Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19557-2

Novel sound event and sound activity detection framework based on intrinsic mode functions and deep learning

Vahid Hajihashemi[1] · Abdorreza Alavigharahbagh[1] · J.J.M. Machado[2] · João Manuel R.S. Tavares[2]

Received: 10 October 2023 / Revised: 13 March 2024 / Accepted: 28 May 2024

Abstract
The detection of sound events has become increasingly important due to the development of signal processing methods, social media, and the need for automatic labeling methods in applications such as smart cities, navigation, and security systems. For example, in such applications, it is often important to detect sound events at different levels, such as the presence or absence of an event in the segment, or to specify the beginning and end of the sound event and its duration. This study proposes a method to reduce the feature dimensions of a Sound Event Detection (SED) system while maintaining the system's efficiency. The proposed method, using Empirical Mode Decomposition (EMD), Intrinsic Mode Functions (IMFs), and extraction of locally regulated features from different IMFs of the signal, shows a promising performance relative to the conventional features of SED systems. In addition, the feature dimensions of the proposed method are much smaller than those of conventional methods. To prove the effectiveness of the proposed features in SED tasks, two segment-based approaches for event detection and sound activity detection were implemented using the suggested features, and their effectiveness was confirmed. Simulation results on the URBAN SED dataset showed that the proposed approach reduces the number of input features by more than 99% compared with state-of-the-art methods while maintaining accuracy. According to the obtained results, the proposed method is quite promising.

Keywords Sound event detection · Feature dimension reduction · Empirical mode decomposition · Intrinsic mode function · Deep learning

✉ João Manuel R.S. Tavares
tavares@fe.up.pt

Vahid Hajihashemi
Hajihashemi.vahid@ieee.org

Abdorreza Alavigharahbagh
abalavi.gh@gmail.com

J.J.M. Machado
jjmm@fe.up.pt

[1] Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

[2] Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Published online: 11 June 2024

1 Introduction

The detection of various events and, consequently, the prevention of many dangers can be greatly enhanced using information obtained from sound signals, one of the most important aids for humans to understand their surroundings. The vision influences many human reactions; however, sounds are often used to announce alerts or make quick decisions, especially in everyday situations. The advantage of the sound signal is that it is not limited to direct vision, and this characteristic is one of the reasons why it is superior to an image in some situations. In many cases, by hearing sounds without seeing the event directly, one can prejudge details or nature of the event and take the necessary action [1].

The emergence of various advances in hardware, software, and Machine Learning (ML) algorithms, mainly of Deep Neural Networks (DNNs), has enabled researchers to perform complex functions. Additionally, the ability of DNNs to address various complex applications has led to changes in many conventional ML methods in many areas, including image, natural language, and sound processing and analysis. While DNNs typically require substantial computational resources for tasks such as image and video processing, the demand is comparatively lower for one-dimensional (1D) data. Working with 1D audio data or corresponding 1D features simplifies the computational requirements, unlike the intricate calculations required for processing images and videos. Consequently, this issue is less prominent in 1D SED systems [2].

A fundamental aspect of sound processing and analysis is detecting sound events, which has several applications, such as in security, medicine, and monitoring of urban events, and can be used simultaneously with information acquired by security and traffic cameras to increase detection accuracy and coverage. For example, in security systems, namely, in situations where an imaging camera cannot fully acquire the scene of an event for some reason, sound signals can be used in parallel to increase the accuracy and efficiency of the event detection system [3, 4]. In most sound event recognition systems based on DL, researchers have attempted to improve their efficiency and accuracy using standard sound features and modifying the structure of the used DL network. Few studies have focused on extracting useful sound features to optimize the performance of those SED systems based on Deep learning (DL). However, many different features can be extracted from sound signals, which usually require less computation than those extracted from images. The inherent instability of the features in the time and noise sensitivity of sound is also higher than those from images. Therefore, it is interesting to increase the efficiency, speed, and accuracy of a DL-based system by using a new feature extraction pattern that is particularly interesting in detecting sound events. In most SED systems, Mel coefficients and standard time-frequency domain features such as wavelets have been used to extract features. This study uses EMD and IMF for feature extraction, confirming their effectiveness in the proposed SED method. The proposed approach reduces the number of features required to detect a sound event while maintaining the system's performance. The main contributions and advantages of the proposed method can be categorized as follows:

1. Compared to conventional methods, the number of features required to detect a sound event is reduced;
2. The EMD method shows more robust characteristics against noise and distortion than other sound features;
3. The ability to detect multiple sound events simultaneously demonstrates the power of the EMD method in SED systems.

The article is organized as follows: In the next section, an overview of the related state-of-art is given, including the advantages and disadvantages of the current approaches; in the third section, the proposed method is described; in the fourth section, results of the proposed method are presented and compared against the ones of other methods; and finally, the findings of the current study are summarized and future works suggested.

2 Literature review

The approaches usually proposed in this area are based on multiclass classifiers because SED is considered a multiclass classification problem. In the field of feature extraction, the most commonly used features have been Mel-based features such as Log-Mel [5, 6], Log-Mel Power Spectrograms (LMS) [7, 8] and Mel Frequency Cepstral Coefficients (MFCC) [9–11]. In addition to MFCC, features such as linear predictive coding [12], discrete cosine transforms [13, 14], wavelet [9, 15], Perceptual Linear Prediction (PLP) [16], Linear Prediction Cepstral Coefficients (LPCC) [17], and Line Spectral Frequencies (LSF) [18] have been used in various studies for SED. MFCC has been used as a usual feature in a wide range of acoustic and sound-based machine-learning methods, for example, in voice disorder detection [19], emotion recognition [20–22], singing voice separation [23], fault detection using acoustic and sound data [24, 25], leak detection [26] and tree cutting events detection [27].

Several researches and experiments show that IMFs extracted using EMD from sound signals show a good response. For example, Pandya et al. [28] used sound signals in conjunction with IMF features and the K-nearest neighbor classifier to detect problems in ball bearings.

Amarnath et al. [56] used IMF to detect faults in a helical gearbox using sound and vibration signals and obtained good results. Zahra et al. [57] used Multivariate Empirical Mode Decomposition to detect seizures from medical electroencephalogram signals with an Artificial Neural Network (ANN) classifier. Bagherzadeh [58] used IMF to predict the sound signal envelope. Cheema and Singh [59, 60] used EMD to capture nonlinear dynamics of phonocardiogram signals to detect stress. Yao et al. [61] applied EMD to extract features from sounds and detect faults in a planetary gearbox by using the Random Forest (RF) classifier. Ning et al. [62] relied on EMD to extract sound features and detect gas pipe leakage from sound data with an RF classifier. Erdogan and Narin [63] used the cough signal, EMD, and a deep neural network to diagnose COVID-19 disease. Vican et al. [64] detected the pulse in the fetal phonocardiography signal by EMD. Therefore, it can be said that the EMD method has been successfully used for sound feature extraction in medicine and industry and has proven its efficiency. In most new SED studies, DL algorithms have shown their superiority in terms of classification accuracy compared to conventional methods. Hence, conventional DL algorithms and Convolutional Neural Networks (CNNs) have been frequently used individually and combined. For example, CNNs were used in [33–35, 42–44, 53]. ResNet was used as a CNN with some modifications in [32, 42, 55]. Recurrent Neural Network (RNN) in combination with a CNN, called CRNN, was considered in several works, such as the ones presented in [29, 37–40, 42, 45–51, 54], for SED.

Meng et al. [5] used a bidirectional gated recurrent unit (BGRU) as an RNN for sound event detection. Politis et al. [65] analyzed the classifiers used in Sound Event Localization and Detection in the DCASE 2019 Challenge and concluded that most were CRNN. Some studies, such as [66], have also used Generative Adversarial Networks (GANs). A limited number of researchers have used hidden Markov models [67], regular neural networks [68], support vector machines [69], cross-correlation [70], and ensemble learning [71], which is far

less than the number of researchers that used DL algorithms. Table 1 identifies state-of-the-art DL-based approaches used in SED systems.

As a summary, the advantages of previous methods are:

1. Widespread Use: MFCCs and spectrogram-based features have been widely used in SED systems for their simplicity and effectiveness;
2. Interpretability: Those features are easily interpretable by humans, aiding in understanding the characteristics of sound events;
3. Established Performance: Due to their extensive usage, there are well-established performance metrics and benchmarks, making comparing different approaches easier.

The disadvantages of previous methods can be summarized as:

Table 1 Deep-learning approaches that have been used in SED systems

Reference	Feature	Classifier	Metric
Mushtaq, Z. and Su, S.-F.: [8]	LMS	CNN	Accuracy
Su, Y et al. [6]	Log-Mel	CNN	Accuracy
Gontier, F et al. [29]	Mel Spectrogram	CRNN	Accuracy
Wang, J et al. [30]	Mel Spectrogram	BGRU	Accuracy
Jose, T et al. [31]	MFCC	LSTM	Accuracy
Esmaeilpour, M et al. [32]	MFCC	ResNet	Accuracy
Kong, Q et al. [33]	Raw signal	CNN	Accuracy
Katsis, L.K et al. [7]	LMS	CNN	F-score
Meng, J et al. [5]	Log-Mel	BGRU	F-score
Lin, L al. [34]	Log-Mel	CNN	F-score
Serizel, R et al. [35]	Log-Mel	CNN	F-score
Gao, L et al. [36]	Log-Mel	CNN	F-score
Nam, H et al. [37]	Log-Mel	CRNN	F-score
Dinkel, H et al. [38]	Log-Mel	CRNN	F-score
Nguyen, T.N.T et al. [39]	Log-Mel	CRNN	F-score
Komatsu, T et al. [40]	Log-Mel	CRNN	F-score
Tonami, N et al. [41]	Log-Mel	DNN	F-score
Johnson, D.S et al. [42]	Mel Spectrogram	CNN	F-score
Chan, T.K et al. [43]	Mel Spectrogram	CNN	F-score
Huang, Y et al. [44]	Mel Spectrogram	CNN	F-score
Turpault, N et al. [45]	Mel Spectrogram	CRNN	F-score
Pankajakshan, A et al. [46]	Mel Spectrogram	CRNN	F-score
Bear, H.L et al. [47]	Mel Spectrogram	CRNN	F-score
De Benito-Gorrón, D et al. [48]	Mel Spectrogram	CRNN	F-score
Pankajakshan, A et al. [49]	Mel Spectrogram	CRNN	F-score
Martín-Morató, I et al. [50]	Mel Spectrogram	CRNN	F-score
Park, H et al. [51]	Mel Spectrogram	CRNN	F-score
Al-Banna, A.-K et al. [52]	MFCC	LSTM	F-score
Turpault, N et al. [53]	Raw signal	CNN	F-score
Turpault, N et al. [54]	Raw signal	CRNN	F-score
Hershey, S et al. [55]	Log-Mel	ResNet	ROC curve

1. High Dimensionality: Conventional features such as MFCCs can result in high-dimensional feature vectors, leading to increased computational complexity and memory requirements;
2. Limited Discriminative Power: While effective for many applications, conventional features may lack the discriminative power needed to distinguish between subtle variations in sound events;
3. Fixed Representations: Features like MFCCs provide fixed representations of sound, which may not capture the dynamic nature of certain events or adapt well to changing environments.

