Research Article

Ester Vidaña-Vila*, Giovanni Brambilla, Rosa Ma Alsina-Pagès

Sound event detection by intermittency ratio criterium and source classification by deep learning techniques

https://doi.org/10.1515/noise-2024-0014
received July 24, 2024; accepted December 09, 2024

Abstract: Urban environments are characterized by a complex interplay of various sound sources, which significantly influence the overall soundscape quality. This study presents a methodology that combines the intermittency ratio (IR) metric for acoustic event detection with deep learning (DL) techniques for the classification of sound sources associated with these events. The aim is to provide an automated tool for detecting and categorizing polyphonic acoustic events, thereby enhancing our ability to assess and manage environmental noise. Using a dataset collected in the city center of Barcelona, our results demonstrate the effectiveness of the IR metric in successfully detecting events from diverse categories. Specifically, the IR captures the temporal variations of sound pressure levels due to significant noise events, enabling their detection but not providing information on the associated sound sources. To fill this weakness, the DL-based classification system, which uses a MobileNet convolutional neural network, shows promise in identifying foreground sound sources. Our findings highlight the potential of DL techniques to automate the classification of sound sources, providing valuable insights into the acoustic environment. The proposed methodology of combining the two above techniques represents a step forward in automating acoustic event detection and classification in urban soundscapes and providing important information to manage noise mitigation actions.

*** Corresponding author: Ester Vidaña-Vila,** Human-Environment Research (HER), La Salle, Universitat Ramon Llull – c/Sant Joan de la Salle, 42, 08022 Barcelona, Catalonia, Spain, e-mail: ester.vidana@salle.url.edu
Giovanni Brambilla: Department of Earth and Environmental Sciences (DISAT), University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126, Milan, Italy, e-mail: giovanni.brambilla@artov.inm.cnr.it
Rosa Ma Alsina-Pagès: Human-Environment Research (HER), La Salle, Universitat Ramon Llull – c/Sant Joan de la Salle, 42, 08022 Barcelona, Catalonia, Spain

Keywords: intermittency ratio, sound event detection, convolutional neural network, sound source identification, urban noise

1 Introduction

Sound perception depends on several factors, not only the acoustical ones. Among these, the temporal structure of sound pressure level (SPL) plays an important role, since human hearing tends to adapt to steady sounds, but remains very sensitive to SPL fluctuations over time, as well as to prominent, salient noise events. Unfortunately, these events occur very often in real life, such as those generated by transport systems, *e.g.*, car pass-bys, which can evoke increased annoyance. Thus, peak values and relative SPL changes can be important in noise perception and its influence on non-auditory health effects, such as annoyance and sleep disturbance. To grasp and describe these SPL fluctuations, several methods have been proposed so far. Many of them look at transients in SPL time history, such as exceedances above fixed or time adaptive thresholds [1–4]; others focus on modeling the hearing perception process of such events [5].

A review of the wide range of algorithms, protocols, or criteria reported in the literature for identifying noise events in the time series of A-weighted SPLs is given by Brown and De Coensel [2]. A small set of parameters was identified [3], which may prove useful in the construction of event-based indicators supplementary to energy-equivalent measures (*i.e.*, L_{Aeq}). A further approach is the detection of noise "notice-events," that is those clearly perceivable and, therefore, potentially affecting exposed people. On this issue, the model proposed in [5] considers aspects of human auditory perception, such as attention strength and habituation to time constants; it is grounded in the hypothesis that long-term perception of environmental sound is determined primarily by short notice events. Thus, the detection of noise events is strongly required to guide noise mitigation actions and clearly demands automatic procedures [6,7].

However, the detection of noise notice-events *per se* is not sufficient for an efficient noise mitigation planning aimed at improving or protecting the quality of the sonic environment. For this purpose, a further issue needs to be addressed: the recognition of the source generating the sound event. Several studies on soundscape have shown that the human response to sound events depends not only on SPL, but also on the type of noise source. For instance, natural sources are rated more acceptable than the mechanical ones. Once more, automatic procedures able to detect the type of sound source, such as distinguishing road traffic from other sources, are strongly needed [8,9]. The procedure developed in [10] showed promising results and it was applied in some noise monitoring networks [11].

A step toward meeting these needs is outlined in the European Noise Directive (END, Directive 2002/49/EC) [12], which mandates EU Member States to create strategic noise maps and action plans for major agglomerations, roads, railways, and airports. A key requirement of the END is to identify and classify noise sources contributing to environmental noise [13], enabling targeted mitigation strategies. Traditional noise mapping techniques do not face this level of granularity required by the END, highlighting the need for innovative approaches that integrate temporal and source-specific analyses.

Within the above issue, this article presents the application of the criterium proposed by the intermittency ratio metric (IR) [4] for acoustic event detection together with deep learning (DL) techniques for the classification of source(s) producing such events (*e.g.*, automatically categorizing the detected events by convolutional neural network (CNN) [14]). The integration of these methods provides a dual-layered analysis – temporal and categorical – that offers more and deeper insights into urban noise. The IR metric identifies prominent noise events based on their temporal characteristics and evolution, while the DL model assigns these events to specific noise sources, widening its comprehension. This synergy aligns directly with the objectives of the END, enabling enhanced noise mapping and more effective noise mitigation actions.

This article is organized as follows: Section 2 explains the methodology proposed to combine the IR for acoustic event detection and the DL model for classification. First, it explains the datasets that have been used, and later it provides details about the application of the IR and the DL models over the data. Next, Section 3 describes the results obtained when applying the proposed methodology over the data, both in terms of acoustic event detection and classification. Finally, Section 4 concludes the article, highlighting the findings of the study and future work directions.

2 Materials and methods

This subsection first explains the methodology proposed in this article in Subsection 2.1. Then, Subsection 2.2 details the datasets used for the study. Next, Subsection 2.3 explains how the IR algorithm has been used for sound event detection. Finally, Subsection 2.4 explains the DL-based algorithm used for sound source classification.

2.1 Proposed methodology

The proposed methodology is outlined in Figure 1. The IR criterium used for sound event detection has been applied on the time history of the SPL data. Thus, the first step requires converting the audio recording file into a time series of SPL values with a preset time resolution. Then, the criterium of sound event detection selects all the SPL values exceeding the preset threshold (Equation (2)). Afterwards, a deep CNN would provide the probability of a source type associated to the event, considering that multiple events can overlap in an urban acoustic environment.

2.2 Dataset used

The dataset used in this article is the one described by Vidaña-Vila [15], composed of audio recordings taken in the city center of Barcelona during two campaigns. The first one took place in Autumn 2020 at noon, when there were some mobility restrictions due to COVID-19 pandemic; the second one was performed in Spring 2021 during the afternoon, when the mobility restrictions were softened. The dataset contains a rich variety of acoustic data, as the soundscape was slightly different due to the effect of: i) the COVID-19 restrictions, ii) the season of the year when data was recorded, and iii) the hour of recording.

Four sensors were deployed at four different spots of a street intersection, the crossroad between Villarroel Street and Diputació Street (Figure 2) located in the Eixample area of Barcelona. The rationale behind the decision of locating the four sensors on a crossroad was that the

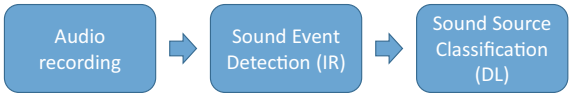

Figure 1: Outline of the methodology proposed to detect and classify the sound events.

four sensors would capture similar loud events (an acoustic event that is loud enough should be recorded in the four spots) while the background sounds would be different in the four spots, enabling us to study the benefit of placing more than one sensor on a street in terms of increasing the accuracy at the time of performing acoustic event classification or not [15]. For this work, this physical redundancy of sensors (*i.e.*, more than one sensor capturing the same acoustic event from a different perspective) is not relevant, as it is not the topic under study. Therefore, using data from the four sensors will just enable us to have a larger and richer variety of data, as different events were occurring in the background of the four locations.

