Digit. Signal Process. 147 (2024) 104431

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp

CARgram: CNN-based accident recognition from road sounds through intensity-projected spectrogram analysis

Alessandro Sebastian Podda [a,*], Riccardo Balia [a], Livio Pompianu [a], Salvatore Carta [a,b], Gianni Fenu [a], Roberto Saia [a]

[a] Dept. of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, Cagliari, 09124, CA, Italy
[b] VisioScientiae S.r.l., Via Francesco Ciusa 46, Cagliari, 09131, CA, Italy

ARTICLE INFO

Keywords:
Convolutional neural networks
Deep learning
Audio analysis
Traffic surveillance
Artificial Intelligence

ABSTRACT

Road surveillance systems play an important role in traffic monitoring and detecting hazardous events. In recent years, several artificial intelligence-based approaches have been proposed for this purpose, typically based on the analysis of the acquired video streams. However, occlusions, poor lighting conditions, and heterogeneity of the events may often reduce their effectiveness and reliability. To overcome the limitations mentioned, scientific and industrial research has therefore focused on integrating such solutions with audio recognition methods. By automatically identifying anomalous traffic sounds, e.g., car crashes and skids, they help reduce false positives and missed alarms. Following this trend, in this work, we propose an innovative pipeline for the analysis of intensity-projected audio spectrograms from streams of traffic sounds, which exploits both (i) a visual approach based on a custom, special-purpose Convolutional Neural Network for the identification of anomalous events on the sound signal; and, (ii) a novel multi-representational encoding of the input, which proved to significantly improve the recognition accuracy of the neural models. The validation results of the proposed pipeline on the public MIVIA dataset, with a 0.96% of false positive rate, showed to be the best performance against the state-of-the-art competitors. Notably, following such findings, a prototype implementation has been deployed on a real-world video surveillance infrastructure.

1. Introduction

Recent *World Health Organization* (WHO) studies report that an average of 1.3 million people die each year due to road accidents, with between 20 and 50 million injured. These findings are not encouraging, so several countries are increasing efforts to manage this problem [1]. An actionable path to counter the phenomenon is the enhancement of instant notification systems so that authorities can be quickly alerted for immediate intervention. In this way, it is possible to increase the chances of survival of the injured and prevent further side events [2,3]. Solutions that can fulfill this need can essentially be classified into two types: (i) *active* ones, i.e., those that require voluntary action, or (ii) *passive* ones, i.e., those that use sensors inside or outside the vehicle can automatically identify the accident event [4,5]. The first category includes devices for speed dialing to emergency numbers, such as the *SOS buttons* installed in modern cars [6,7]. The second, on the other hand, assumes the adoption of *built-in*, e.g. crash sensors or *off-vehicle* monitoring systems, like traffic video surveillance systems [8]. Video surveillance systems are proving to be particularly effective for this purpose since, in addition to being widely used in urban and non-urban settings, they also allow the safeguarding of drivers of vehicles, either old or not equipped with expensive security facilities. However, their use is often limited to a *traditional* unsupervised scenario recording because of the high costs of personnel and resources needed for continuous monitoring. For these reasons, novel Artificial Intelligence and Computer Vision techniques represent a turning point for this sector, as they allow the automation of the scene understanding and alerting activities [9–12]. Since most of these approaches are image-based, they tend to suffer significantly from environmental factors such as poor lighting, weather conditions, and occlusions, to name a few. Accordingly, scientific research has recently focused on integrating such solutions with models capable of analyzing context information, especially the audio signal often associated with the captured video stream.

* Corresponding author.
E-mail addresses: sebastianpodda@unica.it (A.S. Podda), riccardo.balia@unica.it (R. Balia), livio.pompianu@unica.it (L. Pompianu), salvatore@unica.it (S. Carta), fenu@unica.it (G. Fenu), roberto.saia@unica.it (R. Saia).

https://doi.org/10.1016/j.dsp.2024.104431

Available online 13 February 2024

Following this trend, in this study, we propose an original method, based on a visual approach that uses *Convolutional Neural Networks* (CNNs), to detect road hazard events by analyzing the spectrogram of the audio signal collected by traffic surveillance systems. Note that although several contributions in the literature [13–16] adopt similar approaches, the methodology proposed in this paper differs significantly from them. The foundational element of our proposal lies in the fact that a series of transformations, known as *intensity projections*, a popular technique used in medical field [17], are applied to the spectrogram to generate an enhanced encoding of the input - that we refer to as *multi-representational* - capable of both taking into account the original representation of the data and highlighting areas of harmonic and rhythmic breakup. In addition, we also designed a custom CNN architecture to exploit the potential of such enhanced encoding.

Summarizing, the main contributions of this work are:

1. we first experiment with different *Deep Learning*-based approaches to detect and classify road hazard events, especially accidents, from audio signals, determining an initial baseline on the type of architecture that performs best;
2. we then propose a novel multi-representational CNN architecture that exploits, in a synergistic and efficient mode, the visual information obtained from the analysis of the audio spectrogram, as well as its intensity projections on the frequency and time axes;
3. we validate the effectiveness of our method by (i) demonstrating its ability to significantly improve performance with respect to using a standard CNN architecture on the basic spectrogram and (ii) comparing its results against the state-of-the-art competitors on the public *MIVIA Audio Road Events* dataset, showing the superiority of the proposed architecture to detect road hazardous sounds in urban scenarios.

The remainder of the manuscript is organized as follows. Section 2 presents an overview of the related work, while Section 3 describes the proposed method in detail. Section 4 illustrates the experimental setup, with a focus on the pre-processing stage. Then, Section 5 shows the validation results and the comparison with the state-of-the-art. Finally, Section 6 concludes the paper and outlines the future research directions.

2. Related work

Identifying accidents and, more generally, harmful events for road users is a hot research area for the artificial intelligence scientific community. In particular, the most common literature approaches exploit the analysis of video or, less frequently, audio-type signals.