By addressing these limitations, the proposed method aims to overcome the challenges associated with conventional feature extraction techniques. Using IMF and extracting locally based features, the proposed scheme reduces the feature dimensions while maintaining or improving the efficiency and accuracy of SED systems. In the current study, an LSTM was used as the classifier.

3 Proposed method

This study developed two event detection approaches: a segment-based approach and an activity-based approach. In the segment-based event detection approach, a sound clip is cut into multiple fixed-size segments, and the system processes each segment individually. Since multiple events may occur simultaneously, a practical solution in a segment-based event detection system is to train a binary classifier for each event separately. This classifier indicates whether or not an event occurred in a segment. Activity-based event detection specifically detects the start and end of an event in a sound clip and can estimate the duration of the event.

3.1 Segment-based event detection

The proposed method for segment-based event detection is depicted in Figs. 1 and 2. According to Figs. 1 and 2, the proposed segment-based event detection method includes two main parts: feature extraction and classification. IMFs are used in feature extraction, and Long Short-Term Memory (LSTM) or ensemble learning is used for classification.

3.2 Feature extraction

The input sound is divided into several time intervals in the feature extraction phase, depending on the selected approach. The intervals can be chosen with or without overlap, and should

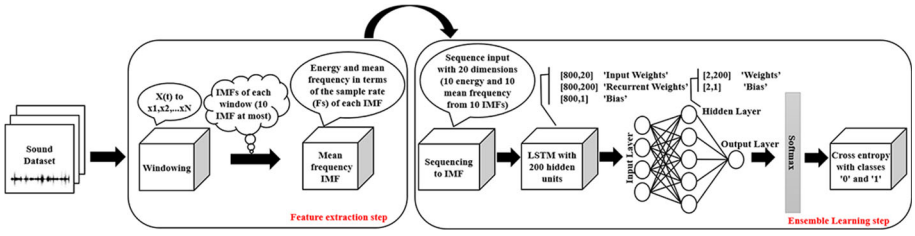

Fig. 1 Block diagram of the proposed method for segment-based event detection based on DL

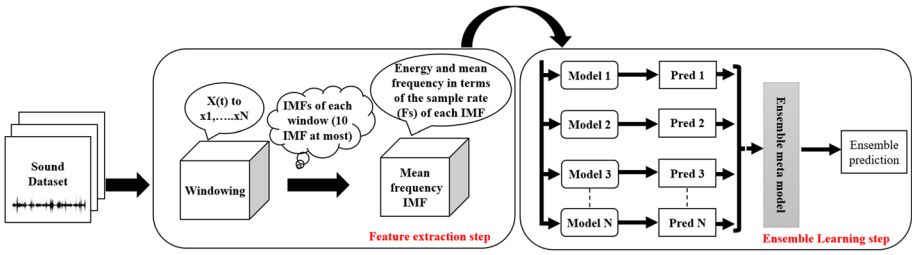

Fig. 2 Schematic of the proposed method for segment-based event detection based on ensemble learning

not be so long that the online capability of the method is compromised or so short that the feature extraction process cannot provide the desired result. Moreover, excessively shortening the time intervals amplifies the impact of noise. It is important to note that all feature extraction methods inherently require a minimum number of samples to extract features, which depends on the sampling frequency and feature type.

Algorithm 1 EMD for IMF Extraction.

Input:
Input signal: x(t)
Constant eps: a small positive value close to 0 (zero)
Initialize $i = 1$, $c_1 = \text{eps}$, $xb = x$
while Until stopping condition is met **do**
  Extract upper envelope env_max and lower envelope env_min using extremum points of x and cubic spline interpolation
  Calculate $temp = \frac{env\_max + env\_min}{2}$
  Compute $c_2 = x - temp$
  if c_2 satisfies IMF conditions **then**
    Set $\text{IMF}_i = c_2$
    Update $x = xb - c_2$
    if x is a monotonic function **then**
      Set $r = x$ and exit loop
    else
      Update $xb = x$, $c_1 = \text{eps}$, $N = i$, and $i = i + 1$
    end if
  else
    if $\frac{(c_1 - c_2)^2}{c_1} < \gamma$ **then**
      Set $r = x$ and exit loop
    else
      Update $x = c_2$, $c_1 = c_2$, and $xb = x$
    end if
  end if
end while
Output:
Output: IMFs

3.3 IMF

Sound is considered a quasi-linear or non-stationary signal; hence, time series-based methods are required to model its nonlinear and nonstationary behavior. Due to the wide application of time series in various fields such as economics, medicine, and industry, many methods have

been proposed to analyze these signals quickly, such as the ones based on the spectrogram, wavelet analysis, Wigner-Ville distribution, evolutionary spectrum, and principal component analysis. Mel coefficients, which were specifically developed based on the human auditory system and are very efficient in feature extraction from sound signals, have also been used. All current methods attempt to identify and extract the inherent characteristics of the nonlinear and nonstationary sound signal that change less over time and depend on the desired output. However, most of these methods have problems with unstable and nonlinear signals, mainly:

1. When a signal is nonstationary, it is generally computed the harmonic components, which require a large amount of data to extract the characteristics of the signal over time.
2. Most of those methods require a linear system to obtain signal information, and in nonlinear systems, a lot of data is needed to model the nonlinear components. In addition, the EMD method produces a collection of IMFs that allows the system to extract instantaneous frequencies from the signal at different time scales.

IMFs are well-functioning Hilbert transforms that can extract the instantaneous frequencies of a system in short periods and model the phenomenon under study on the time-frequency axis, even if they are transient. The main concept used in IMF is the instantaneous frequency, which differs from the time-independent frequency defined in most transforms, such as the Fourier transform. In the concept of instantaneous frequency, the frequency can vary with time, similar to frequency modulation. One must first understand the Hilbert transform to understand the concept of IMF. The Hilbert transform of a signal, $X(t)$, is defined as:

$$Y(t) = \frac{1}{\pi} p.v. \int_{-\infty}^{\infty} \frac{X(t')}{t - t'} dt', \tag{1}$$

where $p.v.$ is the Cauchy principal value. The Hilbert transform of $X(t)$ is combined with the signal itself as a complex function, $Z(t)$:

$$Z(t) = X(t) + iY(t) = \alpha(t)e^{i\theta(t)}, \tag{2}$$

where α and θ are the absolute value and argument of the polar form of the complex function, $Z(t)$. Based on θ, the instantaneous frequency of the signal $X(t)$ can be defined as:

$$\omega = \frac{d\theta}{dt}. \tag{3}$$

Even with the above definition, there is still ambiguity in the definition of instantaneous frequency because calculating the Hilbert transform requires an infinite number of samples. The Hilbert transform is limited within EMD; therefore, an alternative function class is defined as IMF, which can define the instantaneous frequency locally. In the entire period, the number of extrema and zero crossings of IMF should equal the original signal or the maximum difference of this number should be 1 (one). After applying EMD and extracting IMFs, one obtains:

$$\mathrm{X(t)} = \sum_{i=1}^{N} IMF_i(t) + r(t), \tag{4}$$

where N is the number of IMFs and $r(t)$ is the residual signal representing the computational error. In most cases, the residual signal is monotonic and has low amplitude. The pseudocode for EMD is given in Algorithm 1.

In Algorithm 1, γ is assumed to be 0.2, as suggested in the used Matlab software. The maximum number of IMFs in this step is assumed to be 10, and all remaining IMFs in the signal are discarded. After extracting IMFs, each IMF's energy and average frequency are

extracted as the final feature. The energy is calculated using the sum of the square powers of the amplitudes in each IMF. The average frequency is defined by [72] as:

$$Average\ Frequency = \frac{\sum_{j=1}^{M} f_j P_j}{\sum_{j=1}^{M} P_j}, \tag{5}$$

where P_j is the power of the signal at frequency f_j. If a signal has less than 10 IMFs, the energy values and the average frequency of IMFs that do not exist are assumed to be 0 (zero). If a signal has more than 10 IMFs, the first 10 IMFs are included in feature extraction, and the rest is discarded.

3.4 Deep Learning

RNNs are particularly suited for processing serial data, where subsequent samples depend on previous ones. In traditional RNNs, due to their simple structure and limited recurrent coefficients in the hidden layers, weights are updated using the gradient relationship, which fails in the case of long-time series. This limitation in maintaining and understanding long-term patterns is a weakness of RNNs.

Several techniques have been proposed to address this weakness of RNNs, including non-gradient-based training patterns such as simulated annealing and discrete error propagation [73, 74], explicitly introduced time delays [75–77] or time constants [78], and hierarchical sequence compression [79] are among them with each having its limitations and advantages.

3.4.1 LSTM

LSTM belongs to the modified RNN architectures [80]. This DL model is considered the most effective for] simultaneously capturing long-term and short-term patterns. Considering the importance of processing time series and video data whose results depend on current and past data, LSTM is considered one of the most widely used Neural Network (NN) models. Its structure is a modified RNN with the purpose of long-term data retention. In a typical RNN, as the data are updated, the influence of the data on the more distant samples decreases compared to the closer samples until it eventually becomes almost 0 (zero). Figure 3 depicts

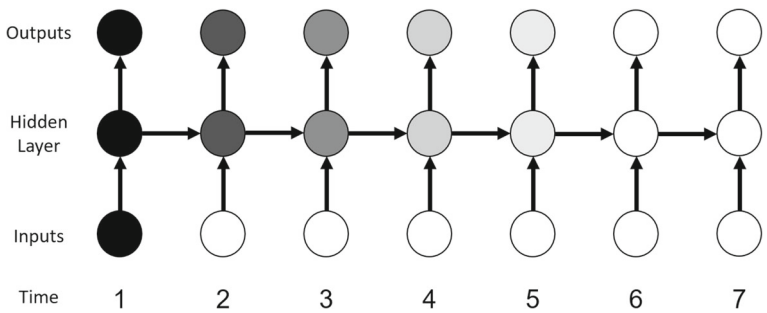

Fig. 3 Reduction of the effect of distant samples in updating the hidden layer of an RNN

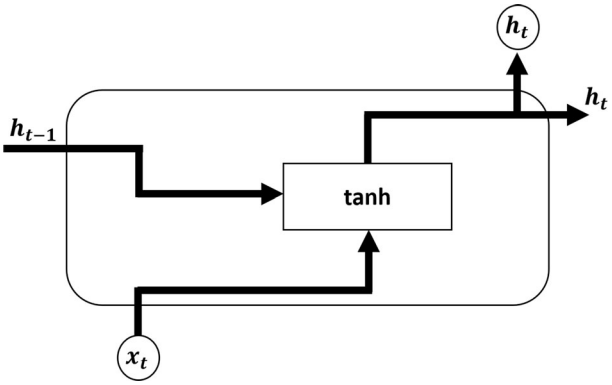

Fig. 4 A typical RNN structure

the reduction of the effect of distant samples in updating the hidden layer of an RNN, which indicates the small impact of distant samples and their ineffectiveness over time.