The hardware used to collect the audio data was a Zoom H5 recorder mounted on a tripod. The device was powered with two Alkaline batteries, enabling to have an autonomy of a few hours, which was enough for the duration of the recording campaign. The microphone used was the one provided by the same Zoom recorder: the XYH-5 microphone capsule with a windscreen [16]. This microphone capsule is composed of two condenser and unidirectional microphones paired at 90° angle, enabling stereo recordings. Therefore, the recordings obtained in the campaign are stereo. The sampling rate of the recordings was 44,100 Hz bit 16-bit depth.

Figure 3: Example of a recorder used mounted over a tripod.

As shown in Figure 3, the recorder was placed at an angle of 45° from the floor and pointing to the street.

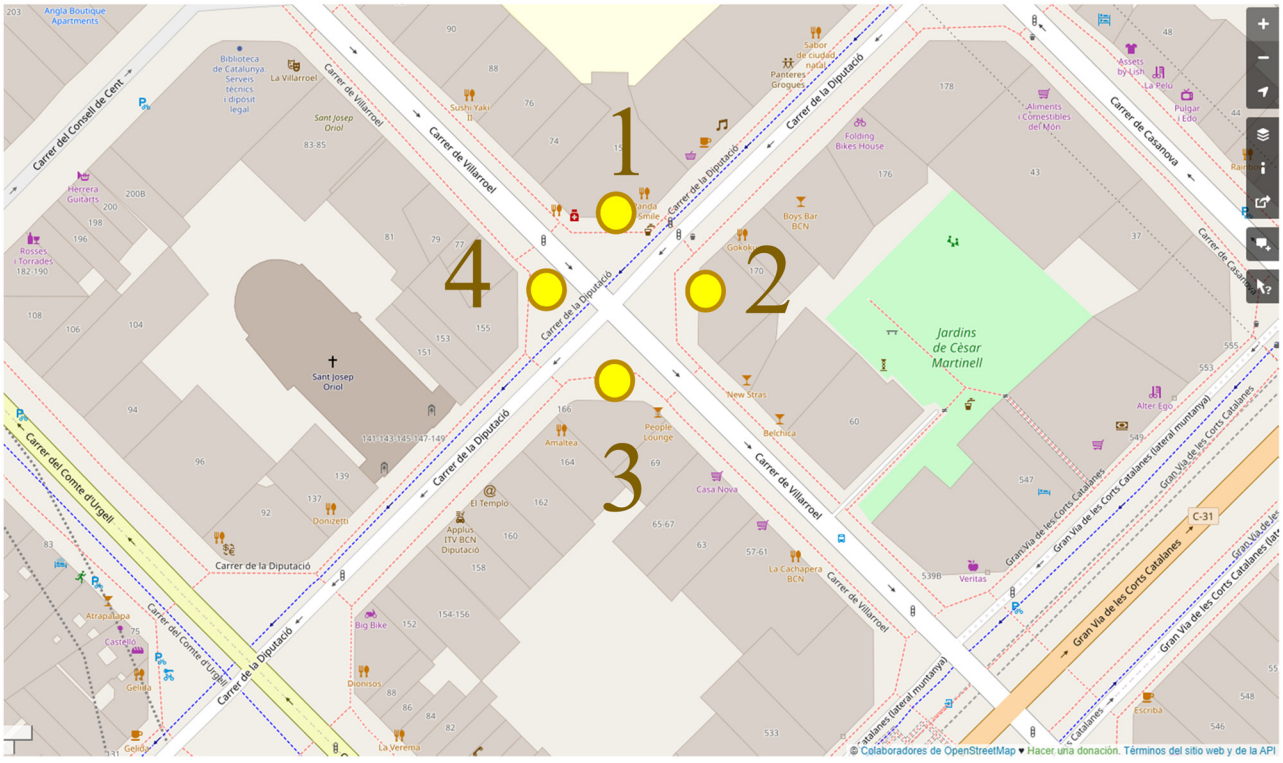

Figure 2: Crossroad in the Eixample area of Barcelona and sound recording spots selected for data collection (edited from OpenStreetMaps).

Table 1: The predetermined 21 categories of the sound sources and corresponding labels of the Eixample dataset

Source category description	Label	Number of occurrences
Background traffic noise	rtn	4,295
Noise from people	peop	912
Car brakes	brak	913
Bird twittering	bird	1,317
Motorcycles	motorc	1,334
Engine idling	eng	1,116
Car door slamming	cdoor	294
Undefined impulsive noise	impls	615
Undefined complex noise	cmplx	158
Trolley	troll	314
Wind	wind	31
Car or motorbike horn	horn	76
Sirens from vehicles	sire	75
Music	musi	38
Bicycle	bike	75
House door	hdoor	85
Bells	bell	51
Waste into the recycling bin	glass	49
Beep from truck on reversing	beep	31
Dog barking	dog	28
Drilling	drill	14

Putting together the acoustic data from both campaigns in all the four recording spots, the dataset has a duration of 5 h in each sensing node, resulting in an aggregated amount of data of 20 h. More details on measurement and recording setup are reported by Vidaña-Vila [15].

In this work, data coming from the different sensors were processed individually, differently to the usage of the dataset in the study by Vidaña-Vila [15]. For the experimental evaluation, 4 h from each sensor (together with data coming from external sources) were used to train the DL model, whereas the remaining hour was used for testing. The detection of sound events was carried out only for the testing hour. Even though the recordings are in stereo mode, only the left channel of them was used, turning them into mono audio files. The decision of conserving just one channel was for simplicity purposes, making it easier to label the events using only one channel. The sounds contained in this dataset are the ones shown in Table 1.

Two complimentary datasets were used for training the automatic classifier CNN, namely the BCNDataset [17] and the UrbanSound8K dataset [18]. On the one hand, the BCNDataset was selected as it contains acoustic data from the same city like the evaluation dataset (the Eixample dataset), and on the other hand, UrbanSound8K dataset was used because all the sounds that are contained there can be heard very clearly, and can be considered foreground sounds.

The events present in the BCNDataset are similar to the ones present in the Eixample dataset. Actually, some labels are the same, such as *rtn, sire, horn, brak, troll, peop* or *musi*. Then, there are categories that are slightly different in both datasets but have a similar meaning in terms of sound. For those cases, the labels were unified to match the taxonomy of the Eixample dataset. These specific labels can be checked on Table 2. Those categories that did not contain acoustic events that could match the ones on the Eixample dataset, specifically the *busd* category, were discarded and tagged with a '–'.

For the UrbanSound8k dataset, the labels were unified as shown in Table 3. Those categories that did not contain acoustic events that could match the ones on the Eixample dataset were discarded and tagged with a '–'.

