. To ensure scale invariance, [84] proposed the feature-pyramid CRNN depicted in Figure 15. This

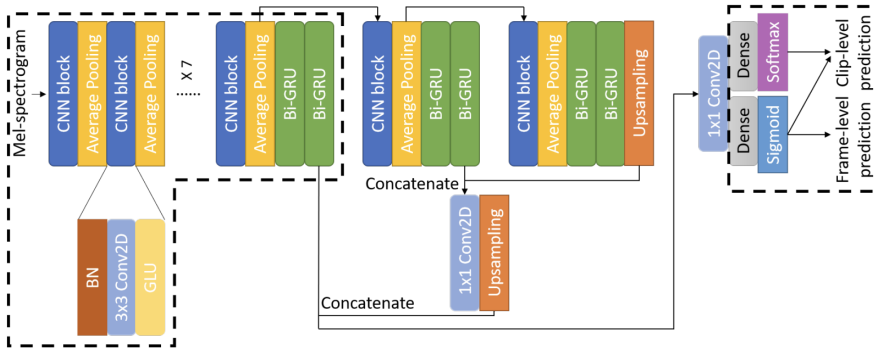

Figure 15: The proposed feature-pyramid CRNN architecture by [84].

involved pooling the second last layer of the CNN with different sizes, upsampling the resulting feature maps, and combining them with the last layer's feature maps.

- Attention-Based Hybrid Model: In hybrid models for SED, the incorporation of an attention mechanism allows the model to concentrate on pertinent audio segments. By attending to specific regions of the audio spectrogram, the model emphasizes significant features while disregarding irrelevant information. Hybrid models rely on transformer architectures, similar to RNNs, to effectively capture both local and global dependencies. The inclusion of attention mechanisms in these architectures further enhances the model's capacity to identify crucial details and adapt to diverse SED tasks. The researchers in [116] employed a model called conformer, which combines CNNs and transformers, to better utilize local features in audio data while capturing global features. The study also included a comparison of performance between conformer and transformer models, as well as an evaluation of their fusion performance.

  The work by [99] combines two distinct models: the sound event detection transformer (SEDT) and a frame-wise model. The SEDT illustrated in Figure 16 is an event-wise model that learns representations at the event level and directly predicts sound event categories and boundaries. On the other hand, the frame-wise model follows the commonly used frame-classification approach, where each frame is classified into event categories, and event boundaries are obtained through post-processing techniques like thresholding and smoothing. In their approach, [137], employed the audio teacher-student transformer (ATST) model, trained with Audioset, for clip-level audio processing. They introduced a dedicated clip-level classification token, which gathers information from the frame-

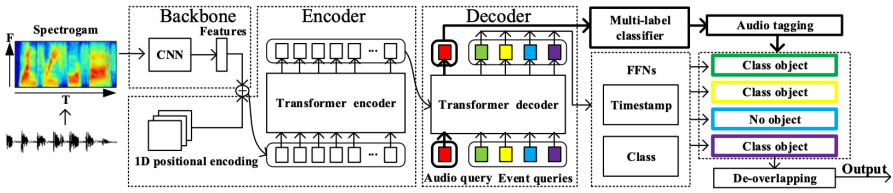

Figure 16: The proposed architecture of sound event detection transformer (SEDT) by [99].

level embeddings, enhancing the model's understanding of audio at the clip level. The research conducted by [150] aimed to prioritize energy efficiency, training speed, and the reduction of carbon emissions. The approach involved extracting embeddings from a range of vision transformer (ViT) models [37], along with the utilization of two neural network-based classifiers. One classifier was designed to predict on- and offsets, while the other employed a simple linear classifier.

- Ensemble Methods: Ensemble methods combine the predictions of multiple models or classifiers to improve SED performance. For instance, an ensemble of CNNs with different architectures or initialization can be trained, and their outputs are aggregated to make the final prediction. Ensemble methods increase the robustness and reliability of the SED system, as diverse models capture different aspects of sound events. In their approach, [95] employed an ensemble of ConvNets with multiple analysis windows. Their system combined the outputs of global-input and separated-input models to make predictions about the timestamps of the input audio. By utilizing ensemble selection methods, they aimed to minimize errors in the process. The work described in [34] emphasizes the potential for substantial improvements through the use of a large ensemble comprising various architectures and frameworks. The researchers developed four distinct networks, namely, SCRATCH, SMALL, PRECISE, and TAG, each designed to excel at a specific task.

- Cascaded Architectures: Cascaded architectures [78, 79] consists of a sequence of models, where each subsequent model refines the predictions of the previous one. This hierarchical approach enables more detailed and accurate detection of sound events. For example, a first model may perform coarse event detection, followed by a series of models that progressively refine the detection, reducing false positives and increasing precision.

- Multi-Task Learning (MTL): In SED, MTL is a beneficial approach, where a model is trained to handle multiple tasks simultaneously. This

setup allows for the exchange of knowledge between tasks, leading to improved performance for each individual task. By utilizing the existing dataset, MTL enables the model to learn from multiple tasks and leverage their relationships. In a research paper, [72], the authors proposed jointly analyzing sound events and acoustic scenes to exploit their interdependencies. Another study [76] focused on utilizing shared layers and weighted loss to capitalize on distinctive high-level acoustic characteristics of different sound events. These examples demonstrate the effectiveness of MTL in SED and highlight its potential to enhance performance through knowledge sharing.

3.4 Post-Processing

In this section, we review the post-processing methods that have been utilized through the DCASE challenge on SED, as summarized in Figure 17. The methods of post-processing aim to refine the output of the SED models and improve the accuracy of event detection. In their initial study, [163] employed a baseline approach where a constant value was used for the median window size in the median filtering process across all classes. [22] further enhanced the approach by optimizing the threshold to 0.4 instead of the standard 0.5 threshold, resulting in an approximate 4% performance improvement. In addition, [152] expanded upon this threshold selection by establishing a minimum event length and minimum event gap of 100 milliseconds. These criteria were utilized to determine the exact start and end times of the detected acoustic events. During that same year, [69] introduced an intriguing approach to enhance the output by leveraging a logical rule. According to their method, if the label at the middle index differs from the other two labels, while the other two are identical, the middle label is classified as misclassified and subjected to smoothing as represented below:

$$s_i = s_{i-1}, \text{ for } (s_{i-1} = s_{i+1}) \text{ and } (s_{i-1} \neq s_i) \tag{1}$$

where s_i is the i^{th} output label in the sequence. Furthermore, they computed the duration of each instance and determined the shortest duration as their threshold. If the duration of a recognized sound event is shorter than the threshold, it is considered non-occurring and subsequently removed.

In the following year, [68] opted for a constant threshold value for each class, which they fine-tuned individually on the validation dataset. These optimized thresholds were subsequently applied to the test and evaluation datasets. Following a similar approach, the researchers [14] employed the genetic algorithm to optimize the threshold. They achieved this by randomly selecting a value, denoted as δ, from a normal distribution and then adding or subtracting this value from the threshold estimate. This modification process

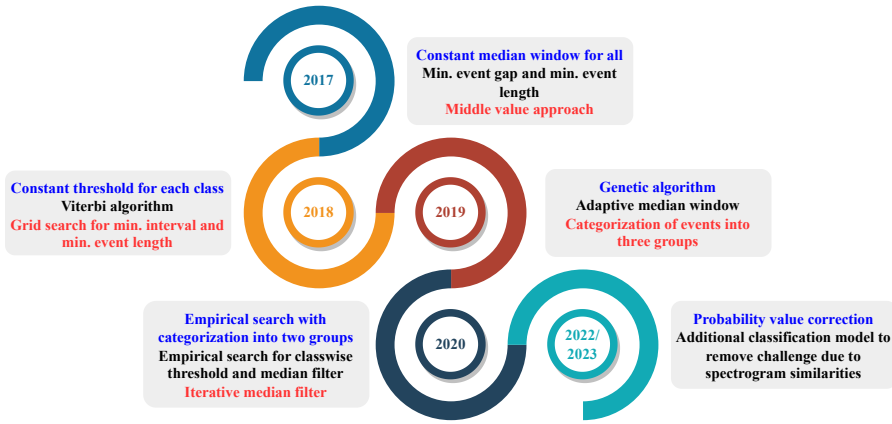

Figure 17: The introduction years of various post-processing techniques.

was designed to be more pronounced around 0.5 and less significant towards the boundaries. The authors also proposed hysteresis thresholding, using two thresholds to determine the onset and offset of an event.

Likewise, in a similar vein, [105] introduced a smoothing technique that integrated grid search to identify a minimum interval of n frames between two events, along with a minimum event length of m frames. In the fascinating research conducted by [101], they utilized the Viterbi algorithm on the frame probabilities of each class to generate binary values by determining the most probable state sequence given the observations. Furthermore, they incorporated median filter sizes that were tailored to the estimated length of the event. Similarly, to accommodate the distinct event categories, [103, 102, 58, 160, 107, 156] implemented a set of median filters with adaptive window sizes (S_{window}) that were determined based on the varying durations observed in different real-life event categories and given as:

$$S_{window} = duration_{average} \times \beta \tag{2}$$

where β was taken as $1/3$ and $duration_{average}$ represents the average duration of each event that occurred in the training set.