Figure 4 shows a typical RNN where the output depends on the previous states and the new input. Figure 5 shows the LSTM structure, which is also essential to state that its cell is more complex than a simple RNN. In Fig. 5, σ denotes the activation function. The cell's input and output activation functions (σ_g and σ_h) are usually hyperbolic tangent functions (tanh) or logistic sigmoid functions, although in some cases, σ_h is the identity function. Dashed lines represent weighted 'peephole' connections. In peephole connections, in addition to the

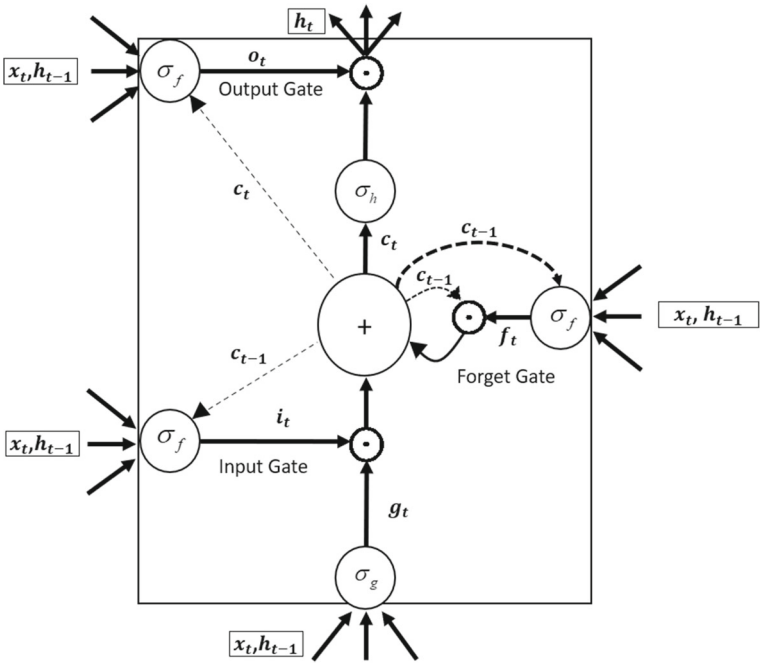

Fig. 5 The usual LSTM network

input and previous internal states, hidden states are also used to control input and output, and forgetting gate activation functions assist in increasing the degrees of freedom and capabilities of LSTM. Forget Gate determines which inputs and previous states affect the output and which should be ignored. The presence of the forget gate allows the cell to learn long-term patterns.

Equations (6) to (11) show the effect of the blocks depicted in Fig. 5 on the LSTM output, where x_t is the input, h_{t-1} the previous state, c_t the previous states of Forget Gate, W the weights, b the bias of each part, σ the activator functions, $\odot$ the Hedmard's multiplication, and the output is h_t [81]:

$$i_t = \sigma_f \left(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} \odot c_{t-1} + b_i\right), \tag{6}$$

$$f_t = \sigma_f \left(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} \odot c_{t-1} + b_f\right), \tag{7}$$

$$g_t = \sigma_g \left(W_{xc} x_t + W_{hc} h_{t-1} + b_c\right), \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{9}$$

$$o_t = \sigma_f \left(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} \odot c_t + b_o\right), \tag{10}$$

$$h_t = o_t \odot \sigma_h \left(c_t\right). \tag{11}$$

Based on our best knowledge, most SED methods used simplified LSTMs [31, 52]. Peephole LSTM has been predominantly used in other domains [82–84]. Hence, the proposed method uses the simplified LSTM shown in Fig. 6. In the used LSTM network, the previous equations were changed as:

$$i_t = \sigma_f \left(W_{xi} x_t + W_{hi} h_{t-1} + b_i\right), \tag{12}$$

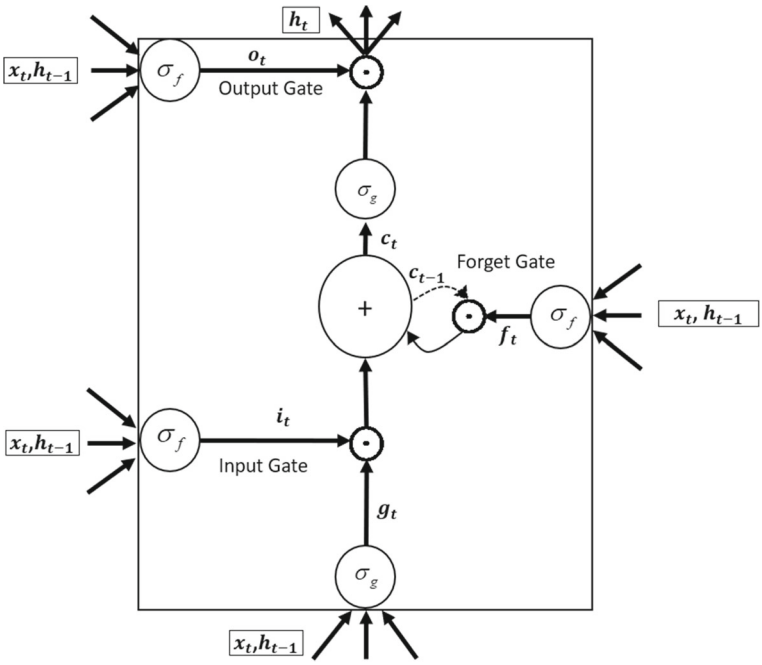

Fig. 6 Simplified LSTM network used in the current study

$$f_t = \sigma_f \left(W_{xf} x_t + W_{hf} h_{t-1} + b_f \right), \tag{13}$$

$$g_t = \sigma_g \left(W_{xc} x_t + W_{hc} h_{t-1} + b_c \right), \tag{14}$$

$$c_t = f_t \odot c_{t-1} + g_t \odot i_t, \tag{15}$$

$$o_t = \sigma_f \left(W_{xo} x_t + W_{ho} h_{t-1} + b_o \right), \tag{16}$$

$$h_t = o_t \odot \sigma_g \left(c_t \right). \tag{17}$$

3.4.2 Fully connected layers

The fully connected layer is equivalent to the hidden layer in typical NNs. This layer combines an affine function and a nonlinear activation function. The affine function is defined as $y = Wx + b$. The nonlinear activation function can be defined from a class such as sigmoid, tanh, or rectified linear unit (ReLU). The fully connected layer in the proposed structure has a no nonlinear function and only has 1 (one) affine function. The connections between the LSTM and middle layers and the middle and output layers are usually made through this layer.

3.4.3 Softmax

The softmax layer or softmax function, also known as softargmax or normalized exponential function, maps the input vector to a set of numbers between 0 (zero) and (1) one. This function provides a smooth and continuous approximation to the differentiable maximum function. The sum of the output numbers of this function, which is a probability distribution, is necessarily equal to 1 (one). This layer is not trained and maps the input to the interval: [0 1]. The formula for the softmax function is:

$$y_i = \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}}, \tag{18}$$

where x_i is the input vector, i.e., the output of the fully connected layer, and y_i is the output corresponding to each input.

3.4.4 Cross entropy loss

Cross entropy is used to maximize the accuracy of the entire classifier. The value of cross entropy increases rapidly when the predicted probability deviates from the actual value. Thus, minimizing cross entropy is equivalent to bringing the predicted probability closer to the real value. A classifier trained using cross entropy is more accurate and effective than a classifier trained using other optimization criteria where the last layer can produce probability values. In information theory and pattern recognition, minimizing cross entropy is equivalent to achieving maximum likelihood. Minimizing the cross entropy is equivalent to minimizing the Kullback-Leibler divergence between the probability distribution of the real output and the probability distribution of the classifier, corresponding to the maximum similarity between the ideal output and the classifier's output.

3.4.5 Ensemble learning

In addition to DL, ensemble learning is used in the proposed method. Ensemble learning methods use a set of weak classifiers instead of a single classifier to improve efficiency. The parameters of an ensemble classifier are the number of weak classifiers, the type of classifiers, and whether they are similar or different. Bootstrap aggregation,i.e., bagging, performed better than other types of ensemble learning in our study. Finally, the results of all classifiers are combined, and the dominant class, i.e., the class selected by most classifiers, is chosen as the final output. The proposed method used decision trees as weak classifiers in the ensemble learning structure.

3.5 Event Activity Detection System

A sound activity detection system determines the start and end of an event or the duration of the event in a sound clip. Due to the complexity of detecting the beginning and end of a sound event and its structural differences from the segment classification, this study takes into account changes in the extracted features instead of the approach normally used in conventional methods. The proposed sound activity detection method is divided into two steps:

1. Detect if there is a sound event in the input clip;
2. And, if there is, find the start and end of the event based on the change in the used features.

Since this is a hierarchical method, the efficiency of both steps, which involve detecting the sound occurrence and correctly labeling the start and end, directly impacts the method's accuracy. The block diagram of the first step is depicted in Fig. 7.

The first step of the activity detection system is similar to that of the segment-based event detection system. In this case, the IMF features of the entire signal are obtained, so one has only 20 features for each clip x(t). According to the studies conducted in this case, ANN is a better choice than DL and ensemble learning methods because of the small number of input features. The second step is initiated if the first part's output is an event's occurrence. Figure 8 depicts the block diagram of the second step's training phase.

Only audio clips containing selected events were used in the training phase. First, the input signal is split into 1-second segments without overlap. The end of the segment before the event is selected as the beginning of the event. The start of the segment after the ending event is considered the end of the event. As in the segment-based approach, IMF features are extracted for each segment. In this way, one has 20 features for each interval. Any feature that changes when a sound event starts or ends can be used to detect event activity. This process is very challenging because background noise and other events occur in different

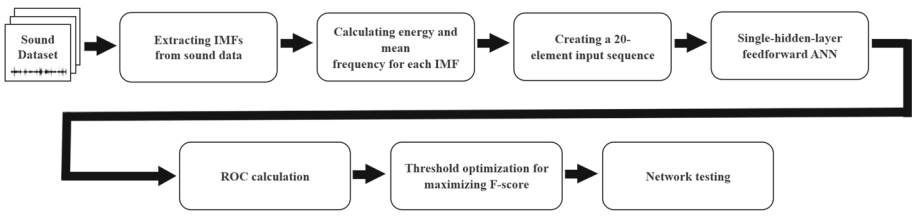

Fig. 7 Block diagram of the first step of the sound activity detection method

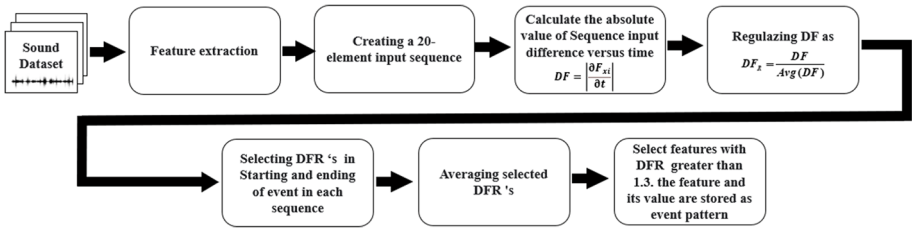

Fig. 8 Block diagram of the training phase of the second step of the sound activity detection method

segments and cause changes in the value of features while having no meaningful relationship to the beginning and end of the event. To solve this problem, an averaging and regularization step is added to the proposed method to separate the effects of noise and other sound events from selected events. In the regularization block, the change in feature is measured relative to the overall signal by dividing the derivative value by the average value of the derivative in the entire signal. The regularization step removes the background noise effectively. In the averaging block, the pattern of the derivative vector is determined by averaging the absolute value of the regularized derivative of the features at the beginning and end of the selected event.