Table 2: Unification of labels from the BCNDataset to the Eixample dataset

Source category description	Label in the BCNDataset	Label in the Eixample dataset	Number of occurrences
A mix of background city noise and music	bkmu	musi	18
The opening or closing of blinds	blin	cmplx	20
People coughing	coug	peop	20
House door	door	hdoor	731
Whistle	whtl	peop	8
Unrecognizable noise	rare	cmplx	671
Background traffic noise	rtn	rtn	1,150
Sirens from vehicles	sire	sire	9
Car or motorbike horn	horn	horn	93
Bus door	busd	—	—
Car brakes	brak	brak	810
Trolley	troll	troll	11
Noise from people	peop	peop	2,256
Music	musi	musi	16

Table 3: Unification of labels from the UrbanSound8K to the Eixample dataset

Source category description	Label in UrbanSound8k	Label in the Eixample dataset	Number of occurrences
Noise from air conditioner	air_conditioner	—	—
Car horn	car_horn	horn	429
Children playing	children_playing	peop	1,000
Dog barking	dog_bark	dog	1,000
Drilling noise	Drilling	drill	1,000
Engine idling	engine_idling	eng	1,000
Gun shot	gun_shot	—	—
Jackhammer	jackhammer	—	—
Siren	siren	sire	929
Street music	street_music	musi	1,000

As it can be observed, adding the UrbanSound8K dataset and BCNDataset to the training set helps mitigating the class imbalance of data in some categories (such as car horns, sirens or street music). However, there are some categories that still have a few amount of samples such as bell sounds (with only 51 samples), beep sounds (with 31 samples), bike sounds (with 75 samples) or wind sounds (with 31 samples). The lack of data in this categories can be therefore mitigated by applying data augmentation techniques such as mix-up augmentation.

2.3 Sound event detection

The digital audio recordings in the ".wav" format have been processed to get the A-weighted SPL time history with Fast (F) weighting according to the IEC 61672-1 standard (exponential integration time τ = 125 ms, that is 8 SPL values per second, 28,800 SPL values in 1 h of recording). An example of such processing is shown in Figure 4, where the blue, red and green dotted lines report the equivalent continuous level L_{Aeq} and the percentile levels L_{A10} and L_{A95} hourly values, respectively. This plot is among the output of a script developed in the R environment [19] to process the L_{AF} time history in order to detect the sound events and to determine the acoustic descriptors given in Table 4.

For the sound event detection task, the criterium used in the metric termed IR has been applied [4]. In particular, the IR value [20], in percentage, is calculated as the ratio of the event-based sound energy $L_{\mathrm{Aeq},T,\mathrm{events}}$ to the overall sound energy $L_{\mathrm{Aeq},T}$, both referred to the time T:

$$\mathrm{IR} = \frac{10^{(L_{\mathrm{Aeq},T,\mathrm{events}}/10)}}{10^{(L_{\mathrm{Aeq},T}/10)}} \times 100[\%]. \tag{1}$$

The time T can be 1 h, the day and/or night period, 24 h and so forth. A sound event is detected whenever its A-weighted SPLs exceed the preset threshold K:

$$K = L_{\mathrm{Aeq},T} + C[\mathrm{dB(A)}], \tag{2}$$

where the constant C is to be set. On the basis of practical experience on transportation noise situations, the Authors proposing IR suggest that C might not be smaller than 0 and not larger than about 10 dB [4]. For low values of C, almost any situation produces a large IR value, whereas high values of C almost always produce low IR, because only in extraordinarily intermittent situations the level rises above the high threshold. To make IR able to distinguish between situations with different degrees of intermittency, the criterium for setting C would be a preferably uniform spread of IR across the range of exposure situations as they occur in the real world. The balance between these extreme

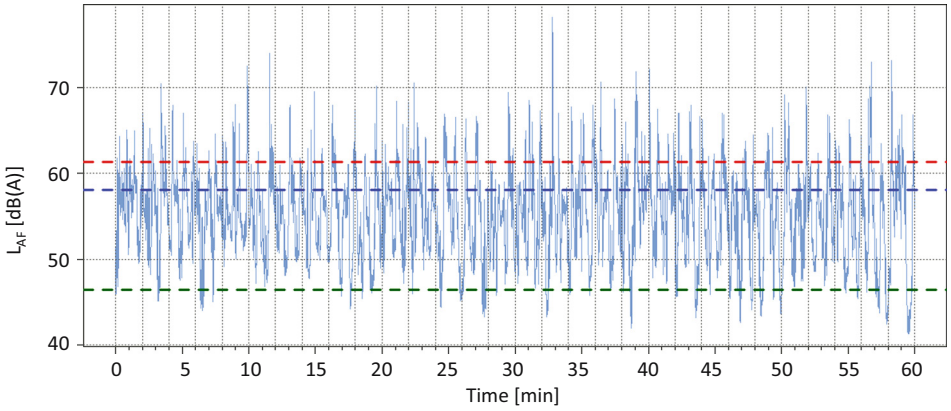

Figure 4: Example of the L_{AF} hourly time history obtained from digital audio recording processing (spot 4). Blue, red and green dotted lines correspond to the equivalent continuous level (L_{Aeq}) and the percentile levels L_{A10} and L_{A95} hourly values, respectively.

Table 4: Acoustic descriptors calculated by the script developed in the R environment

Equivalent continuous sound level L_{Aeq} [dB(A)]	Standard deviation of sound level sL_A [dB(A)]
5th percentile of sound level L_{A5} [dB(A)]	Intermittency ratio IR [%]
10th percentile of sound level L_{A10} [dB(A)]	Onset of sound level for each event [dB(A)]
50th percentile of sound level L_{A50} [dB(A)]	Onset rate of sound level for each event [dB(A)/s]
90th percentile of sound level L_{A90} [dB(A)]	Sound exposure level SEL of each event [dB(A)]
95th percentile of sound level L_{A95} [dB(A)]	Continuous equivalent sound level L_{Aeq} of each event [dB(A)]

cases was investigated by numerical simulations of various traffic situations and resulted in C = 3 dB [4]. This has been the value applied in the present study regardless of the identified sound sources and because mostly of the events were from mixed road traffic.

An IR > 50% means that more than half of the sound dose is caused by "distinct" sound events. In situations with only events that clearly emerge from background noise (*e.g.*, a receiver point close by a railway track), IR yields values close to 100%.The IR metric provides information on the noise temporal structure and can be fruitfully added to the noise energy content $L_{Aeq,T}$ to describe the potential harmful effects on the exposed population. For instance, in the SiRENE study, the IR values were included in the façade noise maps computed for all dwellings in Switzerland [4]. Moreover, the IR provide interesting results to classify urban roads on the basis of traffic noise features [21].

An example of sound event detection is given in Figure 5, where the parameters of sound event(s), reported with red lines, are given as follows:

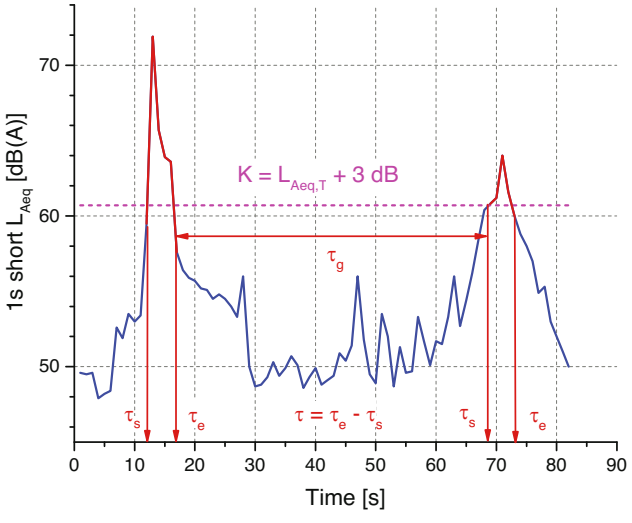

Figure 5: Parameters of the sound event detection based on IR algorithm (adapted from the study by Alsina-Pagès [7]).