In the same year, the researchers [29] categorized the classes into three groups: “impulsive sounds”, “intermediate sounds” and “background sounds”, as represented in Figure 18. To adapt the median filtering technique to the type of category each class belongs to, specific window sizes were assigned to each group: 5, 13, and 41, respectively. Following a similar approach, the authors of [84, 141] divided only into two groups, “background sounds” and “impulsive sound” and then empirically searched for the optimal threshold. In their study [116, 109, 149, 154, 34, 104], the authors extended the empirical

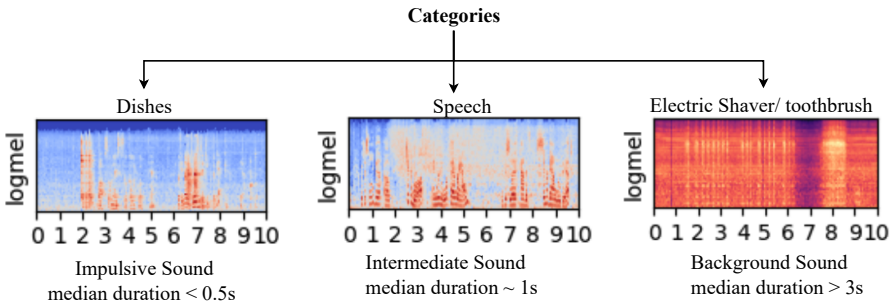

Figure 18: The division of classes into three groups, "impulsive sounds", "intermediate sounds", and "background sounds".

search for the optimal threshold and median filter size. They explored a range of values for the threshold and the median filter size in incremental change. The authors of [18] proposed using an iterative median filter with event-specific window sizes, following a similar implementation by [85]. They also considered neighboring frames as activated if they exceeded a lower threshold of 0.08. Events with a duration shorter than 0.1 seconds were removed as noise, and if the time difference between the offset of the first event and the onset of the second event was less than 0.2 seconds, they were concatenated. Extending on the adaption of median filtering, [117] utilized a median filter to process probabilities and introduced a probability value correction method. This correction involved applying a magnification factor to adjust the existence probability of each class in the final output, with a maximum value of 1.0. Lastly, in their research, [155] noted that certain classes present challenges for the models due to their spectrogram's similarities. To address this issue and compensate for the model's limited ability to distinguish these classes, they trained additional models specifically for further classification. Each detected event belonging to these challenging classes underwent dual classification using the classification model. The probability value of the class verified by both models was increased, while the probability of the opposite class was decreased. Furthermore, they placed special emphasis on the "dishes" class, as it exhibited the shortest duration and posed the greatest difficulty for detection.

In summary, most of the post-processing techniques in SED encompass a range of methods, including filtering, threshold optimization, contextual information utilization, and class-specific refinements. By leveraging these techniques, the contributions have been able to improve the overall performance of SED systems.

3.5 Evaluation Metric

To evaluate the performance of an SED system, computational metrics are employed to compare the system's output with a reference annotation. The condensed form of this information can be found in Table 2. The DCASE challenge held in 2013 comprised two subtasks, and the evaluation metric employed was the acoustic event error rate *(AEER)* [139], as represented below.

$$AEER = \frac{D + I + S}{N} \tag{3}$$

where N represents the current frame's number of events to detect, while D, I, and S correspond to the count of deletions (missing events), insertions (extra events), and event substitutions, respectively. The DCASE challenge in 2016 incorporated the total ER [113] as the metric, by considering both the false positives and false negatives in the detected sound events. To compute the total ER, the total number of errors *(E)* is computed by summing the false positives *(FP)*, false negatives *(FN)*, and substitutions *(S)*:

$$E = FP + FN + S \tag{4}$$

Then the total ER is then obtained by normalizing the total number of errors by the total number of sound events in the ground truth annotation. In the subsequent year, 2017, the challenge transitioned to employing a segment-based error rate [113] as the evaluation metric. This metric focused on aligning the system output and the reference annotation at the segment level, with each segment spanning a duration of 1 second. During the period from 2018 to 2020, the challenge transitioned to using an event-based F1-score for evaluation. This metric provided a more detailed assessment of performance by considering individual events, event boundaries, and overlapping events. To account for slight timing variations, a tolerance of 200 milliseconds was applied, meaning that detected events within 200 milliseconds before or after the true onset time were considered correct detections. This approach allowed for a more lenient evaluation that accommodated practical timing differences without penalizing the system unnecessarily.

Between 2021 and 2022, the challenge adopted polyphonic sound event detection scores (PSDS) [7] as the evaluation metric. This metric addressed the limitations of collars by using an intersection-based approach for matching against the ground truth, as depicted in Figure 19. This approach enhanced the robustness of performance measurements by reducing the impact of labeling subjectivity. Additionally, the metric was computed over a range of operating points (50 operating points) rather than a single system setting, allowing for a more comprehensive evaluation. Moreover, the proposed method offered flexibility by enabling adjustments to evaluation parameters, and accommodating diverse application needs and user experience requirements. In the

Table 2: Evaluation metric employed over the years in the DCASE challenge for the task of SED.

Year	Metric	Description
2013	Acoustic event error rate (AEER)	Calculated for frame-based, event-based, and class-wise event-based metrics.
2016	Total error rate (ER)	Error rate was evaluated in one-second segments over the entire test set.
2017	Segment-based error rate	Calculated in one-second segments over the entire test set
2018-2020	Event-based F1-score	Event-based measures with a 200 milliseconds collar on onsets and 200 milliseconds / 20% of the events length collar on offsets.
2021-2022	Polyphonic sound event detection scores (PSDS)	Computed using 50 operating points (linearly distributed from 0.01 to 0.99) for two different scenarios that emphasize different systems properties
2023	Threshold-independent PSDS	Computed using timestamped scores rather than detected events for two different scenarios that emphasize different systems properties

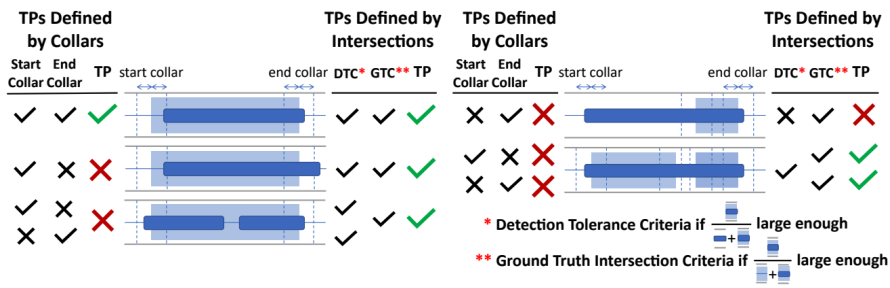

Figure 19: The calculation difference between collars and Detection Tolerance Criteria (DTC)/Ground Truth Intersection Criteria (GTC) for PSDS metric, adopted from [7].

challenge, two different scenarios emphasize different systems properties with scenario-1 (PSDS1) focusing on the need to react fast upon an event detection and scenario-2 (PSDS2) to avoid confusion between classes with the reaction time being less crucial than in the first scenario. For ranking purposes, the metric was an aggregation of $\overline{PSDS1}$ and $\overline{PSDS2}$, with $\overline{PSDS1}$ and $\overline{PSDS2}$ the PSDS on scenarios 1 and 2 normalized by the baseline PSDS on these scenarios, respectively:

$$RankingScore = \overline{PSDS1} + \overline{PSDS2} \tag{5}$$

In 2023, during the final year of the challenge, a new version of PSDS called the threshold-independent variant [42] (t-PSDS) was introduced. This version was created based on the concept that calculating PSDS using a predetermined set

of thresholds could result in a biased assessment of the ultimate measurement. This bias could potentially lead to a significant underestimation of performance if an unfavorable set of thresholds is utilized.

4 Discussion and Summary

In this section, we provide a comprehensive overview of the findings and insights obtained from different leading teams participating in the SED task of the DCASE challenge from the first edition in 2013 till the very recent edition in 2023. We thoroughly examine each year, emphasizing the strategies employed by the top teams, and their system proposals in comparison to the baseline. This allows us to identify emerging trends, novel algorithms, and improved methodologies that have shaped the landscape of SED. Ultimately, in the concluding part of the section, we consolidate our findings and present a comprehensive overview of the general transformations witnessed in SED methods over the years.

4.1 DCASE 2013 Task 2

In the first edition of the DCASE challenge series consisting of two subtasks, the organizers utilized acoustic event error rate (AEER) as the primary metric for the frame-based, event-based, and class-wise event-based evaluations. The baseline [51] was based on NMF, which is common to both subtasks and involves learning a dictionary of spectral basis vectors through NMF on the training data. This fixed dictionary is then used for the NMF decomposition of unlabeled audio files from the development set, generating an activation matrix. By summing and thresholding the activation vectors per class, the activity for different classes is obtained.

4.1.1 DCASE 2013 Task 2-OL

The baseline for DCASE 2013 achieved an AEER of 2.59 as represented in Figure 20. According to results, [146], the leading team significantly improved on this and achieved an AEER of 1.001 by employing GMMs estimated from MFCCs. However, they noted that the utilization of GMMs faced challenges in handling significant variations in characteristic sounds within certain classes and a limited number of training examples. Following that, the subsequent team [135] achieved an AEER of 1.016 by integrating the GBFB feature extraction stage and a two-layer HMM as their back-end classifier. For estimating the time regions of events in a signal, they employed Viterbi decoding. The third team [49] employed the exemplar-based NMF method (Section 3.3.1) that aided in achieving an AEER of 1.084. They represented

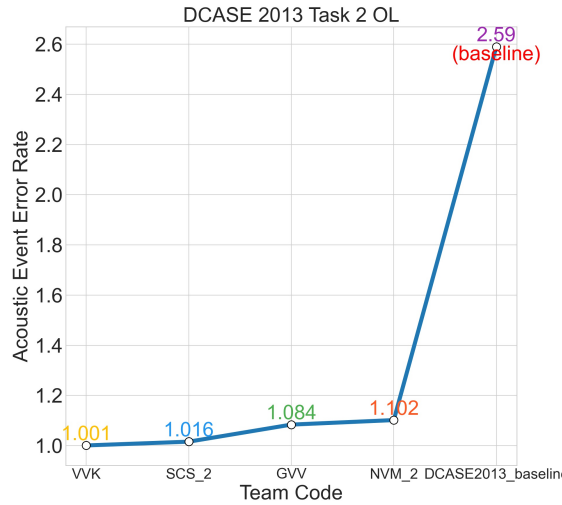

Figure 20: Comparison of system performance among the top teams in DCASE 2013 Task 2 OL with the baseline.

events as a linear combination of dictionary atoms and mixtures as a linear combination of overlapping events. They used a dictionary of atoms extracted from available training data, which was augmented by linear time warping at multiple rates. In their work [126], the fourth team employed a combination of temporal-based, spectral-based, autocorrelation, and multidimensional features to handle the diverse characteristics of sound events in an office environment. To classify these audio features, they utilized a two-layer hierarchical HMM, effectively capturing temporal dependencies within and between sound events. This approach contributed to their achievement of an AEER of 1.102.