In some cases, another sound event coincides with the selected event, and averaging removes the effects of these interfering sound events. When a feature does not change with a selected event, its regularized value average is approximately 1 (one). The threshold value used to select or discard a feature is 1.3. The changed features, features whose averaged regularized value is greater than 1.3, and their rate of change, i.e., the averaged regularized value, are stored as the pattern of the selected event. The activity detection phase is depicted in Fig. 9.

In the final phase, similar to the training phase, the signal is split into non-overlapping one-second segments. For each segment, 10 IMFs are calculated, and the average frequency and energy characteristics of IMFs are extracted as features. In the time domain, a derivative of the obtained features is computed, followed by regularization, as in the training phase. Some regularised features are selected based on the pattern stored in the training phase, and the correlation coefficient between the stored pattern and regularized derivatives of all segments is calculated. The maximum correlation coefficients are selected as the beginning and end of the event. If only one maximum is detected, it is considered the event's starting point, and the event is assumed to last until the end of the clip.

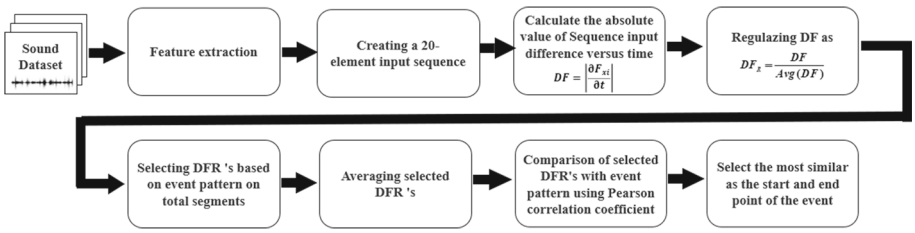

Fig. 9 Block diagram of the final phase of the sound activity detection system

4 Experimental settings

4.1 Dataset

The proposed method was tested on the URBAN-SED dataset[1], which is widely used in this field [85]. The URBAN-SED dataset contains 10000 labeled samples of ten events in the urban area, which are classified as air conditioners, car horns, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren, and street music (Fig. 10). Regarding time, all samples in the dataset have the same length of 10 seconds. The dataset contains a total of 100,000 seconds (approximately 28 hours) of sound, with almost 50,000 events tagged. All sounds contain background and Brownian noise, which can be heard as the typical "hum" of most crowded urban environments.

The dataset was created using the Scaper library, a soundscape synthesis and enhancement library. The sounds of the included events were taken from the UrbanSound8K dataset, which is completely real, and the scaper library added the sounds of urban environments. The labeling was done automatically according to the time of the added event. The UrbanSound8K dataset contains 8732 urban environment events with times shorter than 4 seconds. The UrbanSound8K dataset is a modified version of the UrbanSound dataset, which contains 1302 samples totaling approximately 27 hours. To standardize the comparison between different methods, UrbanSound8K was divided into three subsets: training, testing, and validation, with 6000 samples in the training group and 2000 samples in the testing and validation groups.

4.2 Evaluation metrics

In both segment-based sound event detection and sound activity detection approaches, the developed system was trained separately for each event. In the segment-based approach, a binary classifier determines whether an event has occurred in the segment. In the activity detection approach, each true label for the start and end of an event in a clip is assumed to be a true positive label, making it possible to evaluate the system using binary classification metrics. The following evaluation parameters were used for the segment-based approach before evaluation:

1. **True positive (TP) :** The sound event occurred and was correctly detected;
2. **True negative (TN) :** The sound event did not happen, and the non-event is correctly detected;
3. **False positive (FP) :** The system detected an event that did not occur;
4. **False negative (FN) :** The absence of an event has been detected when the sound event occurred.

In the sound activity detection approach, the evaluation parameters were defined as follows:

1. **True positive (TP) :** The beginning and end of the sound event were correctly detected;
2. **True negative (TN) :** The sound event did not happen, and the non-event was correctly detected;
3. **False positive (FP) :** The system detected an event, but it did not happen, or if the event did occur, the beginning and end of the event were incorrectly marked;

[1] http://urbansed.weebly.com/

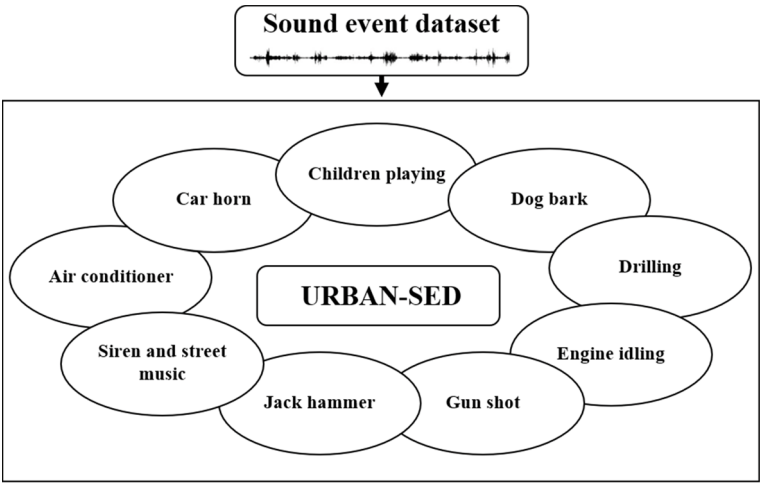

Fig. 10 Events included in the URBAN-SED dataset

4. **False negative (FN) :** Detecting the absence of an event when the sound event occurred.

It is essential to note the imbalance between the two classes: the presence or absence of sound events. After specifying TP, TN, FP, and FN, the Precision (PR), Recall (RE), F-score (F1 or F-score), and Accuracy (ACC) can be calculated. The ideal value for all of these parameters is 1 (one). In the present study, Segment F1 and Event F-scores [38, 86, 87] were used as primary assessment metrics, and precision, recall, and accuracy were used as secondary assessment metrics [38, 86].

4.3 Other parameters

Another important parameter of the proposed method is the segment length. In [42], the segment length was assumed to be 0.1 seconds, and in [55], 1 (one) second. A longer segment length leads to more background noise, and a shorter length decreases valuable information for event detection. Therefore, the choice of segment length is a trade-off between noise and valuable information. In this study, the segment length was assumed to be 1 (one) second. For IMF extraction, the parameters were selected as follows:

1. The Cauchy type convergence criterion (γ in the EMD pseudocode), which is one of the stopping criteria, was set to 0.2;
2. The maximum number of iterations, which is one of the stopping criteria, was set to 100;
3. The maximum number of IMFs, which is one of the decomposition stop criteria, was chosen to be 10;
4. The maximum number of extrema in the residual signal, which is one of the stopping criteria for decomposition, was set to 1 (one):
5. The ratio between signal and residual energy, which refers to the ratio between the energy of the signal at the beginning of the iteration and the average envelope energy, is one of the decomposition's stop criteria and was set to 20;
6. The envelope construction is based on the spline-based interpolation method.

For the RNN, the first parameter is the number of hidden layers of LSTM, which was assumed to be 200. Two Adaptive Moment Estimation (ADAM) and Root Mean Squared Propagation (RMSprop) methods were used to train the RNN, and the results were compared. The parameters of the ADAM method are listed in Table 2. During training in very large datasets, the ADAM method is better than the Gradient Descent method. Gradient Descent involves problems, such as many calculations and failure to reach the global minimum when there are many local minima. By simplifying the calculation of the learning rate for each parameter using the first and second moments of the gradient, the ADAM method reduces the computational volume and memory consumption of conventional stochastic gradient descent. On the other hand, RMSprop is a modified version of gradient descent where the step size for each parameter is adjusted using a decaying average of partial gradients. A decaying moving average allows the algorithm to eliminate early gradients and works based on the most recently observed partial gradients during the search process.

5 Results and discussion

This section aims to analyze the impact of the different parameters on the proposed method's efficiency and compare the final results with the ones obtained by related state-of-the-art methods. In the first step, the results of the segment-based event detection are reported, and the effects of data balancing and some other parameters on accuracy are discussed. In the second step, event-based results are reported and compared with the ones obtained by other methods.

5.1 Segment-based event detection

In the first step, the proposed segment-based event detection method was tested for a segment length of 1 (one) second without overlap using LSTM and ensemble learning. To show the efficiency and effectiveness of the proposed method, the results were compared with the ones using mel features such as MFCC and log-mel, which are the most commonly used features in SED. The URBAN-SED is an unbalanced dataset for all events, and this unbalanced form

Table 2 Parameters of the ADAM method

Parameter	Value
Gradient Decay Factor	0.9000
Squared Gradient Decay Factor	0.9990
Epsilon	1e-08
Initial Learn Rate	1e-03
Learn Rate Drop Factor	0.1000
Learn Rate Drop Period	10
L2 Regularization	1e-04
Gradient Threshold Method	'l2norm'
Gradient Threshold	2
Max Epochs	100
Mini-Batch Size	128
Shuffle	'once'

Table 3 Number of training samples before and after data balancing

Event	Original		Balanced data	
	Negative	Positive	Negative	Positive
Air Conditioner	53746	6254	16156	6254
Car Horn	55453	4547	15203	4547
Children Playing	53875	6125	15995	6125
Dog Bark	54597	5403	16207	5403
Drilling	54149	5851	16599	5851
Engine Idling	53694	6306	16374	6306
Gun Shot	56219	3781	17439	3781
Jackhammer	54100	5900	16570	5900
Siren	53818	6182	16138	6182
Street Music	53684	6316	16174	6316

can significantly reduce the classification efficiency. To solve this problem, similar negative segments were discarded by using the correlation coefficient as a similarity measure. Table 3 presents the number of training data in each class before and after the balancing process, assuming a segment length of 1 (one) second. Table 4 presents the acceptable F-score results of the proposed feature, which indicate a strong correlation between features extracted from IMFs and events. In the worst case that corresponds to a gunshot event, even after the balancing process, there is a significant imbalance in the ratio between the positive and negative samples (Table 3), which may cause the system's low accuracy.