- τ_s and τ_e are the start and end time of the event, corresponding to the instants immediately before and just after the exceedances of the threshold K (Equation (2)) above which an event (red line) is detected accordingly to the IR criterium [4];
- $\tau = \tau_e - \tau_s$ is the event duration;
- τ_g is the time gap between two consecutive events;

For any detected event, the following parameters were also determined (Figure 5):

- τ_{max}, the instant at which the maximum SPL of the event occurs (SPL_{max});
- the onset O of the SPL O = $SPL_{max} - SPL_s$, where SPL_s is the SPL corresponding to the start time τ_s of the event;
- the onset rate OR of the SPL OR = $\frac{O}{(\tau_{max} - \tau_s)}$;
- the sound exposure level (SEL), corresponding to all the acoustic energy of the sound event as if this had occurred within a 1-s time period:

$$SEL = 10\ \log\left(\sum_{i=\tau_s}^{\tau_e} 10^{(SPL_i/10)}\right)[dB(A)]. \quad (3)$$

- the equivalent continuous level (L_{Aeq}):

$$L_{Aeq} = SEL - 10\ \log\ \tau = SEL - 10\ \log(\tau_e - \tau_s)[dB(A)]. \quad (4)$$

Figure 6 shows an example of the plot obtained by the developed R script, where:

- the blue dotted line reports L_β, that is the threshold applied for the sound event detection according to the IR criterium;
- blue crosses correspond to the maximum sound level of each event with duration $\tau \geq 1$ s and onset O ≥ 5 dB (A), usually considered as "notice-event" in terms of perception;
- red circles correspond to the maximum sound level of each event with duration $\tau \geq 1$ s, onset O ≥ 5 dB(A) and onset rate OR ≥ 10 dB/s, respectively; some standards, like the NORDTEST Method [21], classify a sound as impulsive when the onset rate OR ≥ 10 dB(A)/s.

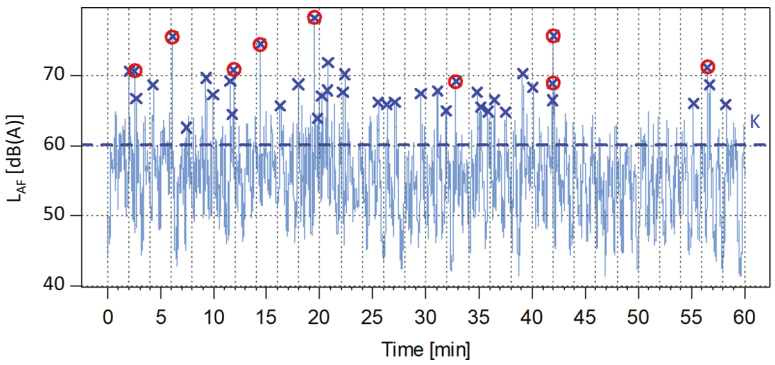

Figure 6: Example of the L_{AF} hourly time history (spot 3) with L_β the threshold applied for the sound event detection according to the IR criterium at K threshold (blue dotted line), maximum sound level of each event with duration $\tau \geq 1$ s and onset O ≥ 5 dB(A) (blue cross) and onset rate OR ≥ 10 dB/s (red circle).

2.4 Sound source classification

The CNN performs a polyphonic classification after the pre-processing stage of the raw audio files and the training process of the neural network.

2.4.1 Audio pre-processing

As inputs of the CNN, the audio files have been converted to spectrograms. Spectrograms are graphic representations of sound with *x*-axis representing time and *y*-axis the frequency and colored scale indicating the SPL. The spectrograms were calculated using the LibROSA Python library, version 0.8.1 [22]. The spectrograms of the raw audio files were determined without applying any pre-processing technique, such as noise reduction, to enable the CNN to classify not only the events in the foreground, but also the events in the background. The recording was windowed in fragments of 4 s each and, then, the short-time Fourier transform was calculated, which computes the discrete Fourier transforms over short overlapping windows [22]. Once the spectrograms were calculated, they are resized to a resolution of 224 × 224, aiming to maintain compatibility with the chosen CNN.

The 4 s windowing was necessary because one of the datasets used for training (the Eixample dataset) is labelled with weak labels of a 4 s resolution. Differently from strong labels, where each acoustic event is defined by the start and end time, weak labels only indicate the presence or absence of a source category in each fragment. In the Eixample dataset, this weak labelling process was carried out by fragments of 4 s, meaning that every 4 s there is a tag with all the categories that are present on it, regardless of their saliency.

A sample spectrogram is shown on Figure 7. Specifically, the spectrogram on the left shows a 4-s fragment containing the sounds of road traffic noise, a car brake, a car horn, and an impulsional sound. Therefore, this spectrogram contains 4 categories as labels, but the position of them inside of the spectrogram is not known. Similarly, the

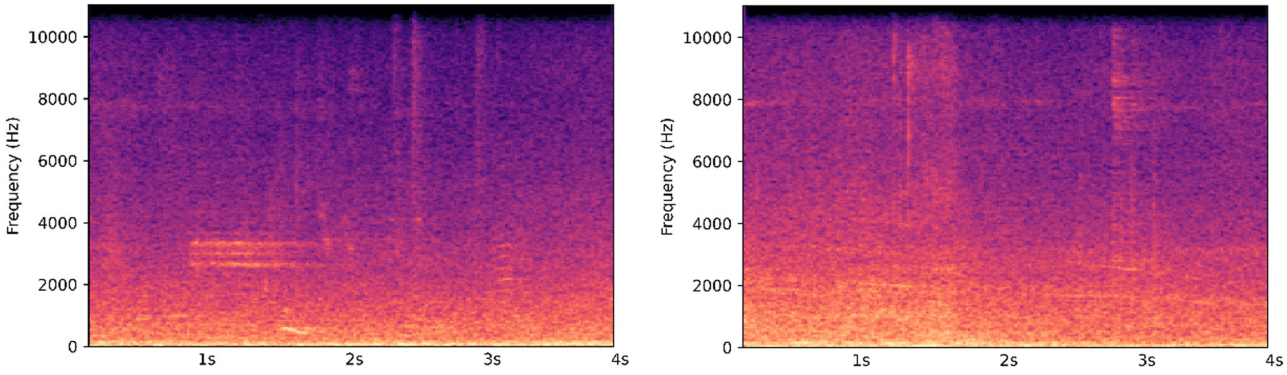

Figure 7: Example of two spectrograms. The left one contains noises from road traffic noise, a car brake, a car horn and an impulsional noise. The right one contains road traffic noise, car brakes and people talking.

spectrogram on the right contains three sounds: road traffic noise, car brakes and people talking.

The other two datasets used for training the neural network, namely the BCNDataset and the UrbanSound8k dataset, are labelled with strong labels.

The sounds of the UrbanSound8k dataset are not masked by other categories and do not enable to perform polyphonic classification directly, as each fragment contains an event of only one category. The fact that UrbanSound8k dataset contains fragments of durations of 4 s (or, in some cases, shorter), makes the selected window length a convenient size to work with. Those fragments that had a duration shorter than the selected size were synthetically modified and repeated until the 4-s duration was achieved.

For the BCNDataset, as the audio fragments contain strong labels, the labels were converted to weak labels by simply splitting the audio files into fragments of 4 s and aggregating all the labels contained in that amount of time.

The spectrograms of all the datasets were obtained using the same parameters. Then, each spectrogram was normalized according to its maximum and minimum value, in such a way that each spectrogram contained only values ranging between 0 and 1.