The number of studies oriented toward event recognition from video is growing due to the widespread deployment of video surveillance infrastructure in modern urban settings. Among them, Thomas et al. [18] propose an optimized framework for perceptual video summarization and categorizing different stages of accidents and types of collisions. On the other hand, Arceda et al. [19] present a three-stage framework in which they first recognize vehicles with a *You Only Look Once* (YOLO) system and then use a *Violent Flow* descriptor along with a *Support Vector Machine* (SVM) to detect their collisions. YOLO-based frameworks for car detection have also been proposed by Wang et al. [20], with the addition of a *Retinex* algorithm to improve image quality in challenging low-light and bad weather conditions. Similarly, the authors in [21] illustrate a solution based on *Convolutional* and *Recurrent* neural networks (CNNs and RNNs) to analyze visual features first and then explore temporal ones. Saravanarajan et al. [21] focuses on single-car crash detection proposing an ensemble of three networks that are involved in feature extraction, identification of regions of interest with a Region proposal network (RPN), and a CNN8L network that predicts the correct bounding box. Overall, the approaches above generally require multiple stages of analysis, and environmental conditions can affect the efficiency of these systems. However, when it comes to classifying audio signals, there are significantly fewer research papers. Moreover, the classification methods often rely on different visual representations of the audio [22,23]. Rovetta et al. [24] propose an SVM-based method to identify outliers in sound streams, such as car crashes, and a *deep neural network* to classify the detected events. Sammarco et al. [25] present *Crashzam*, an audio-based detector deployed as a smartphone application that detects collisions from inside the vehicle and exploits features such as accelerometer and GPS data. On the same trail, the authors in [26] define a framework based on a *Deep Autoencoder* and a *Bidirectional* Long Short-Term Memory (B-LSTM) for hazard events classification. Arslan et al. [27] developed a DNN system for detecting hazardous sounds like screams and car crashes from Mel-frequency cepstrum features.

As regards the studies more focused on acoustic traffic event detection, the work in [28] faces the detection of traffic events in long tunnels through the analysis of audio signals. The authors utilize real-life data collected from a tunnel environment, categorizing audio samples into various traffic events such as tire friction sound and vehicle percussion sound. To enhance efficiency, the paper proposes fast binary spectral features for rapid classification and adopts a deep neural network approach to model acoustic characteristics. Comparative evaluations against state-of-the-art algorithms, including LSTM neural networks and Gaussian mixture models with Mel frequency cepstral coefficients, demonstrate the superiority of the proposed spectral features in terms of accuracy and efficiency for detecting traffic-related audio events. In the investigation conducted in [29], the focus is on incident response in Australian road tunnels. The paper employs Bayesian Networks, a machine-learning technique, to analyze cause-effect relationships between incident variables. The authors use structure learning algorithms and scoring functions to build a network structure validated against multiple indicators. Furthermore, parameter learning is applied to estimate the probability of consequences. The diagnostic and predictive reasoning of Bayesian Networks are then leveraged for what-if scenarios, identifying variables that significantly impact the severity of road tunnel incidents. The findings offer valuable insights for tunnel operators to develop mitigation strategies and reduce the number of injuries resulting from incidents. The work [30] presents an innovative approach to car accident detection and reconstruction named Crashzam. Unlike traditional methods relying on accelerometer sensor analysis, this paper utilizes sound analysis of car impacts reverberating inside the car cabin. The authors introduce an original dataset containing crash sounds, outlining the system design, model, and classification based on features extracted from the time and frequency domain of the audio signal and its spectrogram image. Results indicate the model's ability to accurately identify crash sounds amidst in-car cabin noise, providing a promising avenue for smart connected vehicles to enhance accident detection and reconstruction. In the system for road accident detection proposed in [31], the emphasis is on real-time identification of accidents based on machine learning tools. The system gathers information from adjacent vehicles, utilizing machine learning techniques to differentiate abnormal traffic behavior from normal patterns. By examining traffic behavior, the system aims to identify potential road accidents promptly. This work adds to the growing body of research leveraging machine learning for proactive accident detection, contributing to the overall goal of enhancing road safety.

The literature also features a variety of studies that exploit the MIVIA dataset as a test bed. For instance, in order to analyze multiple representations of the audio data, Mnasri et al. [32] explore the use of a Fully-connected Neural Network (FCNN) and an LSTM, exploiting an autoencoder to initialize the network weights. Analogously, Strisciuglio et al. [33] propose a new feature extractor called *Combination of Peaks of Energy* (COPE) combined with an SVM classifier. Other works employ dedicated CNN architectures: Foggia et al. [34] exploit *MobileNetv2*, a network designed to efficiently run on embedded devices, while Greco et al. [35] leverage on a 21-level CNN, *AReN*, for recognizing unusual

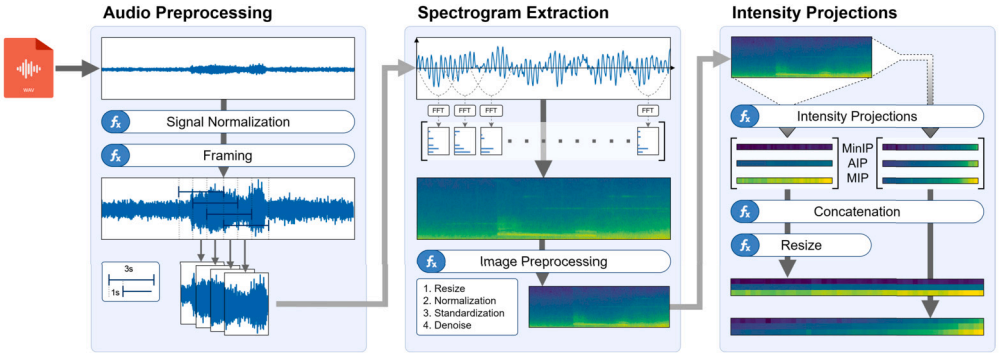


Fig. 1. Summary of the pre-processing, spectrogram extraction, and projection steps.

sounds in audio tracks represented as *gammatonegrams*, i.e. gammatone-filtered spectrograms.