5.2 Sound activity detection

In this experiment, the training and test data were approximately balanced; therefore, a balancing process was unnecessary. Table 5 presents the number of clips of the two classes

Table 4 F-score of proposed features using different classifiers on the URBAN-SED dataset in the segment-based approach

Event	LSTM Adam			LSTM RMSprop			Ensemble learning		
	Train	Test	Validation	Train	Test	Validation	Train	Test	Validation
Air Conditioner	0.58	**0.51**	**0.52**	0.55	0.42	0.45	1	0.48	0.52
Car Horn	0.42	0.32	0.32	0.51	**0.37**	**0.35**	1	**0.37**	**0.35**
Children Playing	0.40	0.32	0.38	0.54	**0.47**	**0.50**	1	0.37	0.38
Dog Bark	0.45	**0.43**	**0.45**	0.48	0.41	0.46	1	0.38	0.42
Drilling	0.55	0.51	0.47	0.56	0.51	0.51	1	**0.54**	**0.51**
Engine Idling	0.54	**0.55**	**0.52**	0.54	0.51	0.49	1	0.52	0.51
Gun Shot	0.04	0.02	0.04	0.26	0.12	0.25	1	**0.14**	**0.30**
Jackhammer	0.53	**0.59**	**0.51**	0.51	0.47	0.49	1	**0.59**	**0.51**
Siren	0.60	**0.48**	**0.60**	0.63	**0.50**	**0.63**	1	**0.52**	**0.60**
Street Music	0.47	0.43	0.45	0.48	0.45	0.45	1	**0.46**	**0.48**
Average	0.458	0.416	0.426	0.506	0.423	0.458	1	0.437	0.458

Best values in bold

Table 5 Number of clips with (P) and without (N) a sound event on the URBAN-SED dataset

Event	Train		Test		Validation	
	N	P	N	P	N	P
Air Conditioner	3697	2303	1242	758	1182	818
Car Horn	3708	2292	1201	799	1257	743
Children Playing	3726	2274	1234	766	1255	745
Dog Bark	3677	2323	1212	788	1204	796
Drilling	3661	2339	1214	786	1220	780
Engine Idling	3685	2315	1227	773	1234	766
Gun Shot	3666	2334	1213	787	1241	759
Jackhammer	3681	2319	1260	740	1271	729
Siren	3709	2291	1238	762	1247	753
Street Music	3684	2316	1257	743	1208	792

for each sound event separately. For this step, different ANN structures were tested, including Patternnet, Cascadeforwardnet, Feedforwardnet, and Fitnet, with Feedforwardnet being the best ANN structure found. Various functions and values were considered when choosing the training function and the number of hidden layers and nodes in each layer. The studied training functions are listed in Table 6. Among the studied functions, trainbr and trainlm showed better results (F-score), which was chosen because of the shorter training time. According to the analysis performed, the structure with 1 (one) hidden layer and 10 nodes showed the best F-score. Table 7 presents the F-score for some of the studied situations.

The results in Table 7 indicate that the feed-forward network with a hidden layer and 10 neurons is the best structure for this step. After training the forward ANN with the above parameters, a Receiver Operating Characteristic Curve (ROC curve) was used to depict the F-score. Figure 11 shows the ROC curve for several events and the chosen threshold level. Maximizing the F-score was prioritized when choosing the threshold level based on the ROC. The main finding, depicted in Fig. 11, where one can observe the magnified area in the middle of the curve, is the divergent behavior of the events. Consequently, during the training phase,

Table 6 ANN training functions that were considered in the first part of the activity detection method

Training Function	Description
trainlm	Levenberg-Marquardt
trainbr	Bayesian Regularization
trainbfg	BFGS Quasi-Newton
trainrp	Resilient Backpropagation
trainscg	Scaled Conjugate Gradient
traincgb	Conjugate Gradient with Powell/Beale Restarts
traincgf	Fletcher-Powell Conjugate Gradient
traincgp	Polak-Ribiere Conjugate Gradient
trainoss	One Step Secant
traingdx	Variable Learning Rate Gradient Descent
traingdm	Gradient Descent with Momentum
traingd	Gradient Descent

Table 7 F-score values obtained for some of the investigated situations

Model	Training function	Hidden layers	Nodes per layer	Average F-score
Feedforwardnet	trainlm	1	2	0.52
Feedforwardnet	trainlm	1	3	0.531
Feedforwardnet	trainlm	1	4	0.534
Feedforwardnet	trainlm	1	5	0.541
Feedforwardnet	trainlm	1	6	0.54
Feedforwardnet	trainlm	1	7	0.545
Feedforwardnet	trainlm	1	8	0.542
Feedforwardnet	trainlm	1	9	0.547
Feedforwardnet	trainlm	1	10	0.556
Feedforwardnet	trainlm	1	11	0.551
Feedforwardnet	trainlm	1	12	0.548
Feedforwardnet	trainlm	1	13	0.531
Feedforwardnet	trainlm	1	14	0.541
Feedforwardnet	trainlm	1	15	0.534
Feedforwardnet	trainlm	2	[10 10]	0.548
Feedforwardnet	trainlm	2	[5 10]	0.54
Feedforwardnet	trainlm	3	[5 10 5]	0.551
Feedforwardnet	trainlm	3	[4 12 4]	0.549
Feedforwardnet	trainbr	1	10	0.555
Feedforwardnet	trainbr	3	[4 12 4]	0.547
Patternet	trainlm	1	10	0.537
Patternet	trainlm	3	[5 10 5]	0.531

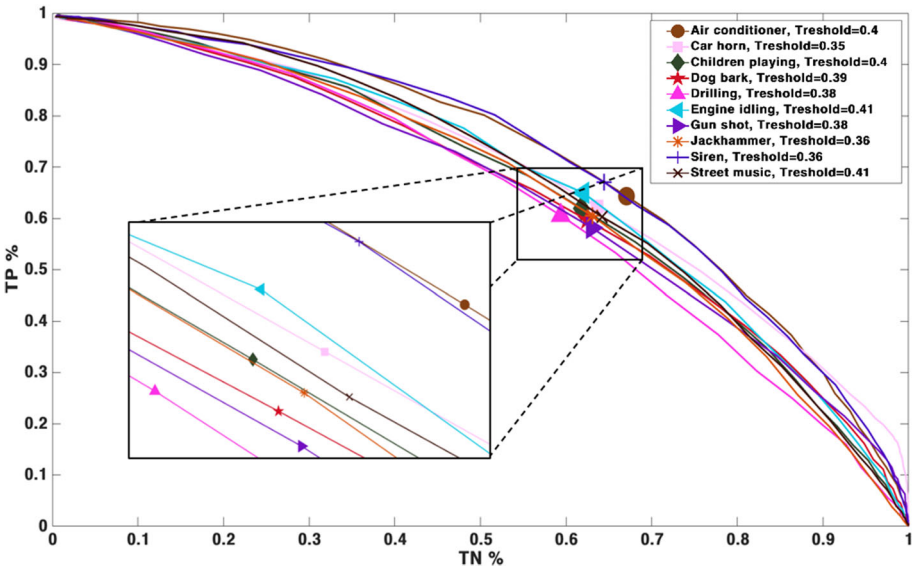

Fig. 11 ROC curves of trained ANNs for different events

it is imperative to compute the threshold for each event independently. The chosen thresholds are indicated in the legend of Fig. 11, positioned at the top right corner.

Table 8 presents the results of the proposed method after selecting the threshold based on the ROC for each event. When a sound event is detected in the first step, the proposed method submits it into the second step to determine the start and end points of the event. Therefore, any error in the first step results in an error in the output, and reporting accuracy metrics for this step seemed redundant as it does not show the overall system's performance. The average absolute value of the regularized derivatives for the start and end segments of the selected event was calculated in the training phase of the second part. Table 9 lists the calculated values for the different events. A threshold value of 1.3 was considered for selecting effective features, and the values above the threshold are indicated in bold in Table 9.

Based on the results presented in Table 9, the following deductions can be stated:

1. Average frequency features are more critical than energy features;
2. 9th and 10th IMF are effective only in the gunshot and in the other events are useless;
3. In detecting a sound event's start and end time, i.e., in the sound activity detection, based on the bold values of the table, it can be realized that only 81 out of 200 features are significant.

In the last phase of the second part, the selected IMF features, represented by the bold values in Table 9, were compared with the corresponding features of all segments of the input clip. The two maximum similarities are marked as the beginning and end of the event. Any false markings in this step should be added to the FP value and subtracted from the TP value. Tables 10 and 11 present the final results of the proposed sound activity detection system.

5.3 Comparison with existing methods

In this section, a comparison with state-of-the-art methods in this field is presented, which confirms the effectiveness of the proposed method. The number of features and the F-score are used in this comparison. The proposed method requires much smaller features (20) than existing ones, making it easier to implement with ML methods. The studies selected for the

Table 8 TN, FP, TP, and FN values of the proposed method after selecting the threshold based on the ROC for each event

Event	Train				Test				Validation			
	TN	FP	TP	FN	TN	FP	TP	FN	TN	FP	TP	FN
Air conditioner	2329	1368	1413	890	740	502	434	324	701	481	512	306
Car horn	2257	1451	1519	773	755	446	528	271	768	489	407	336
Children playing	2253	1473	1320	954	701	533	450	316	773	482	426	319
Dog bark	2395	1282	1427	896	842	370	382	406	801	403	469	327
Drilling	2254	1407	1391	948	694	520	483	303	783	437	429	351
Engine idling	2323	1362	1473	842	818	409	487	286	705	529	527	239
Gunshot	2229	1437	1332	1002	719	494	407	380	815	426	435	324
Jackhammer	2318	1363	1481	838	749	511	419	321	770	501	434	295
Siren	2475	1234	1429	862	843	395	428	334	877	370	459	294
Street music	2304	1380	1405	911	785	472	468	275	729	479	470	322

Table 9 Average absolute value of regularized derivatives at the start and end of the considered events, with values exceeding the threshold level in bold