4.1.2 DCASE 2013 Task 2-OS

This subtask also utilized the same baseline as was used in OL task, which achieved an AEER of 2.804 as represented in Figure 21. The top-performing team [49] participated in both the OL and OS tasks. They observed that their performance was slightly diminished in the OS task achieving an AEER of 1.318, mainly due to the presence of added noise and overlapping events. Noise tends to be less structured, making it more challenging to model accurately. Subsequently, the team that ranked first in the OL task [146] obtained an AEER of 1.888 using a similar approach. However, their performance declined

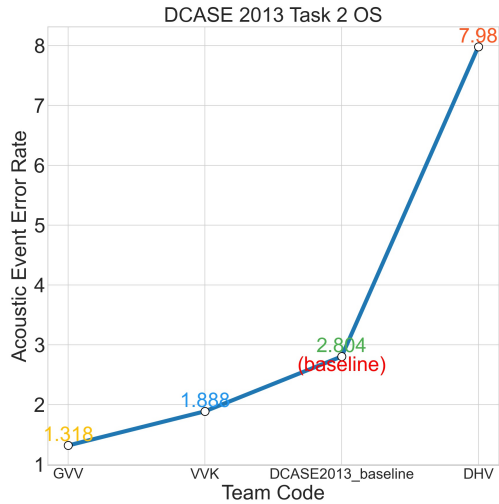

Figure 21: Comparison of system performance among the top teams in DCASE 2013 Task 2 OS with the baseline.

in the OS task, mainly due to the challenges posed by GMMs in effectively modeling significant variation in characteristic sounds and a limited number of training examples. Finally, the team with an AEER of 7.98 employed HMMs and Viterbi decoding, as explained in Section 3.3.1, to determine the most probable event sequence. They extended this approach by incorporating multiple detection passes, resulting in the production of a polyphonic event sequence.

To summarize, the challenge primarily relied on conventional machine learning methods. While they established a baseline for the subsequent 2016 challenge, these traditional models encountered difficulties in capturing long-term dependencies in audio sequences. Additionally, they faced challenges in coping with noisy environments and handling variability in sound events, resulting in reduced detection performance. As a result, there emerged a need to explore models specifically designed to handle data sequences with time dependencies in the following years.

4.2 DCASE 2016 Task 3

In this edition, the challenge used total ER as the evaluation metric, where the evaluation was done over 1 second segments over the entire test set, the trend demonstrated in Figure 22. The baseline [65] for this year was based

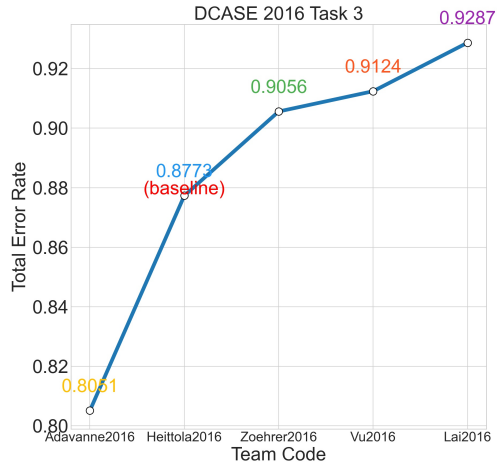

Figure 22: Comparison of system performance among the top teams in DCASE 2016 Task 3 with the baseline.

on MFCC acoustic features and the GMM classifier, building upon the top team from DCASE 2013 Task 2 OL. The training process involves utilizing audio segments that are annotated as belonging to the specific event class for training the positive model. Simultaneously, a negative model is trained using the remaining audio segments. The decision-making relies on calculating the likelihood ratio between the positive and negative models for each class, considering a sliding window of 1 second. In comparison, we observe that the baseline comes second in the overall challenge ranking when compared using total ER.

To surpass the baseline performance [1] with a total ER of 0.8051, the enhanced system adopted spatial and harmonic features in comparison to the mono-channel features employed in the baseline. These features were combined with an LSTM, as explained in Section 3.2.3. This approach drew inspiration from human auditory perception, which utilizes two ears (two channels) to identify and localize surrounding sound events. The team [166] ranked after the baseline with a total ER of 0.9056 which trained a 3-layer gated recurrent neural network (GRNN) with the usage of the log-magnitude spectrogram. The usage of the log-mel spectrogram gave an improvement in the performance over MFCC, as reported in their work. The team ranked fourth [145], achieved a total ER of 0.9124, and incorporated a bi-directional RNN (BiRNN) with 50 hidden units. This BiRNN featured a second hidden layer that learned the input sequence in the reverse direction. On the other hand, the team [93]

that obtained a total ER of 0.9287 utilized both SVM and ANNs with 2–3 layers. This team showcased the distinct performance characteristics of these two classifier types and how they offer complementary information.

In summary, the first edition of the DCASE challenge laid a foundation for future advancements in the field, with the major teams placing significant emphasis on utilizing RNNs to model temporal dependencies. Through their research, they demonstrated how the performance of these models can be influenced by various factors, such as the choice of feature extraction techniques and the number of layers in the neural network architecture.

4.3 DCASE 2017 Task 3

The challenge employed the segment-based ER, which was calculated by considering 1 second segments across the entire test set. The reported trend, depicted in Figure 23, provides an illustration of the observed changes and patterns. For the baseline of the DCASE 2017 task on SED, a multilayer perception (MLP) architecture was adopted as a neural network. The architecture consisted of two dense layers, each comprising 50 hidden units per layer, with a 20% dropout rate. The features utilized in this baseline approach were log mel-band energies. The baseline achieved a segment-based ER of 0.9358 as demonstrated in Figure 23.

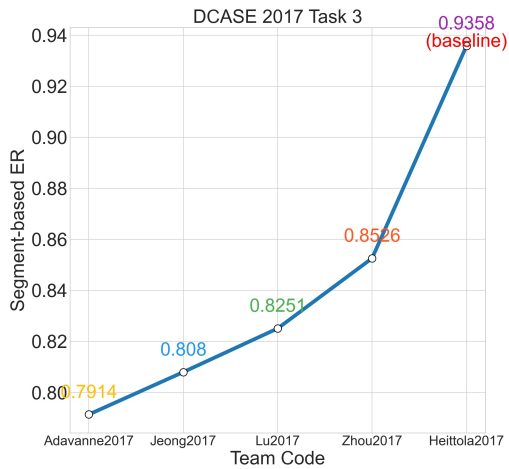

Figure 23: Comparison of system performance among the top teams in DCASE 2017 Task 3 with the baseline.

The top team in 2017 [2] utilized diverse binaural audio features for SED. By processing each feature separately through a stacked combination of convolutional and RNNs, they achieved comparable or improved ERs compared to single-channel features. However, incorporating both audio channels provided a substantial performance boost. The second-ranked team [73] at segment-based ER of 0.808 improved on the baseline with the usage of both short- and long-term audio signals simultaneously as input data as described in Section 3.3.2 in addition to frequent validation with adaptive thresholds and the class-wise early-stopping. The third-ranked system [108] at segment-based ER of 0.8251 continued the usage of multi-label bi-directional GRU as it makes full use of the context information from both directions and explored various data augmentation methods like pitch shift, time stretch, and union deformation. In a similar manner, the fourth-ranked team [163] achieved a performance of 0.8526 by employing LSTM and constructing three distinct channels from the input stereo signals: the right channel, mean channel, and diff channel. Then the team implemented various fusion strategies to integrate the information from these channels.

In this edition of the DCASE challenge, the systems evolved from the top submitted systems of the previous year by incorporating convolutional layers alongside RNNs. Notably, this year witnessed the exploration of additional data augmentation techniques, the adoption of time-of-event feature extraction, and experimentation with multichannel fusion. These advancements signify ongoing efforts to enhance system performance and explore new avenues for improvement.

4.4 DCASE 2018 Task 4

In DCASE 2018 task for SED, the primary metric was switched to an event-based F1-score with a 200 milliseconds collar on onsets and 200 milliseconds / 20% of the events length collar on offsets. This edition's benchmark is established by utilizing two CRNN models, where 64 log mel-band magnitudes serve as the input features. The initial CRNN model comprises three convolution layers and is trained using weak labels, with 20% of the 1,578 clips reserved for validation. This model is employed to predict labels for unlabeled data. Subsequently, the second model is trained using the predictions from the first model, incorporating median filtering to determine the onset and offset of events within each file. This ultimately resulted in an event-based F1-score of 10.8 as reported in Figure 24.

The leading team [74] of this edition's competition accomplished an event-based F1-score of 32.4 by employing a mean-teacher model with a context-gating CRNN. This approach effectively leveraged a substantial volume of unbalanced and unlabeled training data. In the mean-teacher model, the teacher model played a role in utilizing the exponential moving average (EMA)

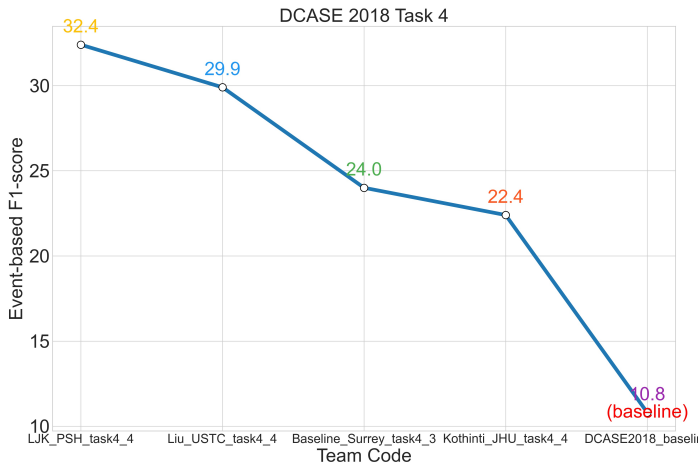

Figure 24: Comparison of system performance among the top teams in DCASE 2018 Task 4 with the baseline.

weights from the student model rather than directly participating in back-propagation. By adopting this technique, the team established a benchmark for future challenges. The team that secured the second position [105] in the rankings with an event-based F1-score of 29.9 enhanced the baseline by converting weak labels into strong labels prior to training. They accomplished this by employing event activity detection, which is based on energy levels as outlined in Section 3.3.2. Furthermore, the team incorporated a capsule-based method and utilized gated convolutional neural networks (CNN) to further improve their approach. At an event-based F1-score of 24, the third-ranked team [87] concentrated their efforts on utilizing CNNs with either 4 layers or 8 layers. The team showcased that the CNN with 8 layers outperformed the CNN with 4 layers, demonstrating the superiority of the deeper architecture in their approach. Lastly, the fourth team [90] came up with an event-based F1-score of 22.4 with a hybrid approach that combines an acoustic-driven event boundary detection with a supervised label inference using a deep neural network based on the baseline.