3. Methods

The proposed method aims to detect and classify road hazard events from audio signals through visual analysis of the corresponding spectrogram using a *Convolutional Neural Network* (CNN). Precisely, the novelty of the proposed method consists of adopting a *multi-representational* approach which exploits information from its intensity projections on the time and frequency axes in addition to considering the spectrogram of the observed signal, in a synergistic and combined way. To accomplish this task, we introduce a custom network architecture hereafter referred to as *Multi-Input CNN*. Along with this solution, the pre-processing step also plays a relevant role, as better discussed in the following Section 3.1.

3.1. Pre-processing

The pre-processing of the data involves both the source audio signals and their graphical representation. It mainly aims to prepare the data for its use by the proposed Deep Learning method by i) preserving some valuable features such as signal strength and ii) reducing the size and range of values to speed up the processing.

More in detail, the first step concerns the source audio files: for each of them, we apply an audio normalization step, scaling the audio signal to bring the highest amplitude peak (in absolute value) to the maximum possible. Next, in order to feed the neural network with fixed-size inputs, in line with the authors of the dataset we sequentially extracted 3-second-long audio frames by means of a sliding window, with a 1-second shift between, so that each frame shares two-thirds of the information with the previous one, in order to prevent the event from being cut off at significant points and to ease detection of events that may occur at the extremes of the frames.

The resulting frames are then used to generate the visual representation of the audio, i.e., its *spectrogram*. It represents the signal spectrum as a function of frequency and time, obtained by applying the *Short-Time Fourier Transform* (STFT). In formulae (polar form):

$$X[k,m] = \sum_{n=0}^{N-1} w[n] \cdot x[n+mH] \cdot e^{-\frac{2i\pi}{N}kn} \tag{1}$$

where k is the frequency index, m the frame index, N the frame size, H the hop size, $x[n+mH]$ the mth frame of the source signal and $w[n]$ the window function.

In short, we first divided the signal into segments; then, we applied the Fast Fourier Transform to them. As a result of this conversion, each audio frame goes from being a one-dimensional vector to a 2D matrix in order to apply image processing techniques. On the practical side, the extraction of spectrograms from the dataset sources involved the use of the `matplotlib` library with default values including a *Hanning window* with NFFT of 256 samples and overlap of 128 samples, applied on 3-second frames sampled at 32 kHz (96,000 samples per frame). Resulting spectrograms were resized to a dimension of 50x300 pixels to speed up the following pre-processing steps and reduce the number of parameters required by the neural network. In the succeeding steps, we normalized the pixel intensity values in the range [0,1] and standardized. We applied both operations in feature-wise mode and computed the parameters on the set of spectrograms that compose the dataset rather than on the single ones, as would happen with a sample-wise approach. Indeed, preliminary empirical tests reported a significant increase in false positives when adopting min-max sample-wise normalization.

As a final pre-processing step, we applied noise reduction to the spectrograms through a Gaussian filter with a `3x3` kernel size.

3.2. Intensity projections

The spectrogram images obtained as a result of the pre-processing operations described in Section 3.1 constitute one of the inputs used to feed our neural model. However, the peculiarity of the proposed method is to exploit additional representations of the data to make the Deep Learning algorithm better able to discern between *normal* background noise and dangerous events. Hence, from the basic spectrogram, the proposed approach extracts two additional representations obtained by performing different *intensity projections* on it [36]. This approach represents a scientific visualization method commonly adopted when manipulating 3D data aimed to project *voxel* values to a 2D plane image, especially in the medical domain. In the context of this work, the size of the spectrograms is reduced from a 2D image to one-dimensional vectors for each applied projection. Specifically, we applied the projections along the vertical (frequency) and horizontal (time) axes in three different modes, namely: i) the *Minimum Intensity Projection* (MinIP); ii) the *Maximum Intensity Projection* (MIP); and iii) the *Average Intensity Projection* (AIP), denoted as it follows:

$$MinIP_i = min_j(I_{i,j}) \tag{2}$$

$$MIP_i = max_j(I_{i,j}) \tag{3}$$

$$AIP_i = \frac{1}{N}\sum_{j=0}^{N} I_{i,j} \tag{4}$$

where $I_{i,j}$ is the intensity of the input image at the pixel position (i,j); min_j and max_j represent the minimum and maximum values of pixels in row i, respectively; N is the total number of columns (i.e., the *x-axis* size) of the input image. As a result, we obtain three one-dimensional vectors, one for each projection, stacked on a new 2D matrix, one for each axis.

Note that the matrix obtained from the projection of the second axis is resized to reduce dimensionality and computation costs when applied to the convolutional layers of the proposed network. Fig. 1 comprehensively summarizes the pre-processing above and the projection steps.