		A	B	C	D	E	F	G	H	I	J
Average frequency	**IMF 1**	1.272	**1.56**	**1.373**	**1.477**	**1.519**	**1.33**	**1.522**	**1.4**	**1.492**	**1.326**
	IMF 2	**1.339**	**1.57**	**1.435**	**1.546**	**1.476**	**1.444**	**1.54**	**1.457**	**1.505**	**1.43**
	IMF 3	**1.502**	**1.566**	**1.431**	**1.49**	**1.471**	**1.584**	**1.601**	**1.561**	**1.348**	**1.565**
	IMF 4	**1.655**	**1.535**	**1.447**	**1.36**	**1.39**	**1.782**	**1.593**	**1.631**	**1.349**	**1.537**
	IMF 5	**1.748**	**1.466**	**1.41**	1.269	**1.304**	**1.81**	**1.574**	**1.636**	**1.323**	**1.418**
	IMF 6	**1.71**	**1.454**	**1.395**	1.248	1.282	**1.752**	**1.727**	**1.535**	1.298	**1.349**
	IMF 7	**1.511**	**1.416**	**1.378**	1.231	1.229	**1.478**	**1.852**	**1.379**	1.169	1.255
	IMF 8	**1.308**	**1.351**	1.285	1.162	1.179	**1.311**	**1.809**	1.281	1.14	1.194
	IMF 9	1.206	1.288	1.243	1.149	1.14	1.222	**1.655**	1.212	1.118	1.145
	IMF 10	1.136	1.212	1.095	1.076	1.092	1.07	**1.446**	1.096	1.082	1.072
Energy	**IMF 1**	1.192	1.274	1.255	**1.359**	1.298	1.28	1.227	1.231	**1.328**	**1.322**
	IMF 2	1.207	1.231	1.214	**1.354**	**1.376**	1.292	1.288	1.271	1.3	**1.341**
	IMF 3	1.171	1.209	1.142	1.281	**1.401**	1.266	1.247	1.295	1.217	**1.311**
	IMF 4	1.146	1.183	1.138	1.141	**1.401**	1.289	**1.312**	1.202	1.249	1.215
	IMF 5	1.108	1.198	1.119	1.16	**1.456**	1.243	**1.332**	1.157	1.235	1.132
	IMF 6	1.137	1.199	1.173	1.133	**1.421**	1.206	1.256	1.27	1.237	1.191
	IMF 7	1.057	1.142	1.132	1.105	1.249	1.174	1.217	1.156	1.189	1.167
	IMF 8	1.05	1.097	1.085	1.08	1.24	1.114	1.21	1.136	1.127	1.134
	IMF 9	0.989	1.036	1.006	1.033	1.06	1.028	1.084	1.04	1.007	1.008
	IMF 10	0.916	0.905	0.956	0.972	0.978	0.987	0.957	0.957	0.992	0.965

(The events listed are: A - Air conditioner, B - Car horn, C - hildren playing, D - Dog bark, E - Drilling, F - Engine idling, G - Gun shot, H - Jackhammer, I - Siren, and J - Street music).
Best values in bold

Table 10 Final results obtained by the proposed sound activity detection system

Event	Train				Test				Validation			
	TN	FP	TP	FN	TN	FP	TP	FN	TN	FP	TP	FN
Air conditioner	2329	1368	1413	890	740	666	270	324	701	673	320	306
Car horn	2257	1451	1519	773	755	721	253	271	768	692	204	336
Children playing	2253	1473	1320	954	701	715	268	316	773	686	222	319
Dog bark	2395	1282	1427	896	842	524	228	406	801	583	289	327
Drilling	2254	1407	1391	948	694	735	268	303	783	616	250	351
Engine idling	2323	1362	1473	842	818	552	344	286	705	681	375	239
Gunshot	2229	1437	1332	1002	719	660	241	380	815	609	252	324
Jackhammer	2318	1363	1481	838	749	694	236	321	770	661	274	295
Siren	2475	1234	1429	862	843	562	261	334	877	555	274	294
Street music	2304	1380	1405	911	785	619	321	275	729	658	291	322

Table 11 Final values of the evaluation metrics obtained by the proposed sound activity detection system

Sound Class	Test				Validation			
	PR	REC	F1	ACC	PR	REC	F1	ACC
Air conditioner	0.29	0.45	0.35	0.51	0.32	0.51	0.4	0.51
Car horn	0.26	0.48	0.34	0.5	0.23	0.38	0.28	0.49
Children playing	0.27	0.46	0.34	0.48	0.24	0.41	0.31	0.5
Dog bark	0.3	0.36	0.33	0.54	0.33	0.47	0.39	0.55
Drilling	0.27	0.47	0.34	0.48	0.29	0.42	0.34	0.52
Engine idling	0.38	0.55	0.45	0.58	0.36	0.61	0.45	0.54
Gunshot	0.27	0.39	0.32	0.48	0.29	0.44	0.35	0.53
Jackhammer	0.25	0.42	0.32	0.49	0.29	0.48	0.36	0.52
Siren	0.32	0.44	0.37	0.55	0.33	0.48	0.39	0.58
Street music	0.34	0.54	0.42	0.55	0.31	0.47	0.37	0.51

Table 12 Details of the state-of-the-art methods used for comparison purpose

Ref	Type of Feature	Feature Length	Classifier	Segment Length
Huang et al. [44]	Mel spectrogram	64,500	CNN	40ms
Tonami et al. [41]	Mel spectrogram	64,500	CRNN	40ms
Ick and McFee [88]	Mel spectrogram	128,862	CNN	...
Ye et al. [89]	Mel spectrogram	64,500	1D Detection Transformer (1D-DETR)	1s
Dinkel et al. [38]	LMS	64,500	Duration robust CRNN	1s
Pankajakshan et al. [49]	Spectrogram	40,500	CRNN	1s
Proposed method	Energy and frequency of IMFs	20	LSTM	1s

Table 13 Average F-score values obtained by the proposed and state-of-the-art methods on the URBAN-SED dataset in the segment-based event detection task

Reference	Average F1
Huang et al. [44]	0.59
Ick and McFee [88]	0.356
Tonami et al. [41]	0.342
Ye et al. [89]	0.657
Dinkel et al. [38]	0.6475
Pankajakshan et al. [49]	0.4103
Proposed method	0.437

comparison were: [38, 41, 44, 49, 88, 89], which also used the Urban SED dataset in the experiments.

As can be perceived from in Table 12, all the studied related methods used Mel features, and the only difference between them is as to the number of Mel bands and extraction details, such as the overlap between segments or the number of short-time Fourier transform. Regarding the number of features, the proposed method with 20 features has fewer features than any of the other methods. In this study, the average F-score of the proposed method, despite the smaller number of features compared to the state-of-the-art methods, showed an acceptable average F-score in segment-based event detection across the entire dataset (Table 13). In addition, the proposed DL model is more straightforward than those in the compared studies.

As to sound activity detection, [44] and [88] did not propose a method, so in Table 14, the average F-score of the proposed method is compared with those of the methods proposed by Ye et al. [89], Tonami et al. [41], Dinkel et al. [38], and Pankajakshan et al. [49]. The results in Table 13, suggest that sthe proposed method outperforms the methods of Ick et al. [88], Tonami et al. [41], and Pankajakshan et al. [49], while showing lower F-score values compared to the methods of Huang et al. [44], Ye et al. [89] and Dinkel et al. [38]. Based on the results in Table 14, it can be realized that the proposed method outperforms the methods of Ick et al. [88], Tonami et al. [41], and Pankajakshan et al. [49], while showing lower F-score values compared to the method of Ye et al. [89].

An important point to note is that the number of features used in the proposed method is significantly fewer than all the related methods listed in Tables 13 and 14. This indicates that the features employed in this study are highly informative yet concise representations of the spectrograms. It seems that employing IMF spectrograms could lead to developing a SED system with much fewer inputs, reduced computational complexity, and acceptable accuracy compared to existing SED methods in both segment-based event detection and sound activity detection tasks. The aim of this research was not only to enhance the accuracy but also to

Table 14 Average F-score values obtained by the proposed and state-of-the-art methods on the URBAN-SED dataset in the sound activity detection task

Reference	Average F1
Ye et al. [89]	0.3727
Dinkel et al. [38]	0.2254
Pankajakshan et al. [49]	0.1113
Tonami et al. [41]	0.214
Proposed method	0.358

reduce the number of input features concurrently. Considering the obtained results, it appears that the proposed method is quite promising.

6 Conclusion

This study proposed a novel feature extraction approach based on IMF features for SED systems. Since the proposed method has fewer features than the Mel coefficients, the most common feature in this field, it can be easily integrated with conventional ML methods. To prove the effectiveness of the proposed features concerning the average frequency and locally regulated energy of IMFs extracted from sound segments, the features were used as input in LSTM, ensemble learning, and ANN structures, and their efficiency was analized in comparison with state-of-the-art methods proposed in this field. Next, a novel approach for detecting sound activity was proposed based on a statistical analysis of features and detection of changes. The proposed approach uses just the features extracted from the IMF and achieves good results in detecting an event's start and end points. Finally, the proposed method was applied to various events in the URBAN SED dataset, and its effectiveness was demonstrated for both segment-based event detection and sound activity detection. Comparison with state-of-the-art methods proposed in this field showed that the proposed features are as effective as those based on Mel coefficients despite their much smaller number. As a limitation, the effectiveness of IMF as the main part of the proposed method may vary depending on factors such as signal-to-noise ratio, the complexity of the sound events, and environmental conditions. Another limitation of the proposed approach is the small number of extracted IMF features, which can cause issues in training DL methods; for instance, in cases of overfitting, the test and validation sets may show significantly lower accuracy compared to the training data.

Future developments could combine the proposed approach with approaches that use Mel coefficients to improve the efficiency of various SED tasks. In addition, IMF properties can be used in other speech languages such as sound persian phoneme articulation [90, 91] or speech synthesis [92]. The analysis of IMF applicability in SED systems, considering varying levels of signal-to-noise ratios, is another suggested topic for future investigation.

Acknowledgements The first author would like to thank "Fundação para a Ciência e a Tecnologia" (FCT), in Portugal, for his PhD grant with reference 2021.08660.BD. This article is partially a result of the project Sensitive Industry (reference 182852), cofunded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement.

Author Contributions Conceptualization, funding acquisition, and supervision by João Manuel R.S. Tavares; investigation, data collection, code implementation, formal analysis, and original draft preparation by Vahid Hajihashemi and Abdorreza Alavi Gharahbagh; writing review and editing by J.J.M. Machado and João Manuel R.S. Tavares.

Funding Open access funding provided by FCT|FCCN (b-on).

Data Availability This research was developed using a publicly available dataset; the corresponding URL is provided as a footnote in Section 4.1.

Declarations

The authors declare no conflict of interest.