The significant advancement made in this edition was the introduction of the mean-teacher model approach, which served as the foundation for the following edition's baseline. This contribution led to a substantial improvement in the event-based F1-score, elevating the baseline from 10.8 to 32.4. Additionally, teams incorporated CNNs alongside RNNs to achieve enhancements over the baseline. These improvements involved experimenting with a capsule-based

method and exploring the effectiveness of CNN architectures with either 4 or 8 layers.

4.5 DCASE 2019 Task 4

This edition's evaluation still relied on the event-based F1-score metric. The baseline for this year was derived from the winning submission of DCASE 2018 Task 4 on SED, which resulted in an event-based F1-score of 25.8. The student model's inputs were the same as the inputs of the teacher model but with the addition of Gaussian noise. To ensure consistency between the teacher and student models, a cost for consistency was introduced. Additionally, a median filtering technique, utilizing a window size of 5 frames, was applied to determine the onset and offset of events for each file.

The leading team [103], achieving an event-based F1-score of 42.7 as highlighted in Figure 25, employed a CNN architecture enhanced with an embedding-level attention pooling module for conducting weakly supervised learning. Additionally, to integrate weakly supervised learning with SSL, the team implemented guided learning by utilizing a professional teacher model (PT-model) to guide a more promising student model (PS-model). Lastly, the team introduced a set of median filters with adaptive window sizes specific to different event categories, as detailed in Section 3.4. The team securing the second position [29] accomplished an event-based F1-score of 42.1 by implementing data augmentation techniques such as time-shift, frequency-shift, and noise addition on the baseline architecture. Additionally, the team divided the sound events into three distinct categories and adjusted the median window size for each classified category. The focus of the third-ranked team [138] was primarily on employing SSL methods. They introduced consistency regularization, applied data augmentation techniques, utilized interpolation consistency training (ICT), and implemented mixup regularization to interpolate between data augmentations. These approaches were developed based on the baseline system to enhance their overall performance. With an event-based F1-score of 39.7, the fourth-ranked team [14] utilized multi-task learning to leverage both synthetic and unlabeled subsets within the same domain. Additionally, they placed significant emphasis on employing multiple post-processing methods to further enhance their results.

In this edition of the DCASE challenge, the teams primarily emphasized the utilization of SSL methods to leverage an unlabeled set. The proposed methods were designed to complement the baseline approach, taking advantage of the abundance of easily accessible, unlabeled data. By incorporating these methods, the teams were able to reduce their dependence on manual annotation, leading to cost and effort savings in acquiring labeled data.

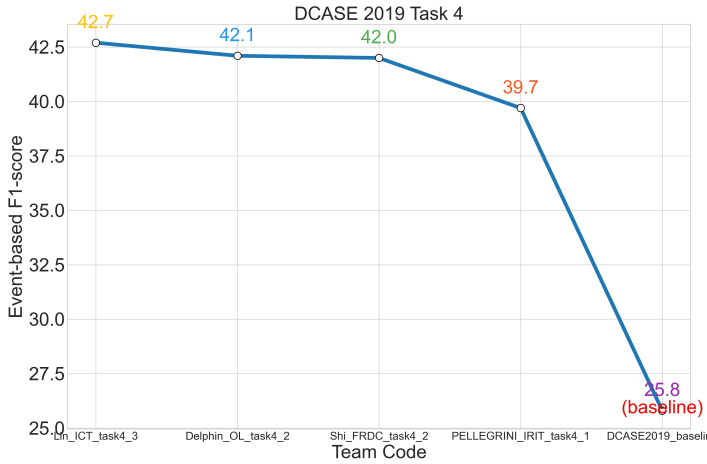

Figure 25: Comparison of system performance among the top teams in DCASE 2019 Task 4 with the baseline.

4.6 DCASE 2020 Task 4

This edition's challenge maintained the use of the event-based F1-score metric as the primary evaluation criterion. Additionally, the challenge introduced the PSDS metric as an optional secondary metric. The baseline approach for this year drew inspiration from the second-best submission of DCASE 2019 Task 4, which was based on the mean-teacher model. Notably, the baseline for 2020 underwent modifications, including changes in the sampling rate, feature extraction hyperparameters, adjustments to the median window, and the introduction of an early stopping mechanism, which resulted in an event-based F1-score of 36.5 illustrated in Figure 26.

The top-performing team, [116], with an event-based F1-score of 51.1 in the challenge, implemented conformer blocks as a replacement for the RNN block in their system. Furthermore, the team conducted experiments with data augmentation techniques and found that time-shifting and mixup techniques yielded positive results for their system. Additionally, the team conducted a post-processing phase as described in Section 3.4, where they searched for optimal threshold values and median filter sizes. The second-ranked team, [58] gained an advantage by addressing the statistical distribution disparity between synthesized and real audio. They achieved this by employing a joint learning approach that combined SED with domain adaptation (DA). With an event-based F1-score of 47.2, the team ranking third [40] introduced a novel

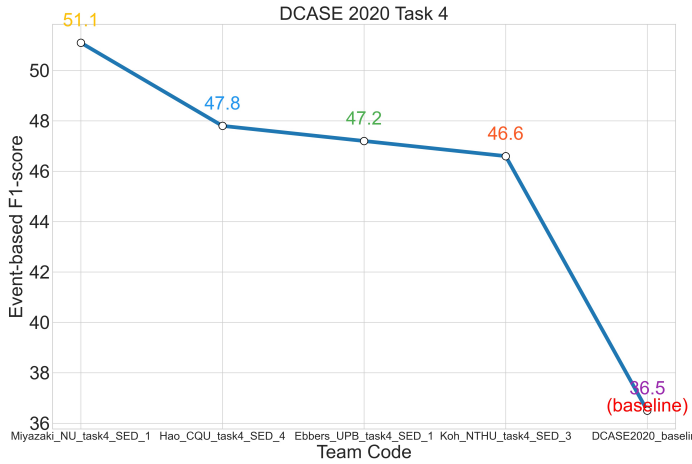

Figure 26: Comparison of system performance among the top teams in DCASE 2020 Task 4 with the baseline.

approach. They proposed a CRNN architecture with two RNNs that share the same preprocessing CNN. Both RNNs were designed to perform audio tagging. Additionally, the team put forward a tag-conditioned CNN, which was trained using pseudo-strong labels, enhancing their overall performance. Lastly, [84] expanded the utilization of information consistency training (ICT) by combining it with shift consistency training (SCT), weakly pseudo-labeling, and their proposed FP-CRNN architecture in Section 3.3.2, to achieve an event-based F1-score of 46.6.

In this edition of the DCASE challenge, participants explored novel architectures to replace the conventional RNN architecture and predominantly adopted a two-stage approach, incorporating the use of pseudo-labeled data in the second stage. The introduction of conformer-based models paved the way for transformer-based architectures, which demonstrated improved capabilities in capturing both global and local context information. Moreover, the proposal to employ domain adaptation highlighted the importance of adapting models to diverse audio sources, leading to enhanced generalization and performance.

4.7 DCASE 2021 Task 4

In this edition of the DCASE challenge, the primary metric was changed to the PSDS metric, which was previously the secondary metric. On the contrary, the event-based F1 score became the secondary metric. The ranking score

was determined by considering the aggregate of these metrics in two different scenarios, as explained in Section 3.5. The baseline used in the DCASE 2020 Task 4 challenge was improved by incorporating mixup for weak and synthetic data, eliminating early stopping, using a different synthetic set, and applying min-max normalization per instance. In this configuration, a total PSDS of 1.11 was achieved, as shown in Figure 27.

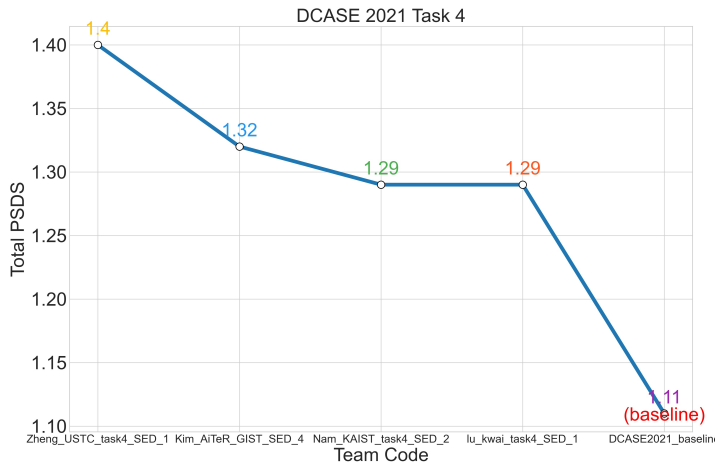

Figure 27: Comparison of system performance among the top teams in DCASE 2021 Task 4 with the baseline.

The top-performing team [162] in this edition of DCASE concentrated on enhancing the localization ability of the CRNN model used in the baseline. They introduced the selective kernel (SK) unit, which is detailed in Section 3.3.2. Moreover, they proposed the utilization of a soft detection output by adjusting the temperature parameter in the sigmoid function, resulting in a notable improvement in the PSDS2 score. Additionally, they continued to incorporate the SCT and ICT techniques from the previous year. The second-ranked team [82] adopted a two-stage architecture similar to the one employed in the DCASE 2020 Task 4. They utilized a residual convolutional recurrent neural network (R-CRNN) in both stages and introduced a self-training-based noisy student model that incorporated feature noises in the second stage. The third-ranked team [120] devoted their efforts to extensive data augmentation, leading to the introduction of a novel technique called filter augmentation. They also proposed two methods to utilize weak predictions of the model: weak prediction masking and weak SED. The combination of these techniques resulted in a significant increase in the PSDS2 score, although there was a

slight degradation in PSDS1. The fourth-ranked team [109] continued to utilize a conformer-based architecture in addition to the CRNN architecture. They conducted experiments with various data augmentation techniques and conducted a search to determine the optimal median window size for each class within the range of 1 to 49.