2.4.2 Sound source classification using DL

Once all the spectrograms were obtained using the same parameters, each audio fragment of 4 s was characterized by a matrix of 224 × 224 positions, and values ranging from 0 to 1. For the acoustic event classification task, and similarly to the work performed by Vidaña-Vila [15], a PyTorch [23] implementation of the MobileNet v2 architecture [24] was selected. This architecture has been used in other works and has proved to achieve good classification results using acoustic data. For instance, in the study by Chen et al. [25], this architecture was used with the UrbanSound8K dataset achieving a classification accuracy over 95%. The authors of the article have also used it in previous works [15,26]. Specifically, in the study by Vidaña-Vila [15], different combinations of the three datasets employed in this article were used to evaluate the classification of acoustic events.

To maintain compatibility with the selected architecture, which expects an RGB image at the input, each spectrogram matrix was replicated 3 times, creating a greyscale image.

As for the training data, the network was trained for 50 epochs with the datasets described in Section 2.2. That is, for training, data from three different sources was mixed:

- UrbanSound8K dataset, fully used for training. This way, training data contained clear examples of the categories that compose this dataset.
- BCNDataset was used for training as well. This dataset contains real-world data collected in a balcony in the city center of Barcelona (also in the Eixample district), meaning that it contains similar sounds to the ones that can be found in the Eixample dataset.
- Eixample dataset, namely, the data collected at the crossroad in the Eixample area of Barcelona; 4 h of recordings per sensor were used for training, while the remaining hour was left for testing of the algorithm.

The testing data were the hour that was excluded from the Eixample dataset, which was used both to evaluate the event detection using the IR and the event classification using the CNN.

3 Results

This Section is divided in two: first, the results obtained using the IR for sound event detection are presented in Subsection 3.1. Then, the source identification results obtained using the CNN are shown in Subsection 3.2.

3.1 Sound event detection

For each 1-h time history of A-weighted SPL, the overall value of the acoustic descriptors calculated by the script developed in the R environment (Table 4) is given in Table 5. The difference of values among the recording spots are small, for instance 1 dB(A) for L_{Aeq} and 0.6 dB(A) for L_{A90}, usually representing the background sound level. The IR values range from 43.7 to 49.0%, indicating that the sound energy of the events roughly contributes to half of the overall energy. The noise climate, determined by the difference $L_{A10} - L_{A90}$, ranges

Table 5: Overall values of acoustic descriptors of each A-weighted SPL 1h-time history

Spot	L_{Aeq} [dB (A)]	L_{A5} [dB (A)]	L_{A10} [dB (A)]	L_{A50} [dB (A)]	L_{A90} [dB (A)]	L_{A95} [dB (A)]	sL_A [dB (A)]	IR [%]
1	57.7	62.8	61.1	54.9	48.4	46.7	4.9	49.0
2	58.2	63.5	61.5	55.5	47.8	46.2	5.2	47.9
3	57.2	62.0	60.5	55.1	47.8	46.0	4.9	43.7
4	58.1	63.4	61.3	55.2	48.0	46.4	5.2	49.0

Table 6: Number and descriptors of sound events detected in each A-weighted SPL 1h-time history

Spot	1		2		3		4	
Type of event	**NE**	**IE**	**NE**	**IE**	**NE**	**IE**	**NE**	**IE**
Number	40	2	54	4	42	9	62	1
Overall τ [s]	191.125	2.625	168.875	4.750	161.500	11.375	183.250	1.875
$\overline{\tau}$ [s]	4.8	1.3	4.2	1.2	3.8	1.3	4.4	1.875
$\overline{L_{\mathrm{A\,max}}}$ [dB(A)]	68.4	73.1	68.1	70.3	68.4	72.8	68.0	78.2
$\overline{O}$ [dB(A)]	8.8	16.0	8.0	16.9	9.8	18.0	7.6	24.5
$\overline{\mathrm{OR}}$ [dB(A)/s]	3.1	12.2	3.1	14.3	5.2	14.4	2.5	13.1
$\overline{\mathrm{SEL}}$ [dB(A)]	79.9	79.1	79.7	76.9	78.4	78.3	79.5	86.3
$\overline{L_{\mathrm{Aeq}}}$ [dB(A)]	64.9	68.5	65.0	66.7	64.5	67.9	64.7	74.3

NE, notice-event; IE, impulsive event.

from 12.7 to 13.7 dB(A), indicating large SPL time variability, as also reported by sL_A.

Based on the IR criterium for sound event detection, namely a sound event is detected whenever its A-weighted SPLs exceed the preset threshold K (Equation (2)), the number of events detected in each 1h-recording are reported in Table 6, considering only the sound events with duration $\tau \geq 1$ s and onset O ≥ 5 dB(A), usually considered as "notice-event" (NE) in terms of perception, and those classified as impulsive (IE) because onset rate OR ≥ 10 dB(A)/s. The average values of some descriptors of the events are also given.

The probability density function of the SEL values is given in Figure 8 for each of the four recording spots. Spots 3 and 4 show a very similar distribution, whereas they are different for spots 1 and 2, the latter showing a bimodal distribution. The dotted lines correspond to the mode value for each distribution.

The R script provides also the A-weighted SPL time history of each detected event. Figure 9 reports such an example where the horizontal red dotted line corresponds to the threshold for the event detection L_β and the inclined blue line connects the SPL at the start time with the maximum SPL of the event.

Referring to the sound recording fragmentation into 4 s windows, the detected sound events were often non-completely included in one 4 s window, because their length was longer. Table 7 reports the percentage of detected notice-events NE spanning over adjacent 4 s windows for each site. Moreover, Figure 10 shows the distribution of the event durations τ observed at each site.

The further step was to listen to the recordings of the detected sound events to recognize the corresponding source(s), according to the classification reported in Table 1. The labelling outcome selected by the listener was taken as reference

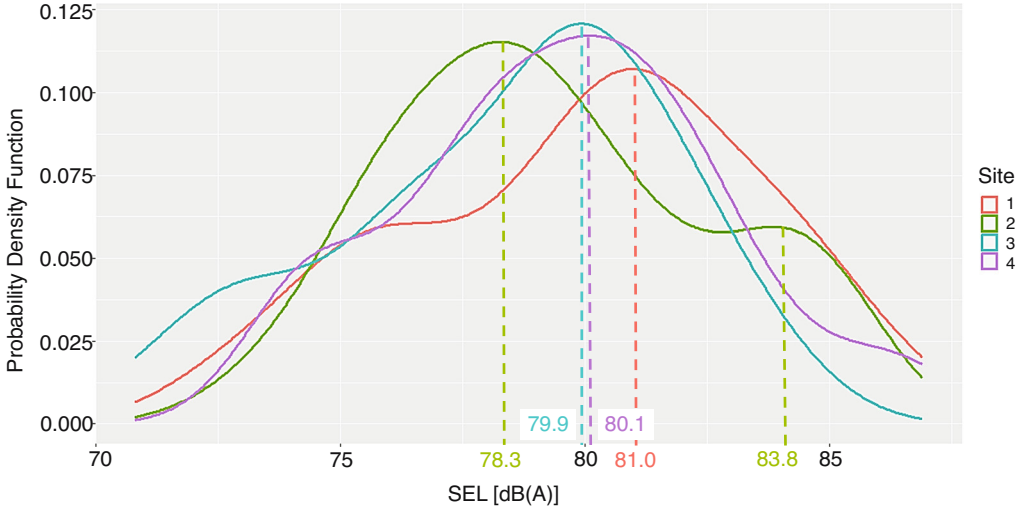

Figure 8: Density plot of SEL values of "notice-event" (NE) observed at the four sites. The legend numbers refer to the locations highlighted on Figure 2.