References

1. Basnyat B, Roy N, Gangopadhyay A, Raglin A (2022) Environmental sound classification for flood event detection. In: 2022 18th Int Conf Intell Envir (IE) pp 1–8. IEEE. https://doi.org/10.1109/IE54923.2022.9826766
2. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR et al. (2021) A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion **76**:243–297. https://doi.org/10.1016/j.inffus.2021.05.008
3. Sathesh S, Maheswaran S, Mohanavenkatesan P, Mohammed Azarudeen M, Sowmitha K, Subash S (2022) Allowance of driving based on drowsiness detection using audio and video processing. In: International Conference on Computational Intelligence in Data Science pp 235–250. Springer. https://doi.org/10.1007/978-3-031-16364-7_18
4. Toma A, Cecchinato N, Drioli C, Oliva G, Ferrin G, Sechi G, Foresti GL (2022) Onboard audio and video processing for secure detection, localization, and tracking in counter-uav applications. Procedia Comput Sci**205**:20–27. https://doi.org/10.1016/j.procs.2022.09.003
5. Meng J, Wang X, Wang J, Teng X, Xu Y (2022) A capsule network with pixel-based attention and bgru for sound event detection. Digit Signal Process. **123**:103434. https://doi.org/10.1016/j.dsp.2022.103434
6. Su Y, Zhang K, Wang J, Zhou D, Madani K (2020) Performance analysis of multiple aggregated acoustic features for environment sound classification. Appl Acoust **158**:107050. https://doi.org/10.1016/j.apacoust.2019.107050
7. Katsis LK, Hill AP, Piña-Covarrubias E, Prince P, Rogers A, Doncaster CP, Snaddon JL (2022) Automated detection of gunshots in tropical forests using convolutional neural networks. Ecol Indic **141**:109128. https://doi.org/10.1016/j.ecolind.2022.109128
8. Mushtaq Z, Su S-F (2020) Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Appl Acoust **167**:107389. https://doi.org/10.1016/j.apacoust.2020.107389
9. Hajihashemi V, Alavigharahbagh A, Oliveira HS, Cruz PM, Tavares JMR (2021) Novel time-frequency based scheme for detecting sound events from sound background in audio segments. In: Iberoamerican Congr Pattern Recognit. Springer, pp 402–416. https://doi.org/10.1007/978-3-030-93420-0_38
10. Waldekar S, Saha G (2018) Classification of audio scenes with novel features in a fused system framework. Digit Signal Process **75**:71–82. https://doi.org/10.1016/j.dsp.2017.12.012
11. Ventura TM, Oliveira AG, Ganchev TD, Figueiredo JM, Jahn O, Marques MI, Schuchmann K-L (2015) Audio parameterization with robust frame selection for improved bird identification. Expert Syst Appl **42**(22):8463–8471. https://doi.org/10.1016/j.eswa.2015.07.002
12. Janjua ZH, Vecchio M, Antonini M, Antonelli F (2019) Irese: An intelligent rare-event detection system using unsupervised learning on the iot edge. Eng Appl Artif Intell **84**:41–50. https://doi.org/10.1016/j.engappai.2019.05.011
13. Grzeszick R, Plinge A, Fink GA (2017) Bag-of-features methods for acoustic event detection and classification. IEEE/ACM Trans Audio Speech Lang Process **25**(6):1242–1252. https://doi.org/10.1109/TASLP.2017.2690574
14. Vafeiadis A, Votis K, Giakoumis D, Tzovaras D, Chen L, Hamzaoui R (2020) Audio content analysis for unobtrusive event detection in smart homes. Eng Appl Artif Intell **89**:103226. https://doi.org/10.1016/j.engappai.2019.08.020
15. Hajihashemi V, Gharahbagh AA, Cruz PM, Ferreira MC, Machado JJ, Tavares JMR (2022) Binaural acoustic scene classification using wavelet scattering, parallel ensemble classifiers and nonlinear fusion. Sensors **22**(4):1535. https://doi.org/10.3390/s22041535
16. Nasiri A, Cui Y, Liu Z, Jin J, Zhao Y, Hu J (2019) Audiomask: Robust sound event detection using mask r-cnn and frame-level classifier. In: 2019 IEEE 31st Inte Conf Tools Artif Intell (ICTAI) pp 485–492 (2019). IEEE. https://doi.org/10.1109/ICTAI.2019.00074

17. Soni S, Dey S, Manikandan MS (2019) Automatic audio event recognition schemes for context-aware audio computing devices. In: 2019 Seventh Int Conf Digit Inf Process Commun (ICDIPC) pp 23–28, IEEE. https://doi.org/10.1109/ICDIPC.2019.8723713
18. Hadi M, Pakravan MR, Razavi MM (2019) An efficient real-time voice activity detection algorithm using teager energy to energy ratio. In: 2019 27th Iranian Conference on Electrical Engineering (ICEE) pp 1420–1424, IEEE. https://doi.org/10.1109/IranianCEE.2019.8786643
19. Verma V, Benjwal A, Chhabra A, Singh SK, Kumar S, Gupta BB, Arya V, Chui KT (2023) A novel hybrid model integrating mfcc and acoustic parameters for voice disorder detection. Sci Rep **13**(1):22719. https://doi.org/10.1038/s41598-023-49869-6
20. Savargiv M, Bastanfard A (2016) Real-time speech emotion recognition by minimum number of features. In: 2016 Artif Intell Robot (IRANOPEN) pp 72–76. IEEE. https://doi.org/10.1109/RIOS.2016.7529493
21. Kwon S, *et al.* (2021) Att-net: Enhanced emotion recognition system using lightweight self-attention module. Appl Soft Comput **102**:107101. https://doi.org/10.1016/j.asoc.2021.107101
22. Mustaqeem Kwon S (2019) A cnn-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**(1):183. https://doi.org/10.3390/s20010183
23. Bastanfard A, Amirkhani D, Naderi S (2020) A singing voice separation method from persian music based on pitch detection methods. In: 2020 6th Iran Conf Signal Process Intell Syst (ICSPIS) pp 1–7. IEEE. https://doi.org/10.1109/ICSPIS51611.2020.9349583
24. Shirdel S, Teimoortashloo M, Mohammadiun M, Gharahbagh AA (2023) A hybrid method based on deep learning and ensemble learning for induction motor fault detection using sound signals. Multimed Tools Appl. p 1–19. https://doi.org/10.1007/s11042-023-15996-5
25. Rustam F, Ishaq A, Hashmi MSA, Siddiqui HUR, López LAD, Galán JC, Ashraf I (2023) Railway track fault detection using selective mfcc features from acoustic data. Sensors **23**(16):7018. https://doi.org/10.3390/s23167018
26. Zhang Z, Xu C, Xie J, Zhang Y, Liu P, Liu Z (2023) Mfcc-lstm framework for leak detection and leak size identification in gas-liquid two-phase flow pipelines based on acoustic emission. Measure **219**:113238. https://doi.org/10.1016/j.measurement.2023.113238
27. Mohmmad S, Sanampudi SK (2023) Tree cutting sound detection using deep learning techniques based on mel spectrogram and mfcc features. In: Proceedings of 3rd Int Conf Adv Comput Eng Commun Syst: ICACECS 2022 pp 497–512. Springer. https://doi.org/10.1007/978-981-19-9228-5_42
28. Pandya D, Upadhyay SH, Harsha SP (2013) Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using apf-knn. Expert Syst Appl **40**(10):4137–4145. https://doi.org/10.1016/j.eswa.2013.01.033
29. Gontier F, Lostanlen V, Lagrange M, Fortin N, Lavandier C, Petiot J-F (2021) Polyphonic training set synthesis improves self-supervised urban sound classification. J Acoust Soc Am **149**(6):4309–4326. https://doi.org/10.1121/10.0005277
30. Wang J, Yao P, Deng F, Tan J, Song C, Wang X (2023) Nas-dymc: Nas-based dynamic multi-scale convolutional neural network for sound event detection. In: ICASSP 2023-2023 IEEE Int Conf Acoust Speech Signal Process (ICASSP) pp 1–5. IEEE. https://doi.org/10.1109/ICASSP49357.2023.10096621
31. Jose T, Mayan JA (2023) Real-time sound detection of rose-ringed parakeet using lstm network with mfcc and mel spectrogram. In: 2023 Annu Int Conf Emerg Res Area: Int Conf Intell Syst (AICERA/ICIS) pp 1–6. IEEE. https://doi.org/10.1109/AICERA/ICIS59538.2023.10420143
32. Esmaeilpour M, Cardinal P, Koerich AL (2020) From sound representation to model robustness. arXiv preprint arXiv:2007.13703. https://doi.org/10.48550/arXiv.2007.13703
33. Kong Q, Xu Y, Iqbal T, Cao Y, Wang W, Plumbley MD (2019) Acoustic scene generation with conditional samplernn. In: ICASSP 2019-2019 IEEE Int Conf Acoust Speech Signal Process (ICASSP) pp 925–929. IEEE. https://doi.org/10.1109/ICASSP.2019.8683727
34. Lin L, Wang X, Liu H, Qian Y (2019) Guided learning convolution system for dcase 2019 task 4. arXiv preprint arXiv:1909.06178 p 134–138. https://doi.org/10.33682/53ed-z889
35. Serizel R, Turpault N, Shah A, Salamon J (2020) Sound event detection in synthetic domestic environments. In: ICASSP 2020-2020 IEEE Int Conf Acoust Speech Signal Process (ICASSP) pp 86–90. IEEE. https://doi.org/10.1109/ICASSP40776.2020.9054478
36. Gao L, Mao Q, Dong M (2021) On local temporal embedding for semi-supervised sound event detection. IEEE/ACM Trans Audio Speech Lang Process. https://doi.org/10.1109/TASLP.2024.3369529
37. Nam H, Kim S-H, Ko B-Y, Park Y-H (2022) Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection. arXiv preprint arXiv:2203.15296. https://doi.org/10.48550/arXiv.2203.15296
38. Dinkel H, Wu M, Yu K (2021) Towards duration robust weakly supervised sound event detection. IEEE/ACM Trans Audio Speech Lang Process **29**:887–900. https://doi.org/10.1109/TASLP.2021.3054313