In this edition, there was a notable trend toward increased utilization of diverse data augmentation techniques, with one team introducing a unique approach known as filter augmentation. This method involved assigning different weights to random frequency regions, aiming to replicate various acoustic conditions. Furthermore, significant developments were made in improving the CNN architecture, particularly in enhancing its localization ability. One noteworthy advancement was the proposal of a residual network, which effectively mitigates overfitting to the training data. The teams also demonstrated that it is relatively easier to improve the PSDS2 score compared to PSDS1. They achieved this by leveraging temperature parameters within the sigmoid function and employing weak training methods.

4.8 DCASE 2022 Task 4

Similar to the previous edition, the PSDS metric was used to assess the systems in this edition of the DCASE challenge. The baseline approach remained consistent with the one employed in DCASE 2021 Task 4. However, a new aspect was introduced to examine the influence of external data. Participants were given the freedom to utilize external data and pretrained models. By incorporating the audioset external set, the baseline achieved a total PSDS score of 1.04 on the evaluation set, as depicted in Figure 28.

The top team [41] utilized forward-backward convolutional recurrent neural networks (FB-CRNNs) for weakly labeled and semi-supervised SED. They generated strong pseudo labels for weakly labeled and unlabeled data and trained (tag-conditioned) bi-directional CRNNs (BiCRNNs) in a strongly supervised manner. Through multiple iterations of self-training, they achieved an impressive total PSDS score of 1.63. The second-ranked team [62] combined multiple strategies to achieve their position and a total PSDS score of 1.57. They utilized ICT, SCT, and filter augmentation techniques. Further, they also investigated the relationship between audioset labels and the target acoustic events. Architectures like SK-CRNN and FDY-CRNN were employed, and pretrained models were used as embeddings to enhance their model's performance. With a total PSDS score of 1.49, the third-ranked team [154] utilized the weak prediction method from the previous year. They employed a fusion approach, combining multiple pretrained models such as PANNs and SSAST, to maximize the utilization of external data and leverage the available information effectively. The team ranked fourth [141] at 1.47, and employed FDY-CRNN as their primary architecture, incorporating a data

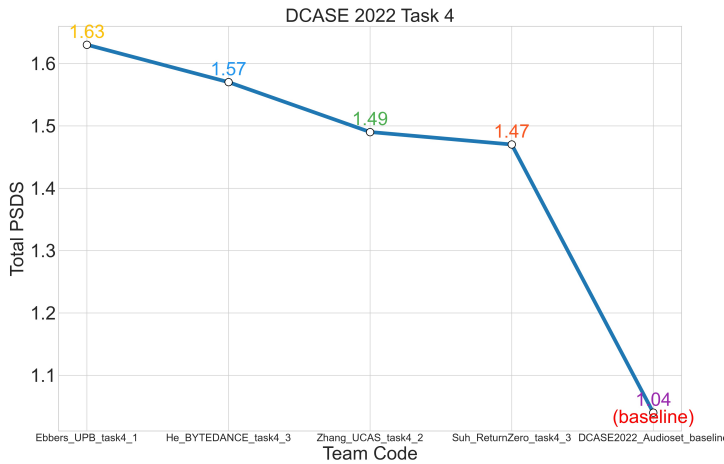

Figure 28: Comparison of system performance among the top teams in DCASE 2022 Task 4 with the baseline.

augmentation pipeline to enhance performance. They made use of a pretrained model to select a specific subset of audioset. To address the data imbalance between active and inactive frames, they utilized asymmetric focal loss (AFL) [71]. Additionally, the team implemented several post-processing techniques, including the use of temperature parameters in the sigmoid function, tuning the median window length, and leveraging the weak SED method from the previous year.

In summary, the 2022 edition involved testing different pretrained models combined with attention-based CRNN architectures to gather pertinent information and adapt to inputs of varying lengths. Additionally, teams explored methods of selecting a subset of the audioset, which contributed to enhanced performance when using an external set. In the 2022 edition of the DCASE challenge, the AFL experiments demonstrated that assigning higher weights to the minority class (sound events) incentivized the model to prioritize the accurate detection of these events. Lastly, the integration of all the proposed advancements from previous challenge years was observed to enhance the overall system.

4.9 DCASE 2023 Task 4A

In this edition, the evaluation metric underwent an update to include a threshold-independent implementation of the PSDS. Additionally, it became

mandatory to include energy consumption data in the reports. A new baseline was introduced this year, which utilizes the pretrained model BEATs. In this established baseline, the frame-level embeddings from BEATs are combined with the existing CRNN baseline classifier using a late-fusion approach. To align the temporal resolution of the frame-level embeddings with the CNN output, adaptive average pooling is applied. By incorporating the pretrained model BEATs, this baseline achieved a total t-PSDS score of 1.52, as depicted in Figure 29.

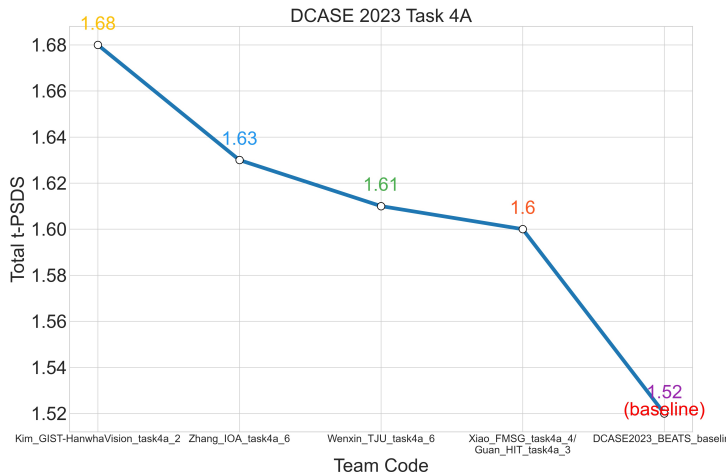

Figure 29: Comparison of system performance among the top teams in DCASE 2023 Task 4A with the baseline.

The team ranked first in the DCASE 2023 Task 4A [80] achieved a total t-PSDS score of 1.68 by implementing a technique called LKA within the FDY method. This approach replaced the conventional convolutions used in the CRNN baseline. By making this modification, the team effectively captured patterns in the time-frequency domain, long-term relationships, and meaningful information in audio signals. Their proposed architecture followed a two-stage training process. In the initial stage, they generated robust pseudo-labels for weakly labeled, unlabeled, and Audioset datasets, which were then used in the second stage of training. The team ranked second [155], with a total t-PSDS score of 1.63, adopted the energy difference-based log-mel spectrogram to enhance the representation of audio features. This method assigned higher weights to frames with sound events and lower weights to frames without sound events. Additionally, they utilized multi-dimensional frequency dynamic convolution (MFDConv) to improve the feature extraction capability of con-

volutional kernels. Furthermore, the team employed a confidence-weighted binary cross-entropy (BCE) loss function to address the challenge of detecting certain short classes. The team ranked third [39], achieving a total t-PSDS score of 1.61, employed the FDY-CRNN architecture, and introduced a novel approach called mutual mean teaching (MMT), which involved collaboratively training two identical networks with different initialization to generate soft pseudo-labels. They also utilized several data augmentation methods, such as ICT and SCT, in combination with various pretrained models. With two teams ranked fourth [156] and [55], both closely followed the third team with a total t-PSDS score of 1.60. One of those teams [156] also utilized the FDY-CRNN architecture. In addition, they incorporated BEATs embeddings, which were developed using a self-curated dataset from Audioset. Furthermore, the team [156] employed numerous aggregation approaches to leverage the strengths of different techniques. They also implemented the AFL function, which adjusted the training weights based on the model's training difficulty, as adopted from previous years. The other team [55], which ranked fourth as well also achieved a total t-PSDS score of 1.60 by incorporating sound activity detection (SAD) as an auxiliary task and training SAD and SED in an MTL framework. The inclusion of SAD aimed to enhance the performance of SED by effectively detecting event boundaries.

In this particular iteration of the challenge, it is evident that the majority of teams opted for a variation of FDY-CRNN combined with a pretrained model like BEATs. Additionally, there was a noticeable trend of incorporating modified loss functions, such as confidence-weighted loss and asymmetrical loss, to assign weights to individual frames. Many teams also made use of various data augmentation techniques and drew inspiration from SSL methods employed in previous editions of the challenge.

5 Future Horizons

In the preceding section, we examined the prominent systems developed each year as summarized in Figure 30 and observed a discernible pattern in their progression. Initially, the baselines employed conventional machine learning methods like NMF and GMM, but later, the focus shifted toward utilizing RNNs to capture temporal dependencies. This was followed by the integration of convolutional layers with RNNs. Notably, the introduction of the mean-teacher model, which exploited the unlabeled dataset, resulted in enhanced performance of SED systems. Teams also explored various CNN architectures to accompany the RNNs. Moreover, the adoption of SSL based on the mean-teacher model gained further attention. This made teams propose alternatives to conventional RNN architectures, which led to the emergence of transformer-based architectures in subsequent years. Alongside these developments, teams

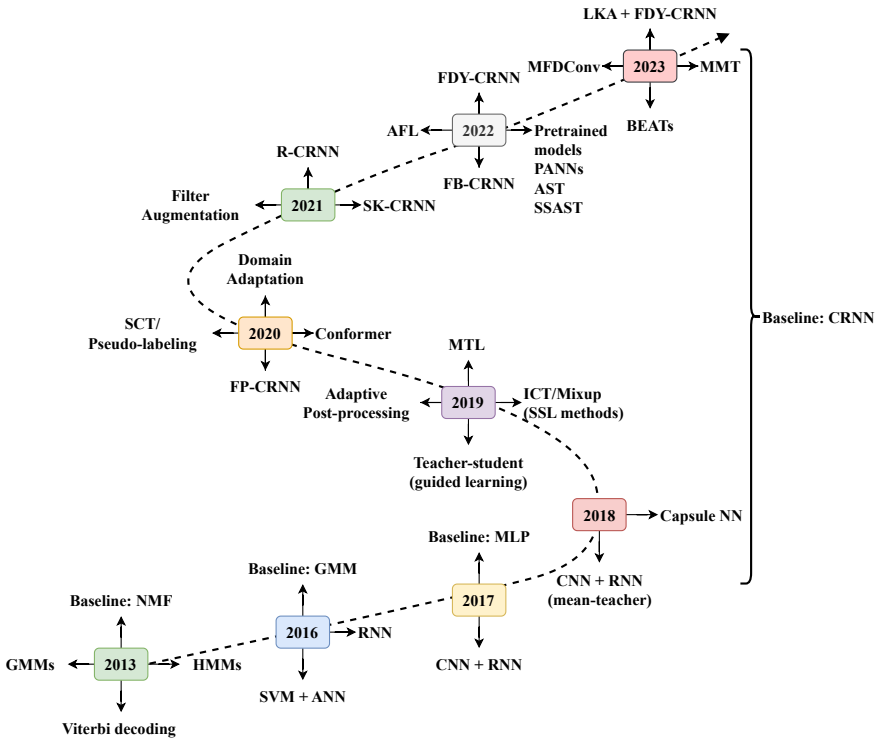

Figure 30: Summarized timeline of various approaches introduced throughout the challenge series.

extensively experimented with diverse data augmentation techniques and introduced novel methods to replicate acoustic variations. The introduction of transformer-based architectures also prompted a greater emphasis on improving the localization capability of CNN architectures. This led to the proposal of new models such as FDY-CRNN, SK-CRNN, and FB-CRNN. Additionally, teams incorporated pretrained models to extract embeddings and combined them with the proposed architectures to enhance the overall performance of SED systems.