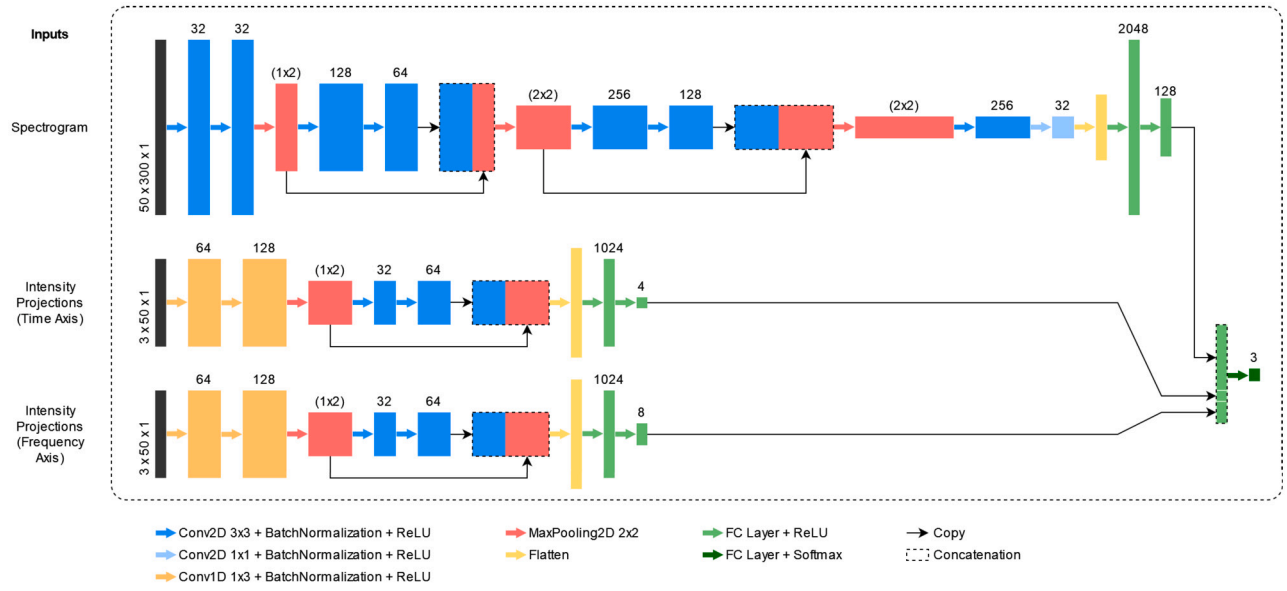


Fig. 2. Diagram of the proposed special-purpose Multi-Input CNN architecture.

3.3. Multi-input CNN architecture

The core of our proposed method consists of a custom *Multi-Input Convolutional Neural Network* focused on processing and classifying graphical representations of an audio signal, i.e. a native spectrogram alongside two derived inputs obtained through its intensity projections (IPs). The architecture, illustrated in Fig. 2, consists of three pipelines that independently process each received input. Each path's intermediate outputs are concatenated to obtain an ensemble step that operates in the last fully connected layer and returns the final prediction in the form of probability values associated with the three possible classes, namely `background_noise` (BN), `car_crash` (CC), and `tire_skiddings` (TS).

The first pipeline, involved in the full spectrogram processing, includes eight 2D convolutional layers with a `3x3` receptive field (except for the last one, detailed below), characterized by a path of feature map expansion and contraction to reduce computational cost growth. The main pattern of such a pipeline consists of three blocks, each composed of two convolutional steps followed by a *Max Pooling* layer. The output of each pooling layer is both i) passed to the first convolutional layer of the subsequent block and ii) concatenated with the output of the second convolutional layer of that block, implementing a *skip connection* technique to limit the vanishing/exploding gradient problems [37]. As mentioned before, the last layer provides a `1x1` kernel to project the feature map into a smaller space and reduce the number of parameters required in the two final fully connected layers, composed of 2048 and 128 nodes, respectively.

On the other hand, the two pipelines that handle the time and frequency IPs share a similar structure. Both include a total of 4 convolutional layers. Two of them are of type 1D and independently process the rows of the input. Then, a pooling layer halves the size of the feature map on the width axis, producing an intermediate output that is both processed by the two subsequent 2D convolutional layers and concatenated to their output. Such a concatenation is then flattened and processed by two consecutive fully connected layers, the first providing 1024 nodes, followed by a smaller one with 4 or 8 nodes, depending on whether the processing pipeline involves the time or frequency IP, respectively.

The outputs of the above pipelines are exploited by the fully connected layer mentioned before, providing three output nodes (corresponding to the number of classes) and employing a *Softmax* activation function to guarantee that returned values are in the range [0,1] and their sum equals 1. We point out that a *Batch Normalization* layer follows each convolutional layer in our Multi-Input CNN and exploits a *Rectified Linear Unit* (ReLU) activation function.

4. Experimental setup

We implemented the proposed methods in *Python 3.8*, by employing the following Audio Processing and Machine/Deep Learning libraries: `pyaudio` 0.2.11, `pydub` 0.25.1, `librosa` 0.9.1, `matplotlib` 3.5.1, `numpy` 1.21.5, `keras` 2.7.0, and `tensorflow` 2.7.0. We ran the experiments on a desktop computer with a 4.10 GHz CPU, 32GB RAM, and an Nvidia GeForce GTX 1060 Max-Q GPU equipped with 6GB dedicated DDR5 RAM and 1280 CUDA cores.

4.1. Dataset

We conducted the experiments on a publicly available dataset known as *MIVIA Road Audio Events* ([38,39]). The basic idea behind this dataset stems from the fact that, although car accidents are rare events, the sounds associated with them can be distinctive and often carry specific characteristics. Breaking and deforming sounds are common factors, often accompanied by the shattering of glass in accidents involving windows or windshields. The deformation of metal and other materials generates additional sounds, including creaking, bending, and crumpling noises. Moreover, before the impact, there may be the unmistakable screeching of tires as vehicles attempt to stop or change direction. Also, human sounds, such as yells or screams, may punctuate the auditory landscape in some cases, reflecting the emotional and alarming nature of the situation. In particular, in traffic surveillance scenarios, the events of interest may occur at varying distances from the microphone, leading to different signal-to-noise ratio levels. Additionally, these events often blend with complex backgrounds comprising various sounds typical of indoor and outdoor environments (e.g., household appliances, crowd cheering, conversations, traffic noise). Hence, the MIVIA dataset is structured to present each audio event across six signal-to-noise ratio levels (5 dB, 10 dB, 15 dB, 20 dB, 25 dB, and 30 dB), layered with diverse combinations of environmental sounds to simulate different ambient settings.