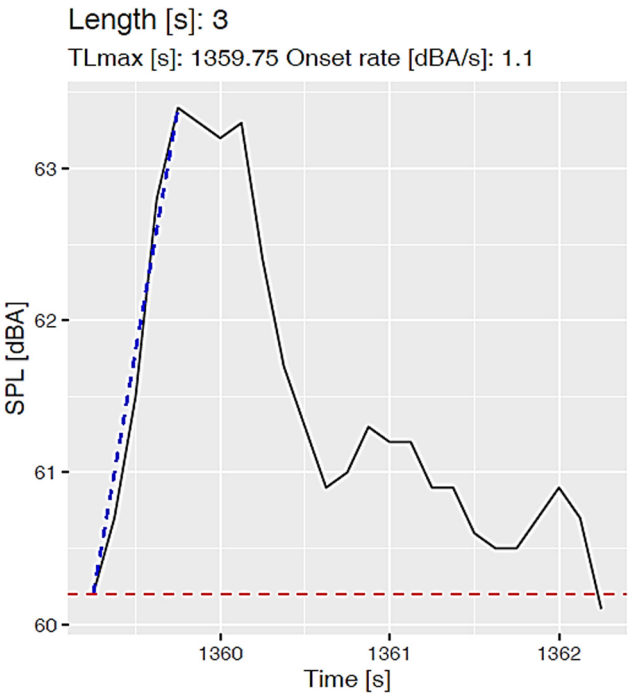

Figure 9: Example of the A-weighted SPL time history of a detected event. The horizontal red dotted line corresponds to the threshold for the event detection L_β and the inclined blue line connects the SPL at the start time with the maximum SPL of the event.

for evaluating the performance of the sound source classification by DL technique.

3.2 Sound source classification

Once the events were detected by the IR, the CNN classified them into different categories. For model evaluation, each event was manually listened to in order to verify whether the neural network correctly classified the category causing the event. This process was carried out by two experts that listened all the events at different moments. After the listening process, the results from both experts were checked and discussed until consensus was reached.

The neural network had 21 output neurons (one for each class described in Table 1), each outputting a probability (between 0 and 1) that the corresponding event was present, using a sigmoid activation function. Then, this output was binarized using a custom threshold for each class aiming to maximize the true positive events and minimize the false negative events for that category. The thresholds were selected from a validation set, which was a subsample of the training dataset (10% of data from BCNDataset and 10% of data from the Eixample dataset) that was not used for training nor testing and in which we did not apply any data augmentation techniques. Therefore, the process for obtaining the thresholds was:

1. First, we passed the validation data through the network, obtaining an output probability values for each class.
2. Then, given that more than one category might be active on each input fragment due to the polyphonic nature of the training data, we calculated the ROC curve for every class, which gave us the true positive rate (TPR) and false negative rate (FNR) in different thresholds.
3. Next, we calculated the geometric mean of the TPR and FNR for every threshold as follows:

$$\text{Geometric\_mean} = \sqrt{\text{TRP} \times (1 - \text{FPR})}. \quad (5)$$

4. Finally, for each class, we selected the threshold that enabled us to have the highest geometric mean.

The results of this procedure resulted in 21 different thresholds, most of them having values of around 0.5. The specific value of threshold per class is: *rtn* = 0.5, *peop* = 0.66, *brak* = 0.56, *bird* = 0.52, *motorc* = 0.42, *eng* = 0.5, *cdoor* = 0.59, *impls* = 0.47, *cmplx* = 0.70, *troll* = 0.56, *wind* = 0.5, *horn* = 0.14, *sire* = 0.54, *musi* = 0.48, *bike* = 0.43, *hdoor* = 0.60, *bell* = 0.59, *glass* = 0.55, *beep* = 0.5, *dog* = 0.54, *drill* = 0.48.

In the testing phase, once the events detected by the IR were passed through the network, the expert listeners would review each fragment and check whether it contained the detected noise sources.

Given the data were collected in a very noisy environment with common overlapping sounds, the manual evaluation categorized each output result of every neuron and for every event as follows:

- **TP:FG**: meaning True Positive: Foreground. This category represents the sources that have been detected both by the CNN and the expert listener and for the noise sources that are occurring in the foreground;
- **TP: BG:** meaning True Positive: Background. This category represents the sources that have been detected both by the CNN and the expert listener but that the event was occurring in the background or was masked by more salient event;
- **FP:** meaning False Positive. This category represents the sources that have been detected by the CNN but that the

Table 7: Percentage of detected notice-events NE spanning over adjacent 4 s windows for each site

Spot	1	2	3	4
N. of overall notice-events NE	40	54	42	62
Percentage of NE spanning over adjacent 4 s windows	55.0%	37.0%	40.5%	50.0%

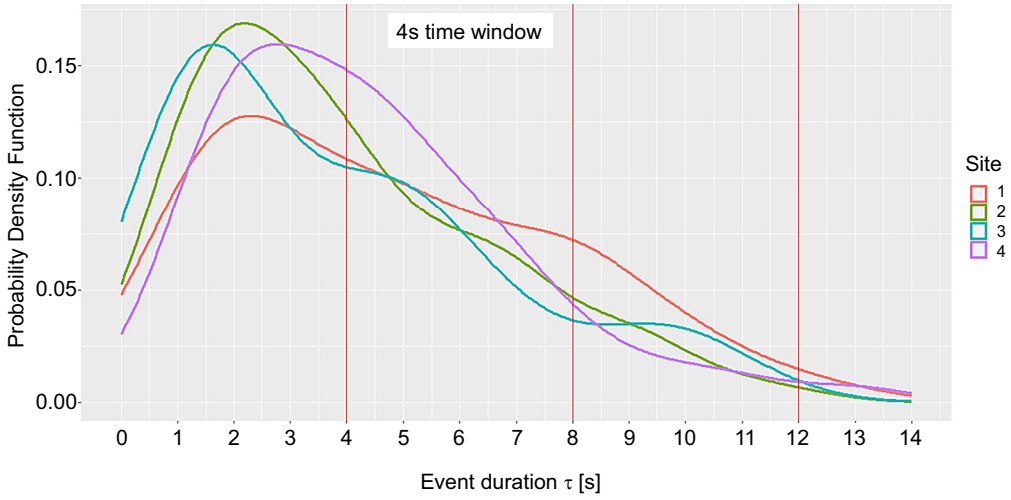

Figure 10: Density plot of event durations τ of "notice-event" (NE) observed at the four sites. The legend numbers refer to the locations highlighted on Figure 2.

expert listener could not hear them. Therefore, this category represents a mistake of the CNN;

- **FN:** meaning False Negative. This category reflects those events that were heard by the expert listener but that the CNN could not detect. Therefore, this category represents a mistake of the CNN too;
- **TN:** meaning True Negative. This category represents all the outputs that both the CNN and the manual listener did not detect. Therefore, this category represents a correct prediction.

As it can be observed, the standard metrics of TP, FP, FN and TN were used, but the True Positive events were divided into two categories: one for the most salient events and one for the events in the background. This way, it can be analyzed whether the positive events are only detected in the foreground or if there are events that are detected but are masked with other sounds.

The aggregated results from the four sensors, as shown in Table 8, provide insights into the performance of the CNN-based classification system, particularly in identifying foreground events, with a sensitivity equal to 0.829 and a specificity of 0.985.

Table 8: Classification results aggregated from the four sensors and all the categories obtained from the CNN

Category	Description	Amount of labels
TP:FG	Foreground: Most salient event	534
TP:BG	Background: Present but not most salient event	74
FP	Detected but not present	48
FN	Present but not detected	110
TN	Not present	6,384

The system demonstrated a commendable ability to detect the most salient sounds, with 534 events classified as TP: FG. These foreground events, which are the prominent parts of the audio, were accurately identified by the CNN. This indicates that the model is effective in capturing and classifying events that strongly influence the sonic environment.