39. Nguyen TNT, Watcharasupat KN, Nguyen NK, Jones DL, Gan W-S (2022) Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection. IEEE/ACM Trans Audio Speech Lang Process **30**:1749–1762. https://doi.org/10.1109/TASLP.2022.3173054
40. Komatsu T, Watanabe S, Miyazaki K, Hayashi T (2022) Acoustic event detection with classifier chains. arXiv preprint arXiv:2202.08470. https://doi.org/10.48550/arXiv.2202.08470
41. Tonami N, Imoto K (2023) Sound event triage: detecting sound events considering priority of classes. EURASIP J Audio Speech Music Process **2023**(1):5. https://doi.org/10.1186/s13636-022-00270-7
42. Johnson DS, Lorenz W, Taenzer M, Mimilakis S, Grollmisch S, Abeßer J, Lukashevich H (2021) Desed-fl and urban-fl: Federated learning datasets for sound event detection. In: 2021 29th Eur Signal Process Conf (EUSIPCO) pp 556–560. IEEE. https://doi.org/10.23919/EUSIPCO54536.2021.9616102
43. Chan TK, Chin CS (2021) Multi-branch convolutional macaron net for sound event detection. IEEE/ACM Trans Audio Speech Lang Process **29**:2972–2985. https://doi.org/10.1109/TASLP.2021.3110649
44. Huang Y, Wang X, Lin L, Liu H, Qian Y (2020) Multi-branch learning for weakly-labeled sound event detection. In: ICASSP 2020-2020 IEEE Int Conf Acoust Speech Signal Process (ICASSP) pp 641–645. IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053023
45. Turpault N, Serizel R (2020) Training sound event detection on a heterogeneous dataset. arXiv preprint arXiv:2007.03931. https://doi.org/10.48550/arXiv.2007.03931
46. Pankajakshan A, Bear HL, Subramanian V, Benetos E (2020) Memory controlled sequential self attention for sound recognition. arXiv preprint arXiv:2005.06650. https://doi.org/10.48550/arXiv.2005.06650
47. Bear HL, Nolasco I, Benetos E (2019) Towards joint sound scene and polyphonic sound event recognition. arXiv preprint arXiv:1904.10408. https://doi.org/10.48550/arXiv.1904.10408
48. De Benito-Gorrón D, Ramos D, Toledano DT (2021) A multi-resolution crnn-based approach for semi-supervised sound event detection in dcase 2020 challenge. IEEE Access. **9**:89029–89042. https://doi.org/10.1109/ACCESS.2021.3088949
49. Pankajakshan, A., Bear, H.L., Benetos, E.: Polyphonic sound event and sound activity detection: A multi-task approach. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 323–327 (2019). IEEE. https://doi.org/10.1109/WASPAA.2019.8937193
50. Martín-Morató I, Mesaros A, Heittola T, Virtanen T, Cobos M, Ferri FJ (2019) Sound event envelope estimation in polyphonic mixtures. In: ICASSP 2019-2019 IEEE Int Conf Acoust Speech Signal Process (ICASSP) pp 935–939. IEEE. https://doi.org/10.1109/ICASSP.2019.8682858
51. Park H, Yun S, Eum J, Cho J, Hwang K (2019) Weakly labeled sound event detection using tri-training and adversarial learning. arXiv preprint arXiv:1910.06790. https://doi.org/10.48550/arXiv.1910.06790
52. Al-Banna A-K, Fang H, Edirisinghe E (2021) A novel attention model across heterogeneous features for stuttering event detection. Expert Syst Appl **244**:122967. https://doi.org/10.1016/j.eswa.2023.122967
53. Turpault N, Wisdom S, Erdogan H, Hershey J, Serizel R, Fonseca E, Seetharaman P, Salamon J (2020) Improving sound event detection in domestic environments using sound separation. arXiv preprint arXiv:2007.03932. p 1–5. https://doi.org/10.48550/arXiv.2007.03932
54. Turpault N, Serizel R, Salamon J, Shah AP (2019) Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In: Workshop on Detection and Classification of Acoustic Scenes and Events. pp 253–257. https://doi.org/10.33682/006b-jx26
55. Hershey S, Ellis DP, Fonseca E, Jansen A, Liu C, Moore RC, Plakal M (2021) The benefit of temporally-strong labels in audio event classification. In: ICASSP 2021-2021 IEEE Int Conf Acoust Speech Signal Process (ICASSP) pp 366–370. IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414579
56. Amarnath M, Krishna IP (2014) Local fault detection in helical gears via vibration and acoustic signals using emd based statistical parameter analysis. Measure **58**:154–164. https://doi.org/10.1016/j.measurement.2014.08.015
57. Zahra A, Kanwal N, Rehman N, Ehsan S, McDonald-Maier KD (2017) Seizure detection from eeg signals using multivariate empirical mode decomposition. Comput Biol Med **88**:132–141. https://doi.org/10.1016/j.compbiomed.2017.07.010
58. Bagherzadeh SA (2018) An improved signal envelope estimation method for analysis of acoustic signals emitted by remotely piloted helicopters. Appl Acoust **135**:8–21. https://doi.org/10.1016/j.apacoust.2018.01.018
59. Cheema A, Singh M (2019) Psychological stress detection using phonocardiography signal: An empirical mode decomposition approach. Biomed Signal Process Control **49**:493–505. https://doi.org/10.1016/j.bspc.2018.12.028
60. Cheema A, Singh M (2019) An application of phonocardiography signals for psychological stress detection using non-linear entropy based features in empirical mode decomposition domain. Appl Soft Comput **77**:24–33. https://doi.org/10.1016/j.asoc.2019.01.006
61. Yao J, Liu C, Song K, Feng C, Jiang D (2021) Fault diagnosis of planetary gearbox based on acoustic signals. Appl Acoust **181**:108151. https://doi.org/10.1016/j.apacoust.2021.108151

62. Ning F, Cheng Z, Meng D, Wei J (2021) A framework combining acoustic features extraction method and random forest algorithm for gas pipeline leak detection and classification. Appl Acoust **182**:108255. https://doi.org/10.1016/j.apacoust.2021.108255
63. Erdoğan YE, Narin A (2021) Covid-19 detection with traditional and deep features on cough acoustic signals. Comput Biol Med **136**:104765. https://doi.org/10.1016/j.compbiomed.2021.104765
64. Vican I, Kreković G, Jambrošić K (2021) Can empirical mode decomposition improve heartbeat detection in fetal phonocardiography signals? Computer Methods and Programs in Biomedicine. **203**:106038. https://doi.org/10.1016/j.cmpb.2021.106038
65. Politis A, Mesaros A, Adavanne S, Heittola T, Virtanen T (2020) Overview and evaluation of sound event localization and detection in dcase 2019. IEEE/ACM Trans Audio Speech Lang Process **29**:684–698. https://doi.org/10.1109/TASLP.2020.3047233
66. Xia X, Togneri R, Sohel F, Huang D (2018) Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection. IEEE Trans Multimed **21**(6):1359–1371. https://doi.org/10.1109/TMM.2018.2879750
67. Liu Y, Zhang E, Jia X, Wu Y, Liu J, Brewer LM, Yu L (2023) Tracheal sound-based apnea detection using hidden markov model in sedated volunteers and post anesthesia care unit patients p 1–10. https://doi.org/10.1007/s10877-023-01015-3
68. Pandey C, Baghel N, Gupta R, Dutta MK (2023) Nocturnal sleep sounds classification with artificial neural network for sleep monitoring. Multimed Tools Appl p 1–17. https://doi.org/10.1007/s11042-023-16190-3
69. Svatos J, Holub J (2023) Impulse acoustic event detection, classification, and localization system. IEEE Trans Instrum Meas **72**:1–15. https://doi.org/10.1109/TIM.2023.3252631
70. Hajihashemi, V., Gharahbagh, A.A., Machado, J., Tavares, J.M.R.: Audio event detection based on cross correlation in selected frequency bands of spectrogram. In: World Conference on Information Systems and Technologies, pp. 182–191 (2023). Springer. https://doi.org/10.1007/978-3-031-45651-0_19
71. Hajihashemi V, Gharahbagh AA, Machado J, Tavares JMR (2023) Audio event detection based on cross correlation in selected frequency bands of spectrogram. In: World Conf Infor Syst Technol pp 182–191. Springer. https://doi.org/10.1007/978-3-031-45651-0_19
72. Phinyomark A, Thongpanja S, Hu H, Phukpattaranont P, Limsakul C (2012) The usefulness of mean and median frequencies in electromyography analysis. Computational intelligence in electromyography analysis-A perspective on current applications and future challenges **81**:67
73. Bengio Y, Frasconi P (1993) Credit assignment through time: Alternatives to backpropagation. Adv Neural Inf Process Syst **6**
74. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw **5**(2):157–166. https://doi.org/10.1109/72.279181
75. Lang, KJ, Waibel AH, Hinton GE (1990) A time-delay neural network architecture for isolated word recognition. Neural Netw **3**(1):23–43. https://doi.org/10.1016/0893-6080(90)90044-L
76. Lin T, Horne B, Tiňo P, Giles C (1995) Learning long-term dependencies is not as difficult with narx networks. Adv Neural Inf Process Syst **8**
77. Plate TA (1992) Holographic recurrent networks. Adv Neural Inf Process Syst **5**
78. Mozer MC (1991) Induction of multiscale temporal structure. Adv Neural Inf Process Syst **4**
79. Schmidhuber J (1992) Learning complex, extended sequences using the principle of history compression. Neural Comput **4**(2):234–242. https://doi.org/10.1162/neco.1992.4.2.234
80. Hochreiter S, Schmidhuber J (1996) Lstm can solve hard long time lag problems. Adv Neural Inf Process Syst **9**
81. Mustaqeem Kwon S (2020) Clstm: Deep feature-based speech emotion recognition using the hierarchical convlstm network. Math **8**(12):2133. https://doi.org/10.3390/math8122133
82. Wang L, Cao H, Yuan L (2022) Gated tree-structured recurnn for detecting biomedical event trigger. Appl Soft Comput **126**:109251. https://doi.org/10.1016/j.asoc.2022.109251
83. Muosa AH, Ali A 920220 Internet routing anomaly detection using lstm based autoencoder. In: 2022 Int Conf Comput Sci Softw Eng (CSASE) pp 319–324. IEEE. https://doi.org/10.1109/CSASE51777.2022.9759613
84. Zhou F, Zhang Z, Chen D (2021) Real-time fault diagnosis using deep fusion of features extracted by parallel long short-term memory with peephole and convolutional neural network. Proceedings of the Institution of Mechanical Engineers, Part I: J Syst Control Eng **235**(10):1873–1897. https://doi.org/10.1177/0959651820948291
85. Salamon J, MacConnell D, Cartwright M, Li P, Bello JP (2017) Scaper: A library for soundscape synthesis and augmentation. In: 2017 IEEE Work Appl Signal Process Audio Acoust (WASPAA) pp 344–348. IEEE. https://doi.org/10.1109/WASPAA.2017.8170052

86. Mesaros A, Heittola T, Virtanen T (2016) Metrics for polyphonic sound event detection. Appl Sci **6**(6):162. https://doi.org/10.3390/app6060162
87. Ebbers J, Haeb-Umbach R (2021) Self-trained audio tagging and sound event detection in domestic environments. In: Proc 6th Detect Classif Acoust Scenes Events 2021 Work (DCASE2021)
88. Ick C, McFee B (2021) Sound event detection in urban audio with single and multi-rate pcen. In: ICASSP 2021-2021 IEEE Int Conf Acoust Speech Signal Process (ICASSP) pp 880–884. IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414697
89. Ye Z, Wang X, Liu H, Qian Y, Tao R, Yan L, Ouchi K (2021) Sound event detection transformer: An event-based end-to-end model for sound event detection. arXiv preprint arXiv:2110.02011. https://doi.org/10.48550/arXiv.2110.02011
90. Bastanfard A, Kelishami AA, Fazel M, Aghaahmadi M () A comprehensive audio-visual corpus for teaching sound persian phoneme articulation. In: 2009 IEEE Int Conf Syst Man Cybernet pp 169–174. IEEE. https://doi.org/10.1109/ICSMC.2009.5346591
91. Bastanfard, A., Fazel, M., Kelishami, A.A., Aghaahmadi, M.: The persian linguistic based audio-visual data corpus, ava ii, considering coarticulation. In: Advances in Multimedia Modeling: 16th Int Multimed Model Conf MMM 2010, Chongqing, China, January 6-8, 2010. Proceedings 16, pp 284–294. Springer. https://doi.org/10.1007/978-3-642-11301-7_30
92. Savargiv M, Bastanfard A (2014) Study on unit-selection and statistical parametric speech synthesis techniques