Given that most SED systems rely on labeled datasets to enhance their performance, the previously described subtask of the challenge placed a significant emphasis on SSL methods to mitigate the requirement for extensive labeled data. As a result, a new subtask, 4B, was introduced as part of the DCASE 2023 Task 4. This specific subtask focused on utilizing soft labels with a temporal resolution of 1 second. Training on such datasets helped accommodate human uncertainty regarding categories or provided a natural

representation of diverse opinions during annotation. Hence, this subtask opened up new avenues for future research in the field of SED. Additionally, we present a few other directions for further research, as depicted in Figure 31 and outlined below:

- Low complexity models: The field of SED has undergone the emergence of new models with a large number of parameters. As a result, there is a growing demand to address practical applications, where limited resources or the need for real-time processing pose challenges to deploying complex and computationally intensive SED systems. In this research field, it is crucial to create models that minimize memory usage and computational operations while maintaining the performance of SED. Some approaches to developing lightweight SED models include techniques such as model compression, pruning, quantization, and network architecture design.

- Low energy consumption: Due to the current global environmental crisis, there is a growing impetus to create energy-efficient technologies. In the field of upcoming SED model research, focusing on reducing energy consumption would help decrease the overall demand for energy, resulting in reduced carbon emissions and a more sustainable future. Moreover, the development of energy-efficient models would extend battery life, allowing users to utilize their devices for longer durations without the need for frequent recharging.

- Zero-shot learning: The SED domain necessitates a significant amount of labeled data to train the system. However, this approach typically incurs a high cost for annotations and also is time-consuming. Therefore, investigating zero-shot learning would enable the model to scale and adapt to identifying unfamiliar sound events without explicit training. By utilizing the semantic connections between known sound event categories and their associated attributes, the model can effectively detect novel or uncommon sound events. Consequently, this approach would decrease the effort required for annotations and enhance the model's ability to adjust to new environments, providing greater flexibility.

- Real-world evaluation metrics: The conventional evaluation metrics, such as the F1-score, frequently prove inadequate in comprehensively addressing the intricacies involved in real-life situations. Consequently, there is a requirement to establish novel metrics that focus on quantifying a system's resilience to environmental fluctuations, the temporal precision of real-time clip localization for a more thorough understanding of the system's ability to identify sound events in real-time, and enhancing the measurement of computational and scalability efficiency through evaluation metrics.

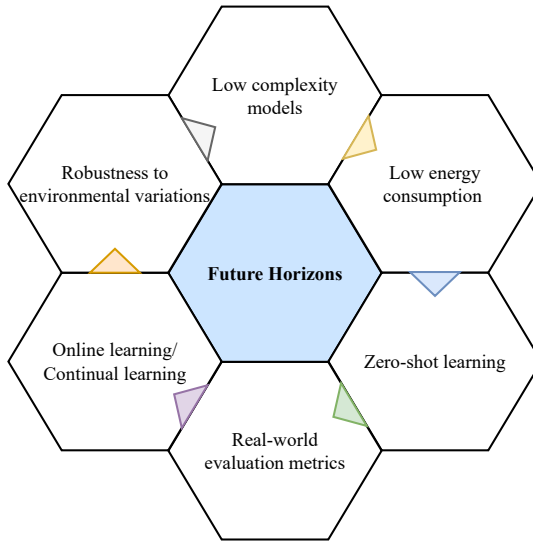

Figure 31: Summarized future horizons for SED.

- Online learning/ Continual learning: In conventional offline scenarios, models undergo training with static datasets. However, in the real world, the environment is constantly changing and diverse, which necessitates the ability to adapt to these variations. Consequently, there arises a need for the model to autonomously update itself when new data is introduced. Furthermore, it becomes essential to seamlessly incorporate new sound categories into the detection system without the need for a complete retraining of the model. Consequently, research in this domain would facilitate the gradual enhancement of the model's knowledge.

- Robustness to environmental variations: An additional essential aspect to consider is the model's capacity to cope with changes in the environment. This field of study could involve investigating data augmentation methods and developing innovative architectures that can adjust to varying acoustic conditions. Lastly, an approach worth considering is domain adaptation, which aims to minimize the disparities between training and target acoustic domains by devising suitable techniques for model adaptation.

6 Conclusion

This work presents an overview of the contributions made to the DCASE challenge series in the field of SED. We begin by examining the problem formulation and applications of SED. We then delve into the evolution of the DCASE challenge, including changes in the dataset, feature extraction, and modeling approaches. Furthermore, we discuss the progression of post-processing techniques employed in the challenge to enhance the accuracy of detected events. We provide a comprehensive description of the evolving evaluation metric used to measure results and highlight the advantages of each iteration over its predecessor. Using this evaluation metric, we conduct a detailed analysis of the top-performing teams in each challenge edition. Lastly, we explore potential future directions for advancements in SED from the view of real-world applications.

References

[1] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features", *tech. rep.*, DCASE 2016 Challenge, 2016.

[2] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features", *tech. rep.*, DCASE 2017 Challenge, 2017.

[3] J. B. Allen, "Cochlear modeling", *IEEE Acoustics, Speech, and Signal Magazine*, 2(1), 1985, 3–29.

[4] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C. H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1", *IEEE Signal Processing Magazine.*, 26(3), 2009, 75–80.

[5] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution", *Communications of the ACM*, 2019, 68–77.

[6] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities", *Computational Analysis of Sound Scenes and Events*, 2018, 373–97.

[7] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 61–5.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers", *Workshop on Computational Learning Theory*, 1992, 144–52.

[9] A. Bregman, *Auditory scene analysis: The perceptual organization of sound*, The MIT Press, 1990.

[10] X. Cai and H. Dinkel, "A lightweight approach for semi supervised sound event detection with unsupervised data augmentation", *Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, 2021, 35–9.

[11] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6), 2017, 1291–303.

[12] E. Cakir and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection", *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, 27–31.

[13] E. Çakir, "Deep neural networks for sound event detection", *PhD thesis*, Tampere University, 2019.

[14] L. Cances, T. Pellegrini, and P. Guyot, "Multi-task learning and post-processing optimization for sound event detection", *tech. rep.*, DCASE 2019 Challenge, 2019.

[15] L. Cances, T. Pellegrini, and P. Guyot, "Sound event detection from weak annotations: Weighted-GRU versus multi-instance-learning", *Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, 2018, 64–8.

[16] T. K. Chan and C. S. Chin, "A comprehensive review of polyphonic sound event detection", *IEEE Access*, 8, 2020, 103339–73.

[17] T. K. Chan, C. S. Chin, and Y. Li, "Non-negative matrix factorization convolutional neural network (NMF-CNN) for sound event detection", *Detection and Classification of Acoustic Scenes and Events 2019 Workshop*, 2019, 40–4.

[18] T. K. Chan, C. S. Chin, and Y. Li, "Semi-supervised NMF-CNN for sound event detection", *tech. rep.*, DCASE 2020 Challenge, 2020.

[19] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, 646–50.

[20] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers", *International Conference on Machine Learning (ICML)*, 2022, 1–16.

[21] W.-Y. Chen, C.-L. Lu, H.-F. Chuang, Y.-H. C. Cheng, and B.-C. Chan, "Sound event detection system using pre-trained model for DCASE 2023 Task 4", *tech. rep.*, DCASE 2023 Challenge, 2023.

[22] Y. Chen, Y. Zhang, and Z. Duan, "DCASE 2017 sound event detection using convolutional neural network", *tech. rep.*, DCASE 2017 Challenge, 2017.

[23] I. Choi, K. Kwon, S. H. Bae, and N. S. Kim, "DNN-based sound event detection with exemplar-based approach for noise reduction", *Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, 2016, 16–9.

[24] W.-G. Choi and J.-H. Chang, "Confidence regularized entropy for polyphonic sound event detection", *Detection and Classification of Acoustic Scenes and Events 2022 Workshop*, 2022.

[25] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. Mataric, "Where am I? Scene recognition for mobile robots using audio features", *IEEE International Conference on Multimedia and Expo (ICME)*, 2006, 885–8.

[26] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *Readings in Speech Recognition*, 1990, 65–74.

[27] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior", *IEEE Signal Processing Magazine*, 2016, 81–94.

[28] G. Dekkers, S. Lauwereins, B. Thoen, M. Adhana, H. Brouckxon, B. V. den Bergh, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network", *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, 1–5.

[29] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4", *tech. rep.*, DCASE 2019 Challenge, 2019.

[30] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions", *IEEE Signal Processing Letters*, 18(2), 2011, 130–3.

[31] J. Dennis, H. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform", *Pattern Recognition Letters*, 34(9), 2013, 1085–93.

[32] J. Dennis, T. Dat, and E. S. Chng, "Analysis of spectrogram image methods for sound event classification", *Interspeech*, 2014, 2533–7.