However, the system also displayed limitations, primarily in the detection of background sources. Despite successfully classifying 74 TP: BG events, the CNN missed 110 events during manual listening analysis, mainly in the background. While this discrepancy exists, it is important to note that the system's ability to correctly identify background sources is still notable. The fact that 74 background events were accurately identified indicates that the model is capable of capturing and classifying sources that are less prominent in the audio.

Actually, considering the system as a whole, as the events are detected through the IR, it is coherent to think that the classifier should focus on detecting the most prominent sound, which is the one that has made the IR trigger the event.

Also, the number of false negatives exceeds the count of false positives, indicating a higher tendency for the system to fail in recognizing background sources compared to the identification of sources not present in the audio file.

For a more detailed analysis of the classification results, Table 9 shows the same categories (True Positives in the Foreground, Tue Positives in the Background, False Positives, False

Table 9: Results per class and per sensor

		rtn	peop	brak	bird	motorc	eng	cdoor	impls	cmplx	troll	wind	horn	sire	musi	bike	hdoor	busd	bell	glass	beep	dog	drill
S1	TP: FG	71	7	4	0	23	0	0	1	0	0	0	0	0	0	0	4	7	0	0	0	0	0
	TP: BG	0	9	1	3	3	1	0	2	2	0	0	0	0	0	4	0	0	0	0	0	0	0
	FP	0	0	0	0	0	2	0	0	0	0	0	2	0	0	1	0	2	0	0	0	0	0
	FN	0	5	0	0	4	0	0	0	1	2	4	0	1	0	2	2	3	1	0	0	0	2
	TN	0	50	66	68	41	68	71	68	68	69	67	69	70	71	64	65	59	70	71	71	71	69
S2	TP: FG	83	16	8	2	23	0	0	0	2	0	1	1	0	0	4	0	1	1	0	0	0	0
	TP: BG	0	6	11	1	3	0	0	2	0	0	0	0	0	0	3	2	0	0	0	0	0	0
	FP	0	3	0	0	4	0	2	1	1	2	0	2	0	0	0	0	0	0	0	0	0	0
	FN	0	2	2	0	0	0	0	0	0	3	5	0	0	0	8	0	1	0	0	0	0	1
	TN	0	56	62	80	53	83	81	80	80	78	77	80	83	83	68	81	81	82	83	83	83	82
S3	TP: FG	67	0	6	0	14	0	2	5	0	3	0	2	0	0	1	4	6	0	0	0	0	0
	TP: BG	0	4	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
	FP	0	0	0	0	1	0	0	5	1	6	0	2	0	0	0	0	0	0	0	0	0	0
	FN	0	10	1	0	1	0	0	0	1	3	1	0	0	0	5	1	6	0	0	0	0	2
	TN	0	53	59	67	51	67	65	57	64	54	66	63	67	67	60	62	55	67	67	67	67	65
S4	TP: FG	104	2	7	2	39	0	0	1	1	1	0	1	0	0	1	1	4	1	0	0	0	0
	TP: BG	0	2	4	0	1	0	0	0	1	1	0	0	0	0	4	0	0	0	0	0	0	0
	FP	0	0	0	0	6	0	1	1	1	0	0	2	0	0	0	0	0	0	0	0	0	0
	FN	0	15	0	1	0	0	0	0	1	0	2	1	1	0	5	1	1	0	0	0	0	2
	TN	0	85	93	101	58	104	103	102	100	102	102	100	103	104	94	102	99	103	104	104	104	102

Negatives and True Negatives) segregated by type of event and per class. This way, it is possible to analyze which of the categories are better classified.

As it can be observed, there are some sources that were not present in the testing data, and therefore it was not possible to evaluate if the algorithm is able to classify them when they are present. These are: *musi*, *glass*, *beep* and *dog*. However, it is still relevant to include them in the table to check that the system does not produce False Positive events in that category. The lack of those events in the testing hour occurred as the hour selected for testing the system contains real-world data, and during that time of the data gathering campaign, those sources were not present.

As it can be observed, all the fragments of events contained the *rtn* source, with some overlapping sources as well. The model was able to predict successfully all those events. Moreover, the system was capable of detecting also most of the *motorc* events, showing a good capability of detecting motorized sources. Same happens with the bus door category or even the house door source, where most of the events are correctly classified.

Also, in general, the system tends to have less False Positive detections than False Negatives. With the data used for evaluation, it can be seen that none of the *wind*, *drill* or *sire* sounds present in the data were actually detected by the CNN. When listening to the events, it can be concluded that it is because these few samples were actually masked by other sounds. Therefore, this reveals that the system struggles to identify all of the polyphonic sounds of a sample when they are highly masked.

As of consistency between sensors, it seems that the data gathered in the four different sensors have the same behavior in the CNN, showing that the system is robust against using data coming from different physical sound devices.

If the TP:FG and the TP:BG are added together, we can obtain the general True Positive values of the system, enabling the calculation of standard metrics such as the Precision, Recall or *F*1-score of the system per class. Actually, for the sake of better understanding the behavior of the classifier, Table 10 shows the classification metrics of those categories which contained samples (at least, 5) in the testing set and aggregating the classification data of the four sensors.

Table 10 reveals key insights into the classifier's performance: certain categories, such as *rtn*, *brak*, and *bird*, achieved exceptionally high *F*1-scores and the maximum Precision score, with *rtn* even attaining perfect scores across all metrics. Following this categories, *motorc* also demonstrates a strong overall performance, with an *F*1-score bigger than 90%. The performance of the classes *rtn*, *brak,* and *motorc* suggest that the classification model excels at classifying road traffic-related sounds, which is one of the categories with more interest in the END [15].

Conversely, categories such as *troll* and *horn* display lower performance, indicating more difficulties in accurately detecting these types of events. This reflects challenges in accurately detecting these types of events, possibly due to their more subtle or masked characteristics within the audio data used for testing and the fact that there are only few events of these categories in the testing set (only 14 *troll* events and 5 *horns*).

We must remember that the limited number of events in certain categories, such as *troll* and *horn*, can be attributed to several factors tied to the data collection and selection process: First, the testing data were derived from real-world, continuous recordings. Unlike synthetic or artificially balanced datasets, real-world data inherently reflects the natural distribution of sounds, which may not include all categories equally. Then, the testing data were chosen as a continuous time segment, representing a specific period of the day in which the IR criteria can be applied to detect the acoustic events, which is the target of this work.

In summary, while there is room for refinement, particularly in the detection of background sources, the CNN-based classification system demonstrates promising capabilities in identifying sources in the acoustic environment, achieving an overall aggregated metric surpassing the 84% of *F*1-score (both micro and macro averaged) among those categories that were more present in the testing data. Addressing these limitations – probably by using more training data in those categories that are less present in the dataset such as *musi*, *glass*, *beep* or *dog* – will lead to a more robust and reliable tool for assessing and managing environmental noise.

Table 10: Precision, Recall and *F*1-score per category and aggregate metrics

Category	Precision (%)	Recall (%)	*F*1-score (%)
rtn	100	100	100
peop	93.8	58.9	72.4
brak	100	93.3	96.5
bird	100	88.8	94.1
motorc	90.6	95.5	93
impls	61.1	100	75.9
troll	42.9	42.9	42.8
horn	33.3	80	47.1
bike	94.7	47.4	63.2
hdoor	100	73.3	84.6
busd	90	62.1	73.5
Macro-average	90.6	84.2	84.3
Micro-average	87.5	87.5	87.5

4 Conclusions

The proposed methodology in this study combines the IR criterium for acoustic event detection with DL techniques for the classification of sound sources associated with these events. The rationale behind this combination is to provide a tool that can automatically detect and classify polyphonic acoustic events significantly affecting soundscape quality.