More in deep, such a dataset distinguishes between car crashes and tires skidding and consists of 57 audio files 60 seconds long and sampled at 32 kHz, recorded with an *Axis P8221Audio* module and an

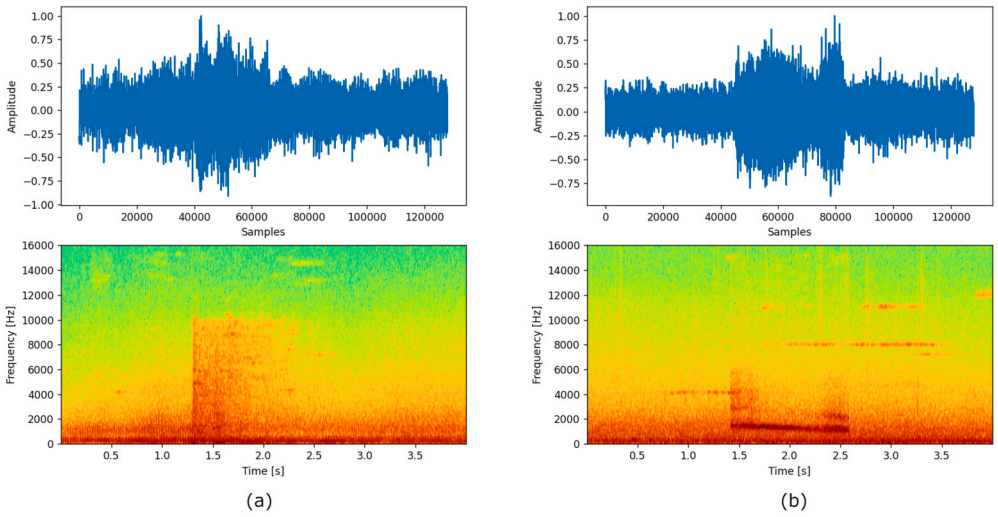


Fig. 3. Examples of spectrograms extracted from dataset's positive events: (a) car crash, (b) tire skidding.

Axis T83 omnidirectional microphone for audio surveillance applications. For the purpose of this work, we clarify that the sampling rate was not adjusted, to let us compare the results with literature competitors; however, downsampling might marginally speed up the real-time preprocessing steps, without a significant impact on performance, although main computations are applied to the spectrogram. These files contain 400 events, of which 200 are labelled with class `car_crash` and 200 with class `tire_skidding`. Fig. 3 shows examples of spectrograms for car crashes (on the left) and tires skidding (on the right). The files are pre-organized into four folders with 100 events each in view of the cross-validation experiments, which we adopted as suggested by the dataset authors. With this approach, we used three folds for training the network in each iteration, and we divided the fourth one into two equal parts, respectively, used as validation and test sets.

4.2. Metrics

In order to quantitatively validate the performance of the proposed method and adequately compare it with the existing works, we adopted the same experimental protocol proposed by the dataset authors [40], which involves the following metrics:

- the *True Positive Rate* (TPR), i.e. the ratio of correctly identified positive events (TP, *true positives*) over all the positive events (P);

$$TPR = TP/P \tag{5}$$

- the *False Positive Rate* (FPR), defined as the ratio of events classified as positive when only background noise frames are present;

$$FPR = FP/P \tag{6}$$

- the *Miss Rate* (MR), computed as the number of undetected events (FN, *false negatives*) over the total number of positive events (P);

$$MR = FN/P \tag{7}$$

- the *Error Rate* (ER), i.e. the number of misclassified events over the total number of positive events.

$$ER = (FN_{TS} + FN_{CC})/P \tag{8}$$

where FN_{CC} are the events classified as `car_crash` when the correct label was `tire_skidding` and FN_{TS} are the events classified as `tire_skidding` when the correct label was `car_crash`.

Additionally, we also provide the *Receiver Operating Characteristic* (ROC) curves and the *Area Under the Curve* (AUC) measures to compare the performance of the proposed multi-representational model against the single-input CNN implementations.

4.3. Hyperparameters

For our experiments, we adopt the following hyperparameters: the *SGD* optimizer with a learning rate set to 0.01 and momentum set to 0.89, the *Mean Logarithmic Squared Error* (MLSE) loss function, a batch size of 16 samples and 25 training epochs. In addition, to avoid overfitting, we exploit an *early stopping* strategy to interrupt the training process if the validation loss does not improve within 15 epochs of patience and a *checkpoint callback* to retrieve the model with the lowest validation loss cost at the end of the training process.

5. Results and discussion

We experimentally validated the proposed system on a dataset designed explicitly for road surveillance applications known as MIVIA Road Audio Events, described in Section 4.1. We obtained experimental results by applying a *4-fold cross-validation* strategy provided by the dataset subdivision and sharing the same training hyperparameters among the models. We assessed the performance of the proposed multi-representational approach in two regards: its contribution compared to single-input methods and its competitiveness against relevant benchmarks from existing literature.