[33] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge", *tech. rep.*, DCASE 2013 Challenge, 2013.

[34] H. Dinkel, Z. Yan, Y. Wang, M. Song, J. Zhang, and W. Wang, "A large multi-modal ensemble for sound event detection", *tech. rep.*, DCASE 2022 Challenge, 2022.

[35] M. Dobreva-McPherson, Y. Kim, and S. Ross, "Automated metadata generation", *Digital Curation Reference Manual*, 2013.

[36] K. Dohi, K. Imoto, H. Noboru, and N. Daisuke, "Description and discussion on DCASE 2023 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring", *tech. rep.*, DCASE 2023 Challenge, 2023.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", *International Conference on Learning Representations (ICLR)*, 2021.

[38] K. Drossos, S. Gharib, P. Magron, and T. Virtanen, "Language modelling for sound event detection with teacher forcing and scheduled sampling", *Detection and Classification of Acoustic Scenes and Events 2019 Workshop*, 2019, 59–63.

[39] X. Duo Wenxin1 Fang and J. Li, "Semi-supervised sound event detection system for DCASE 2023 Task 4A", *tech. rep.*, DCASE 2023 Challenge, 2023.

[40] J. Ebbers and R. Haeb-Umbach, "Forward-backward convolutional recurrent neural networks and tag-conditioned convolutional neural networks for weakly labeled semi-supervised sound event detection", *Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, 2020, 41–5.

[41] J. Ebbers and R. Haeb-Umbach, "Pre-training and self-training for sound event detection in domestic environments", *tech. rep.*, DCASE 2022 Challenge, 2022.

[42] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, 1021–5.

[43] R. Ebbers Janek Haeb-Umbach, "Self-trained audio tagging and sound event detection in domestic environments", *tech. rep.*, DCASE 2021 Challenge, 2021.

[44] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE Challenge 2016: Acoustic scene classification and sound event detection in real life recording", *tech. rep.*, DCASE 2016 Challenge, 2016.

[45] J. L. Elman, "Finding structure in time", *Cognitive Science*, 14(2), 1990, 179–211.

[46] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2022, 829–52.

[47] Y. Gan, Z. Qiao, J. Wu, X. Cai, and M. Wu, "Semi-supervised sound event detection based on pretrained models for DCASE 2023 Task 4A", *tech. rep.*, DCASE 2023 Challenge, 2023.

[48] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, 1–4.

[49] J. F. Gemmeke, L. Vuegen, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach for audio event detection", *tech. rep.*, DCASE 2013 Challenge, 2013.

[50] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, 776–80.

[51] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection", *European Signal Processing Conference (EUSIPCO)*, 2013, 1–5.

[52] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, 47(1-2), 1990, 103–38.

[53] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer", *Interspeech*, 2021, 571–5.

[54] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer", *AAAI Conference on Artificial Intelligence*, 36(10), 2022, 10699–709.

[55] Y. Guan and Q. Shang, "Semi-supervised sound event detection system for DCASE 2023 Task 4", *tech. rep.*, DCASE 2023 Challenge, 2023.

[56] Y. Guo, M. Xu, i. Wu, Y. Wang, and K. Hoashi, "Multi-scale convolutional recurrent neural network with ensemble method for weakly labeled sound event detection", *Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, 2018, 98–102.

[57] J. M. Gutiérrez-Arriola, R. Fraile, A. Camacho, T. Durand, J. L. Jarrín, and S. R. Mendoza, "Synthetic sound event detection based on MFCC", *Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, 2016, 30–4.

[58] J. Hao, Z. Hou, and W. Peng, "Cross-domain sound event detection: From synthesized audio to real audio", *tech. rep.*, DCASE 2020 Challenge, 2020.

[59] R. Harb and F. Pernkopf, "Sound event detection using weakly labelled semi-supervised data with GCRNNs, VAT and self-adaptive label refinement", *Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, 2018, 83–7.

[60] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Bi-directional LSTM-HMM hybrid system for polyphonic sound event detection", *Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, 2016, 35–9.

[61] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, "BLSTM-HMM hybrid system combined with sound activity detection network for polyphonic sound event detection", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, 766–70.

[62] K. He, X. Shu, S. Jia, and Y. He, "Semi-supervised sound event detection system for DCASE 2022 Task 4", *tech. rep.*, DCASE 2022 Challenge, 2022.

[63] T. Heittola, E. Çakır, and T. Virtanen, "The machine learning approach for analysis of sound scenes and events", *Computational Analysis of Sound Scenes and Events*, 2018, 13–40.

[64] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection", *EURASIP Journal on Audio, Speech, and Music Processing*, 2013, 1.

[65] T. Heittola, A. Mesaros, and T. Virtanen, "DCASE 2016 baseline system", *tech. rep.*, DCASE 2016 Challenge, 2016.

[66] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation", *Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011, 36–40.

[67] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition", *IEEE Signal Processing Magazine*, 29(6), 2012, 82–97.

[68] Y. Hou and S. Li, "Semi-supervised sound event detection with convolutional recurrent neural network using weakly labelled data", *tech. rep.*, DCASE 2018 Challenge, 2018.

[69] Y. Hou and S. Li, "Sound event detection in real life audio using multi-model system", *tech. rep.*, DCASE 2017 Challenge, 2017.

[70] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 7132–41.

[71] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of sound duration and inactive frames on sound event detection performance", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 860–4.

[72] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multi-task learning of sound events and scenes with soft scene labels", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 621–5.

[73] I.-Y. Jeong, S. Lee, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network", *tech. rep.*, DCASE 2017 Challenge, 2017.

[74] L. JiaKai, "Mean teacher convolution system for DCASE 2018 Task 4", *tech. rep.*, DCASE 2018 Challenge, 2018.

[75] J. Jie, "Anomalous sound detection based on self-supervised learning", *tech. rep.*, DCASE 2023 Challenge, 2023.

[76] T. Khandelwal and R. K. Das, "A multi-task learning framework for sound event detection using high-level acoustic characteristics of sounds", *Interspeech*, 2023.

[77] T. Khandelwal and R. K. Das, "Dynamic thresholding on FixMatch with weak and strong data augmentations for sound event detection", *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, 428–32.

[78] T. Khandelwal, R. K. Das, A. Koh, and E. S. Chng, "FMSG-NTU submission for DCASE 2022 Task 4 on sound event detection in domestic environments", *tech. rep.*, DCASE 2022 Challenge, 2022.

[79] T. Khandelwal, R. K. Das, A. Koh, and E. S. Chng, "Leveraging audio-tagging assisted sound event detection using weakified strong labels and frequency dynamic convolutions", *IEEE Statistical Signal Processing Workshop*, 2023.

[80] J. W. Kim, S. W. Son, Y. Song, H. K. Kim, I. H. Song, and J. E. Lim, "Semi-supervised learning-based sound event detection using frequency dynamic convolution with large kernel attention for DCASE Challenge 2023 Task 4", *tech. rep.*, DCASE 2023 Challenge, 2023.

[81] N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on convolutional recurrent neural networks with semi-supervised loss function for DCASE Challenge 2020 Task 4", *tech. rep.*, DCASE 2020 Challenge, 2020.

[82] N. K. Kim and H. K. Kim, "Self-training with noisy student model and semi-supervised loss function for DCASE 2021 Challenge Task 4", *tech. rep.*, DCASE 2021 Challenge, 2021.

[83] Y. Kiyokawa, S. Mishima, T. Toizumi, K. Sagi, R. Kondo, and T. Nomura, "Sound event detection with ResNet and self-mask module for DCASE 2019 Task 4", *tech. rep.*, DCASE 2019 Challenge, 2019.

[84] C.-Y. Koh, Y.-S. Chen, S.-E. Li, Y.-W. Liu, J.-T. Chien, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks", *tech. rep.*, DCASE 2020 Challenge, 2020.

[85] Q. Kong, Y. Cao, T. Iqbal, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems", *tech. rep.*, DCASE 2019 Challenge, 2019.

[86] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, 2880–94.

[87] Q. Kong, I. Turab, X. Yong, W. Wang, and M. D. Plumbley, "DCASE 2018 Challenge baseline with convolutional neural networks", *tech. rep.*, DCASE 2018 Challenge, 2018.

[88] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time-frequency segmentation from weakly labelled data", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4), 2019, 777–87.

[89] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint separation-classification model for sound event detection of weakly labeled data", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, 321–5.

[90] S. Kothinti, K. Imoto, D. Chakrabarty, S. Gregory, S. Watanabe, and M. Elhilali, "Joint acoustic and class inference for weakly supervised sound event detection", *tech. rep.*, DCASE 2018 Challenge, 2018.

[91] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout", *Interspeech*, 2022, 2753–7.

[92] S. Krstulović, "Audio event recognition in the smart home", *Computational Analysis of Sound Scenes and Events*, 2018, 335–71.

[93] Y.-H. Lai, C.-H. Wang, S.-Y. Hou, B.-Y. Chen, Y. Tsao, and Y.-W. Liu, "DCASE report for Task 3: Sound event detection in real life audio", *tech. rep.*, DCASE 2016 Challenge, 2016.

[94] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, 401(6755), 1999, 788–91.

[95] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input", *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, 74–9.

[96] J. Lee, J. Park, S. Kum, Y. Jeong, and J. Nam, "Combining multi-scale features using sample-level deep convolutional neural networks for weakly supervised sound event detection", *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, 69–73.

[97] K. Li, P. Cai, and Y. Song, "Li USTC team's submission for DCASE 2023 Challenge Task 4A", *tech. rep.*, DCASE 2023 Challenge, 2023.

[98] K. Li, X. Zheng, and Y. Song, "A two-stage training method for DCASE 2022 Challenge Task 4", *tech. rep.*, DCASE 2022 Challenge, 2022.

[99] Y. Li, Z. Guo, Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi, "A hybrid system of sound event detection transformer and frame-wise model for DCASE 2022 Task 4", *Detection and Classification of Acoustic Scenes and Events 2022 Workshop*, 2022.

[100] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks", *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, 80–4.

[101] W. Lim, S. Suh, and Y. Jeong, "Weakly labeled semi-supervised sound event detection using CRNN with inception module", *Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, 2018, 74–7.