Using a dataset collected in the city center of Barcelona (and particularly a busy street with both traffic and leisure sounds), the results show that the IR criterium has been able to successfully detect events from a wide variety of source categories. This highlights the effectiveness of the IR metric in capturing the temporal irregularity characteristics of perceived noise notice-events, contributing to the accurate identification of various sound sources.

Results also support a significant capability of the CNN-based classification system to handle the identification of acoustic events with an aggregated macro and micro *F*1-score surpassing 84%, and it is especially successful classifying foreground events. Concretely, the system successfully classified 534 True Positive Foreground sound sources, representing the most salient sounds in the audio recordings. However, the system showed more limitations in detecting background sources. While it correctly classified 74 sources in the background, it missed 110 sources that were heard in the background during a manual expert listening analysis for source identification. This discrepancy suggests a higher tendency for the system to fail to recognize background sources compared to sources not present in the audio file. Notwithstanding this inaccuracy, the described methodology represents a step forward in automating acoustic event detection and classification in urban sound environment, providing information on sound events in terms of their occurrences and sources, which aligns to the research objectives of the (END, Directive 2002/49/EC) [12], targeting to create tools that enable effective noise mitigation actions.

Addressing these limitations will lead to a more robust and reliable tool for assessing and managing environmental noise. Future work should focus on the analysis of the specific characteristics of the missed background events to identify patterns or features that the CNN might have overlooked. Additionally, exploring methods to enhance the model's sensitivity to both foreground and background events, such as incorporating more diverse training data or refining the neural network architecture, could lead to improvements in performance.

Furthermore, it would be beneficial to investigate the relationship between the characteristics of the detected events and their perceived impact on human listeners. Understanding how different types of events contribute to the overall soundscape experience can provide useful information for more efficient noise mitigation strategies, tailored to specific environments.

Acknowledgements: The authors would also like to thank the Departament de Recerca i Universitats (Generalitat de Catalunya) under Grant Ref. 2021-SGR-01396 for the funding of HER (Human-Environment Research) research group.

Funding information: The authors thank the support of the Catalan Government (Departament de Recerca i Universitats) for the grant 2021 SGR 01396 given to the HER group.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results and approved the final version of the manuscript. EVV contributed in data curation, formal analysis, investigation, methodology, software, validation, visualization, and writing of the original draft. GB contributed in data curation, formal analysis, investigation, methodology, conceptualization, and writing of the original draft. RMAP contributed in conceptualization, supervision, validation, and writing (reviewing and editing) of the work.

Conflict of interest: Authors state no conflict of interest.

References

[1] Brown AL, Banerjee D, Tomerini D. Noise events in road traffic and sleep disturbance studies. In Proceedings of the Internoise 2012. New York, NY, USA: 19–22 August 2012.

[2] Brown AL, De Coensel B. A study of the performance of a generalized exceedance algorithm for detecting noise events caused by road traffic. Appl Acoust. 2018;138:101–14. doi: 10.1016/j.apacoust.2018.03.031.

[3] De Coensel B, Brown AL. Event-based Indicators for road traffic noise exposure assessment. In Proceedings of the Euronoise 2018, Crete, Greece; 27–31 May 2018. p. 485–90.

[4] Wunderli JM, Pieren R, Habermacher M, Vienneau D, Cajochen C, Probst-Hensch N, et al. Intermittency ratio: A metric reflecting short-term temporal variations of transportation noise exposure. J Expo Sci Env Epidemiol. 2016;26:575–85.

[5] De Coensel B, Botteldooren D, De Muer T, Berglund B, Nilsson ME, Lercher P. A model for the perception of environmental sound based on notice-events. J Acoust Soc Am. 2009;126:656–65. doi: 10.1121/1.3158601.

[6] Orga F, Alías F, Alsina-Pagès RM. On the impact of anomalous noise events on road traffic noise mapping in urban and suburban environments. Int J Env Res Public Health. 2018;15:13. doi: 10.3390/ijerph15010013.

[7] Alsina-Pagès RM, Benocci R, Brambilla G, Zambon G. Methods for noise event detection and assessment of the sonic environment by the harmonica index. Appl Sci. 2021;11:8031. doi: 10.3390/app11178031.
[8] Alías F, Orga F, Alsina-Pagès RM, Socoró JC. Aggregate impact of anomalous noise events on the WASN-based computation of road traffic noise levels in urban and suburban environments. Sensors. 2020;20(3):609.
[9] Alsina-Pagès RM, Alías F, Socoró JC, Orga F, Benocci R, Zambon G. Anomalous events removal for automated traffic noise maps generation. Appl Acoust. 2019;151:183–92.
[10] Socoró JC, Ribera G, Sevillano X, Alías F. Development of an anomalous noise event detection algorithm for dynamic road traffic noise mapping. In Proceedings of ICSV 2015. Florence, Italy: 12–16 July 2015.
[11] Bellucci P, Peruzzi L, Zambon G. LIFE DYNAMAP: Making dynamic noise maps a reality. In Proceedings of Euronoise 2018, Crete, Greece; 27–31 May 2018. p. 1181–8.
[12] EU. Directive 2002/49/EC of the European Parliament and the Council of 25 June 2002 relating to the assessment and management of environmental noise. J Eur Commun. 2002;L189:12–25.
[13] World Health Organization. Environmental noise guidelines for the European region; 2018.
[14] Schmidhuber J. Deep learning in neural networks: an overview. Technical Report IDSIA-03-14/arXiv:1404.7828v4. 2014.
[15] Vidaña-Vila E, Navarro J, Stowell D, Alsina-Pagès RM. Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors. Sensors. 2021;21:7470. doi: 10.3390/s21227470.
[16] XYH-5 Capsule. https://zoomcorp.com/en/us/accessories/mic-capsules-foot-switches-and-pedals/XYH-5/. Visited on: 19/09/2024
[17] Vidaña-Vila E, Duboc L, Alsina-Pagès RM, Polls F, Vargas H. BCNDataset: Description and analysis of an annotated night urban leisure sound dataset. Sustainability. 2020;12:8140. doi: 10.3390/su12198140.
[18] Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM international conference on Multimedia; 2014. p. 1041–4.
[19] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2021.
[20] Brambilla G, Confalonieri C, Benocci R. Application of the Intermittency Ratio metric for the classification of urban sites based on road traffic noise events. Sensors. 2019;19:5136. doi: 10.3390/s19235136.
[21] Nordtest Method NT ACOU 112. Acoustics – Prominence of impulsive sounds and for adjustment of LAeq. Taastrup, Denmark: Nordtest; 2002.
[22] McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science Conference. Vol. 8, 2015.
[23] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS2019), Vancouver, Canada; 2019.
[24] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017.
[25] Chen J, Zhang F, Li Y. A sound event recognition method of crop shear scrap falling state based on log-mel spectrogram and MobileNetV2. In 2023 42nd Chinese Control Conference (CCC). IEEE; 2023. p. 6946–51.
[26] Vidaña-Vila E, Navarro J, Borda-Fortuny C, Stowell D, Alsina-Pagès RM. Low-cost distributed acoustic sensor network for real time urban sound monitoring. Electronics. 2020;9:2119. doi: 10.3390/electronics9122119.