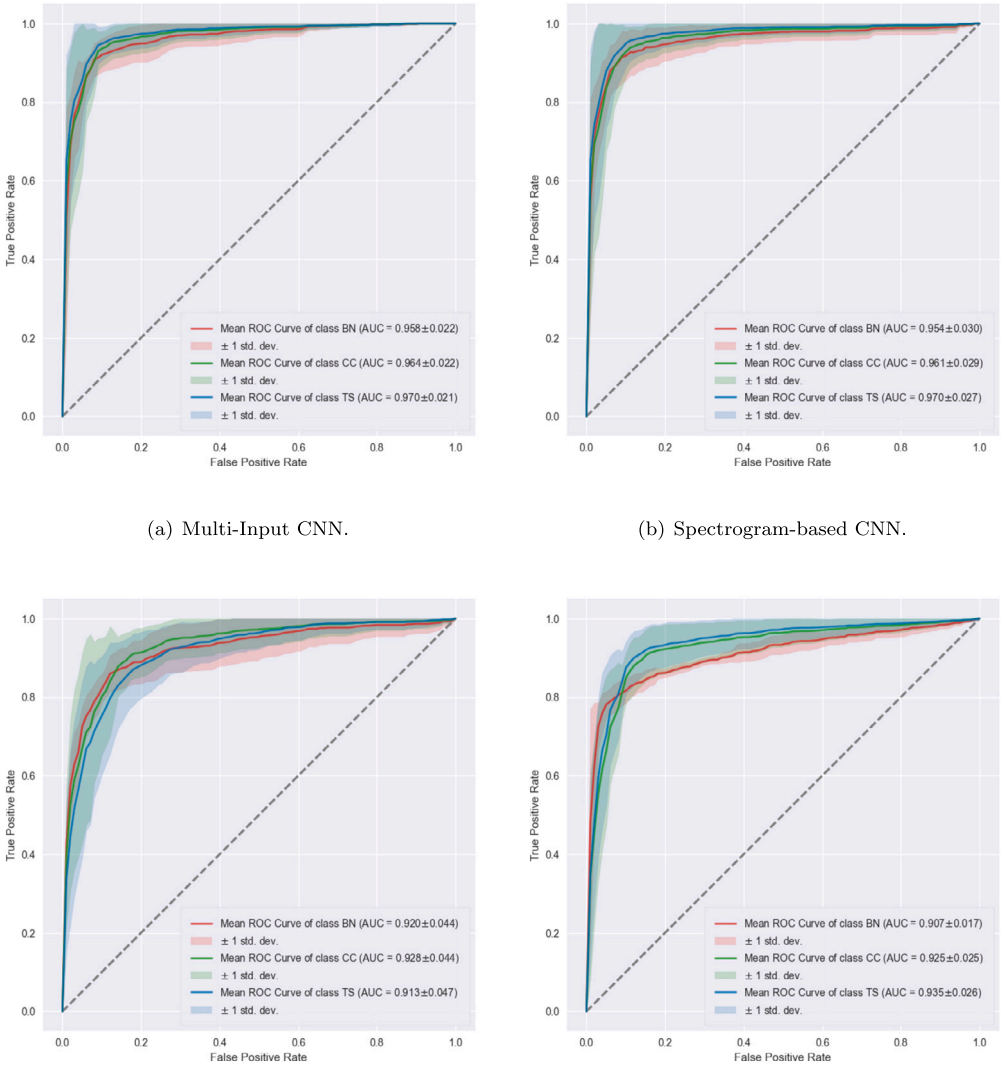
5.1. Multi-input pipeline vs. single-input approaches

Table 1 illustrates the results of comparing the proposed multi-representative approach with single-input networks. For clarity, note that the single-input networks share the same structure as the subnets embedded within the multi-input network described in Section 3.3, and each of them makes use of a classification head of 1024, 32, and 3 nodes. The results show a significant drop in performance for the network associated with intensity projection on the time axis. In particular, the approach did not detect all events correctly, leading to a high rate of missed events with 7.32% MR and multiple events misclassified, despite being recognized as events of interest with 5.85% ER. We expected this behavior as the type of event is more easily distinguished when looking at the frequencies that make up the event, even for a human.

This is done in the classification of events using the intensity projections of the frequency axis, by which the model is able to recognize

Table 1
Results obtained by considering different architectures and input data types on the MIVIA Road Audio Events dataset (†: higher is better; ‡: lower is better).

Method	Data Type	ACC†	TPR†	FPR‡	MR‡	ER‡
CNN	Time IP	77.97 ± 8.01	86.81 ± 8.01	3.92 ± 2.55	7.32 ± 6.27	5.86 ± 2.56
CNN	Frequency IP	83.52 ± 2.24	97.07 ± 3.51	4.37 ± 2.16	1.95 ± 4.60	0.98 ± 0.83
CNN	Spectrogram	90.03 ± 3.24	98.53 ± 0.84	1.94 ± 2.35	0.97 ± 0.97	0.50 ± 0.86
Voting	Spectrogram + IPs	90.09 ± 3.19	98.53 ± 2.35	1.44 ± 1.48	0.97 ± 0.97	0.50 ± 0.86
Multi-Input CNN	Spectrogram + IPs	90.64 ± 2.65	**100.00** ± 0.00	**0.96** ± 1.66	**0.00** ± 0.00	**0.00** ± 0.00



(a) Multi-Input CNN.

(b) Spectrogram-based CNN.

(c) Time-IP-based CNN.

(d) Frequency-IP-based CNN.

Fig. 4. Mean of ROC curves over 4 folds.

97.07% of the positive events in the test set, with a ratio of ER and MR of 0.98% and 1.95% respectively, but a higher number of false positives than any other model with 4.37% FPR. Considering that the data is based on a projection of the spectrograms, the results were higher than we expected. In fact, the network dedicated to the whole spectrogram shares a similarly high recognition rate of 98.53%, but it performs better in terms of other metrics, achieving an ER and MR under 1% and an FPR of 1.94%. We obtained even better predictive capabilities by combining all these networks in the multi-input model that works as an ensemble with a shared classification head, resulting in 100% recognition with consequent no missed or misclassified events and a false positive ratio of only 0.96%, surpassing the performance of soft voting of the three single-input models, which results differ from the performance of the spectrogram-based CNN with only a reduction in false positives.

In Fig. 4, Receiver Operating Characteristic (ROC) curves, analytical tools used to evaluate the performance of classifiers, are presented. Note that these evaluations are independent of the experimental protocol proposed by the authors of the MIVIA Road Audio Events dataset. Each distinct curve within the illustration corresponds to a specific class category, thus adopting a “one-versus-all” strategy underlying class-based analysis. We ensure statistically robust results by implementing

Table 2
AUC values for different classes and architectures combinations (average of 4 folds).