[102] W. Lim, S. Suh, S. Park, and Y. Jeong, "Sound event detection in domestic environments using ensemble of convolutional recurrent neural networks", *tech. rep.*, DCASE 2019 Challenge, 2019.

[103] L. Lin and X. Wang, "Guided learning convolution system for DCASE 2019 Task 4", *tech. rep.*, DCASE 2019 Challenge, 2019.

[104] C.-C. Liu, T.-H. Kuo, C.-P. Chen, C.-L. Lu, B.-C. Chan, Y.-H. Cheng, and H.-F. Chuang, "CHT+NSYSU sound event detection system with pretrained embeddings extracted from beats model for DCASE 2023 Task 4", *tech. rep.*, DCASE 2023 Challenge, 2023.

[105] Y. L. Liu, J. Yan, Y. Song, and J. Du, "USTC-NELSLIP system for DCASE 2018 Challenge Task 4", *tech. rep.*, DCASE 2018 Challenge, 2018.

[106] Y. Liu, J. Tang, Y. Song, and L. Dai, "A capsule-based approach for polyphonic sound event detection", *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, 2018, 1853–7.

[107] Y. Liu, C. Chen, J. Kuang, and P. Zhang, "Semi-supervised sound event detection based on mean teacher with power pooling and data augmentation", *tech. rep.*, DCASE 2020 Challenge, 2020.

[108] R. Lu and Z. Duan, "Bidirectional GRU for sound event detection", *tech. rep.*, DCASE 2017 Challenge, 2017.

[109] R. Lu, W. Hu, D. Zhiyao, and J. Liu, "Integrating advantages of recurrent and transformer structures for sound event detection in multiple scenarios", *tech. rep.*, DCASE 2021 Challenge, 2021.

[110] P. Mayorga, D. Ibarra, V. Zeljkovic, and C. Druzgalski, "Quartiles and mel frequency cepstral coefficients vectors in hidden Markov-Gaussian mixture models classification of merged heart sounds and lung sounds signals", *International Conference High-Performance Computation & Simulation (HPCS)*, 2015, 298–304.

[111] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 Challenge setup: Tasks, datasets and baseline system", *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, 85–92.

[112] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings", *European Signal Processing Conference (EUSIPCO)*, 2010, 1267–71.

[113] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection", *Applied Sciences*, 6(6), 2016, 162.

[114] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection", *European Signal Processing Conference (EUSIPCO)*, 2016, 1128–32.

[115] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial", *IEEE Signal Processing Magazine*, 38(5), 2021, 67–83.

[116] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation", *Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, 2020, 100–4.

[117] S. Mizobuchi, H. Ohashi, A. Izumi, and N. Kodama, "Mizobuchi PCO team's submission for DCASE 2022 Task 4 - Sound event detection using external resources", *tech. rep.*, DCASE 2022 Challenge, 2022.

[118] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection", *Interspeech*, 2022, 2763–7.

[119] H. Nam, S.-H. Kim, D. Min, B.-Y. Ko, S.-D. Choi, and Y.-H. Park, "Frequency dependent sound event detection for DCASE 2022 Challenge Task 4", *tech. rep.*, DCASE 2022 Challenge, 2022.

[120] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions", *tech. rep.*, DCASE 2021 Challenge, 2021.

[121] M. Nandwana and T. Hasan, "Towards smart-cars that can listen: Abnormal acoustic event detection on the road", *Interspeech*, 2016, 2968–71.

[122] A. Nasiri, "Deep learning based sound event detection and classification", *PhD thesis*, University of South Carolina, 2021.

[123] T. N. T. Nguyen, D. L. Jones, and W. S. Gan, "On the effectiveness of spatial and multi-channel features for multi-channel polyphonic sound event detection", *Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, 2020, 115–9.

[124] T. N. T. Nguyen, N. K. Nguyen, H. Phan, L. Pham, K. Ooi, D. L. Jones, and W.-S. Gan, "A general network architecture for sound event localization and detection using transfer learning and recurrent neural network", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 935–9.

[125] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2022, 1749–62.

[126] M. E. Niessen, T. L. M. Van Kasteren, and A. Merentitis, "Hierarchical sound event detection", *tech. rep.*, DCASE 2013 Challenge, 2013.

[127] W. Nogueira, G. Roma, and P. Herrera, "Automatic event classification using front end single channel noise reduction, MFCC features and a support vector machine classifier", *tech. rep.*, DCASE 2013 Challenge, 2013.

[128] P. J. Pereira, G. Coelho, A. Ribeiro, L. M. Matos, E. C. Nunes, A. Ferreira, A. Pilastri, and P. Cortez, "Using deep autoencoders for in-vehicle audio anomaly detection", *Procedia Computer Science*, 192, 2021, 298–307.

[129] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks", *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020, 763–72.

[130] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, 77(2), 1989, 257–86.

[131] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, 158–61.

[132] B. Raj and A. Kumar, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data", *International Joint Conference on Neural Networks (IJCNN)*, 2017, 3475–82.

[133] M. Rajapakse and L. Wyse, "Generic audio classification using a hybrid model based on GMMs and HMMs", *International Multimedia Modelling Conference*, 2005, 1–6.

[134] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications", *Workshop on Mobile Computing Systems and Applications*, 1994, 85–90.

[135] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction", *tech. rep.*, DCASE 2013 Challenge, 2013.

[136] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments", *Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, 2018, 19–23.

[137] N. Shao, X. Li, and X. Li, "ATST self-supervised plus RCT semi-supervised sound event detection: Submission to DCASE 2022 Challenge Task 4", *tech. rep.*, DCASE 2022 Challenge, 2022.

[138] Z. Shi, "HODGEPODGE: Sound event detection based on ensemble of semi-supervised learning methods", *tech. rep.*, DCASE 2019 Challenge, 2019.

[139] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events", *IEEE Transactions on Multimedia*, 17(10), 2015, 1733–46.

[140] D. Stowell, "Computational bioacoustic scene analysis", *Computational Analysis of Sound Scenes and Events*, 2018, 303–33.

[141] S. Suh and D. Y. Lee, "Data engineering for noisy student model in sound event detection", *tech. rep.*, DCASE 2022 Challenge, 2022.

[142] N. W. Z. Terence, T. H. Dat, H. T. Hoa, and E. S. Chng, "Adaptive semi-supervised tree SVM for sound event recognition in home environments", *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, 2013, 1–4.

[143] N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis", *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2019, 253–7.

[144] F. Vesperini, D. Droghini, E. Principi, L. Gabrielli, and S. Squartini, "Hierarchic conv nets framework for rare sound event detection", *European Signal Processing Conference (EUSIPCO)*, 2018, 1497–501.

[145] T. H. Vu and J.-C. Wang, "Acoustic scene and event recognition using recurrent neural networks", *tech. rep.*, DCASE 2016 Challenge, 2016.

[146] L. Vuegen, B. Van Den Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. Van hamme, "An MFCC-GMM approach for event detection and classification", *tech. rep.*, DCASE 2013 Challenge, 2013.

[147] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE Press, 2006.

[148] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation", *IEEE Transactions on Automation Science and Engineering*, 5(1), 2008, 25–31.

[149] Y. W. Wang, C.-P. Chen, C.-L. Lu, and B.-C. Chan, "Semi-supervised sound event detection using multiscale channel attention and multiple consistency training", *Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, 2021, 75–9.

[150] Y. Wang, H. Dinkel, Z. Yan, J. Zhang, and Y. Wang, "PEPE: Plain efficient pretrained embeddings for sound event detection", *tech. rep.*, DCASE 2023 Challenge, 2023.

[151] W. Xia and K. Koishida, "Sound event detection in multichannel audio using convolutional time-frequency-channel squeeze and excitation", *Interspeech*, 2019, 3629–33.

[152] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Class wise distance based acoustic event detection", *tech. rep.*, DCASE 2017 Challenge, 2017.

[153] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Random forest classification based acoustic event detection", *IEEE International Conference on Multimedia and Expo (ICME)*, 2017, 163–8.

[154] S. Xiao, "Pretrained models in sound event detection for DCASE 2022 Challenge Task 4", *tech. rep.*, DCASE 2022 Challenge, 2022.

[155] S. Xiao, J. Shen, A. Hu, X. Zhang, P. Zhang, and Y. Yan, "Sound event detection with weak prediction for DCASE 2023 Challenge Task 4A", *tech. rep.*, DCASE 2023 Challenge, 2023.

[156] Y. Xiao, T. Khandelwal, and R. K. Das, "FMSG submission for DCASE 2023 Challenge Task 4 on sound event detection with weak labels and synthetic soundscapes", *tech. rep.*, DCASE 2023 Challenge, 2023.

[157] R. Xie, C. Shi, L. Zhang, and H. Li, "Semi-supervised sound event detection using pretrained model", *tech. rep.*, DCASE 2022 Challenge, 2022.

[158] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. G. Okuno, "Environmental sound recognition for robot audition using matching-pursuit", *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2011, 1–10.

[159] J. Yan and Y. Song, "Weakly labeled sound event detection with residual CRNN using semi-supervised method", *tech. rep.*, DCASE 2019 Challenge, 2019.

[160] T. Yao, C. Shi, and H. Li, "Sound event detection in domestic environments using dense recurrent neural network", *tech. rep.*, DCASE 2020 Challenge, 2020.

[161] C.-Y. Yu, H. Liu, and Z.-M. Qi, "Sound event detection using deep random forest", *tech. rep.*, DCASE 2017 Challenge, 2017.

[162] X. Zheng, H. Chen, and Y. Song, "Zheng USTC team's submission for DCASE 2021 Task 4 – semi-supervised sound event detection", *tech. rep.*, DCASE 2021 Challenge, 2021.

[163] Q. Zhou and Z. Feng, "Robust sound event detection through noise estimation and source separation using NMF", *Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, 138–42.

[164] Q. Zhou, Z. Feng, and E. Benetos, "Adaptive noise reduction for sound event detection using subband-weighted NMF", *Sensors*, 19(7), 2019, 3206.

[165] x. Zhu and S. Xinghao, "Multi-scale network based on split attention for semi-supervised sound event detection", *Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, 2021, 155–9.

[166] M. Zöhrer and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification and acoustic event detection", *tech. rep.*, DCASE 2016 Challenge, 2016.