Approach	Class BN	Class CC	Class TS
Multi-Input CNN	0.958 ± 0.022	0.964 ± 0.022	0.970 ± 0.021
Spectrogram-based CNN	0.954 ± 0.030	0.961 ± 0.029	0.970 ± 0.027
Time-IP-based CNN	0.920 ± 0.044	0.928 ± 0.044	0.913 ± 0.047
Frequency-IP-based CNN	0.907 ± 0.017	0.925 ± 0.025	0.935 ± 0.026

this strategy and applying 4-fold cross-validation. The shaded regions surrounding each curve serve as visual indicators, effectively depicting the confidence intervals surrounding the class-specific results. Table 2 summarizes the values reported in Fig. 4, to facilitate reading.

Consistent with the comprehensive outcomes elaborated in Table 1, it becomes discernible that the proposed multi-input methodology retains an advantage over its single-input counterparts. This distinction is particularly evident through the encapsulated Area Under Curve (AUC) values, briefly summarizing the holistic classifier performance.

Specifically, the multi-input approach achieves commendable AUC values, registering at 0.958 for the BN class, 0.964 for the CC class, and an elevated 0.97 for the TS class. A noteworthy observation arises from comparing these results with those derived from the single-input strategies. Among these, the spectrogram analysis-driven approach emerges as the most robust, yielding commendable AUC values of 0.954 for the BN class, 0.961 for the CC class, and an equivalent of 0.97 for the TS class. However, it is important to note, that this method reveals observable statistical variability, indicated by a comparatively more significant standard deviation of approximately 0.029, in contrast to the narrower 0.022 observed within the results of the multi-input method.

In Fig. 4, we illustrate the *Receiver Operating Characteristic (ROC)* curves to evaluate single and multi-input classifiers outside the experimental protocol proposed by the MIVIA Road Audio Events dataset authors. Each colored curve is associated with one class by employing a *one-vs-all* strategy and by averaging the results obtained over the 4-folds (the colored area represents the confidence interval of the individual results for each class): the x-axis reports the FPR values, and the y-axis the TPR values.

The results, consistent with Table 1, show an overall superiority of the proposed multi-input method, with an *Area Under Curve* (AUC) of 0.958 for the BN class, 0.964 for the CC class and 0.97 for the TS class. With regard to the single-input methods, the one based on spectrogram analysis achieves the best performance, reporting AUC values of 0.954 for the BN class, 0.961 for the CC class and 0.97 for the TS class. It is important to note, however, that this method reveals observable statistical variability, indicated by a comparatively larger standard deviation of approximately 0.029, in contrast to the narrower 0.022 observed within the results of the multi-input method.

In Fig. 5, we summarize the results of the ROC curves by averaging across classes, providing an overall view of the performance of all the classifiers discussed confirming that the multi-input network records the best AUC value at a lower variability of results.

5.2. Comparison against state-of-the-art approaches

Table 3 below compares the proposed method and state-of-the-art competitors on the MIVIA Road Audio Events dataset. Specifically, we consider *AReN* [35] and *COPE* [33], which mainly perform a gammatonegram analysis, and *MobileNet* [34], which instead focuses on the spectrogram. Among these, as outlined in the previous Section 2, AReN is more recent and performs better, as it shows 100% True Positive Rate and 2.01% in terms of False positive Rate on the MIVIA dataset. It also obtains 0% in terms of Miss Rate and Error Rate. This result appears to be slightly worse for MobileNet, which, however, with 99.5% TPR, 0% Miss Rate and 0.5% Error Rate is still very effective (considering that it is developed for deployment on embedded devices, therefore equipped

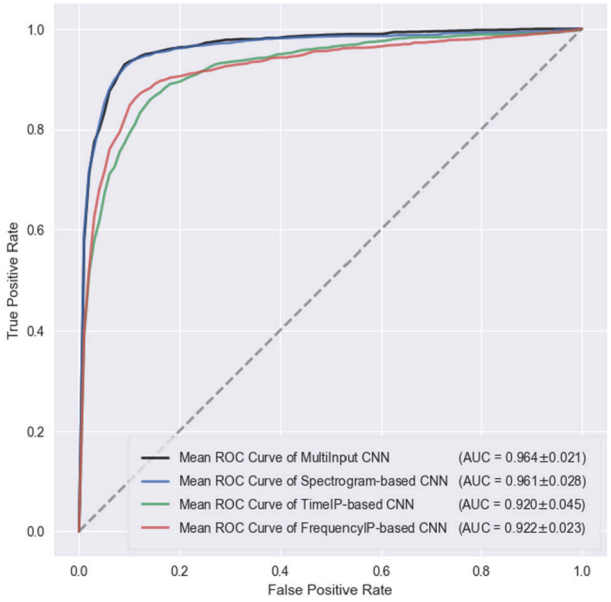


Fig. 5. Comparison of the ROC Curves for the considered architectures and input data types combinations.

Table 3
Results comparison between the proposed approach and the relevant literature competitors (†: higher is better; ‡: lower is better).

Method	Data Type	TPR†	FPR‡	MR‡	ER‡
AReN [35]	Gammatonegram	**100.00**	2.01	**0.00**	**0.00**
MobileNet [34]	Spectrogram	99.50	3.76	**0.00**	0.50
COPE [33]	Gammatonegram	94.00	3.95	4.75	1.25
Proposed	Spectrogram + IPs	**100.00**	**0.96**	**0.00**	**0.00**

with limited hardware). In contrast, with a TPR of 94%, COPE seems to be less performing. Overall, the proposed method, based on a multi-representative analysis of the spectrogram and its intensity projections, proved to be in line with the state-of-the-art solutions in this case study, leading to an improvement given by the halving of false alarms. We conjecture that this aspect is related to the increased ability of the proposed pipeline to better understand the complex input representation of the audio stream. In fact, it also emerges from the result of previous Table 1 that the application of a simple ensembling technique (i.e., the *voting*) on the output of single-input models, which separately analyze the spectrogram and its intensity projections, does not achieve the performance obtained instead by means of the proposed custom architecture, which shows a greater ability to identify the audio anomalies and reduce the false positives, due to its synergistic behavior.

We notice that, although several seminal work in literature [41–43] point out that the use of spectrograms and/or other 2D features might be more effective in classifying audio signals, many existing studies address the problem stated in this work through end-to-end systems based on 1D-based audio analysis methods. Hence, for the sake of completeness, we also decided to compare the performance of our solution against two public 1D-CNN models [44,45].

Table 4 shows the results of such a comparison. The reported values were obtained by training the aforementioned models on the MIVIA dataset's raw signals, with the same fragmentation used for the spectrogram extraction. The hyperparameters used were the *Adam* optimizer, the *binary cross-entropy* loss function, and a batch size of 32 samples. Although such models demonstrate a good performance in the analysis of the raw input, the obtained results seem to confirm the literature evidence, suggesting that these methods would not be able to outperform the 2D-based approaches proposed in this work.

Table 4
Comparison between our approach (best configurations) and 1D convolution models for audio classification (†: higher is better; ‡: lower is better).

Method	Data Type	ACC†	TPR†	FPR‡	MR‡	ER‡
Mansar [44]	Raw Signal	85.88 ± 2.93	96.10 ± 2.71	2.89 ± 3.96	2.92 ± 2.15	0.98 ± 0.98
Abdoli [45]	Raw Signal	87.50 ± 3.31	92.19 ± 3.54	0.97 ± 0.97	5.37 ± 1.55	2.43 ± 2.09
Proposed	Spectrogram	90.03 ± 3.24	98.53 ± 0.84	1.94 ± 2.35	0.97 ± 0.97	0.50 ± 0.86
Proposed	Spectrogram + IPs	90.64 ± 2.65	100.00 ± 0.00	0.96 ± 1.66	0.00 ± 0.00	0.00 ± 0.00

5.3. Prototype implementation

A prototype of the described pipeline - denoted with *CARgram* - implemented in *Python 3.8*, and which exploits the model pre-trained on the MIVIA dataset, has been deployed as part of an existing AI-based urban video surveillance system, under the project *SafeSpotter*.[1] Such a system leverages on a series of 4 fixed monocular cameras to monitor dangerous traffic junctions in a municipality ($\sim 20,000$ inhabitants) of the metropolitan city of Cagliari (Italy), in order to determine anomalies, hazardous behaviors and accidents.

In such a system, the captured video streams are interpreted by a Computer Vision-based artificial intelligence module based on *YOLO* [46] and heuristic algorithms, deputed to identify the aforementioned anomalies. However, the accident recognition component has from the beginning been prone to false positives: on the one hand, this is a consequence of choosing to minimize missed alarms (in order to improve rescue response time); on the other hand, it is due to the fact that occlusions and perspective make it complex to determine vehicle and/or pedestrian collision with high accuracy.

In this context, CARgram was used to support the visual AI module, by analyzing the contextual audio stream acquired by the cameras. When an incident type anomaly was detected, the live output of CARgram was then evaluated in order to confirm (or not) the detection of a hazardous sound.

Although the prototype was evaluated over a period of 6 months, with only one incident occurring (and correctly identified by the SafeSpotter system), during this period the integration of the CARgram module resulted in a reduction of false positives associated with accident alerts by $\sim 10\%$. Despite the limited time window and case history of events, this provided promising insights into the robustness of the proposed pipeline even in a real-world use scenario.

6. Conclusions

In urban surveillance, there is an ever-increasing need for advanced scientific and technological solutions to ensure higher and higher levels of service quality and safety for citizens. In this work, we have proposed an innovative method that exploits Deep Learning and Computer Vision techniques for automatically detecting abnormal traffic events (particularly collisions and hard braking) from audio signals acquired through environmental microphones. In particular, this work is the first to propose a multi-representational analysis of the information from the audio signal, exploiting in a synergistic mode its spectrogram and the intensity projections on the time and frequency axes derived from it. Therefore, for this purpose, we proposed a custom neural network (called Multi-Input CNN), whose architecture is designed to process the above signal representations synergistically. The results obtained through experimental validation on the public MIVIA dataset, with true positive rate values of 100% and false positive rate values of 0.96%, confirm the overall goodness of the proposed approach, as well as its superiority to both the respective single-input methods and literature competitors. Some limitations remain. First, we highlight that the MIVIA dataset covers only two classes of hazardous events, and the analysis thus requires to be extended to more general contexts. Furthermore, the proposed solution is oriented toward fixed urban settings and, therefore, needs to be tested, e.g., on autonomous driving vehicles: this represents an important future research direction.

[1] https://monserrato.etrasparenza.it/index.php?id_oggetto=11&id_doc=380586.

CRediT authorship contribution statement

Alessandro Sebastian Podda: Conceptualization, Methodology, Writing – original draft preparation, Validation. **Riccardo Balia:** Data curation, Writing – original draft preparation, Software. **Livio Pompianu:** Conceptualization, Visualization, Writing – reviewing and editing. **Salvatore Carta:** Conceptualization, Supervision, Validation. **Gianni Fenu:** Supervision, Validation. **Roberto Saia:** Conceptualization, Writing – reviewing and editing

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We used the public MIVIA dataset (available under request to the owners).

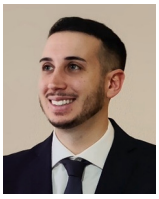
Acknowledgments

This research has been partially supported by the “Bando Aiuti per progetti di Ricerca e Sviluppo – POR FESR 2014-2020 – Asse 1, Azione 1.1.3” – project *Safespotter* – CUP C36H17000000006, and by the “Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5” - project *e.INS* – CUP F53C22000430001.

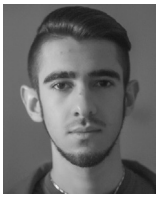
References

[1] WorldHealthOrganization, Road traffic injuries report, www.who.int, 2021.
[2] H.R. Champion, J. Augenstein, A.J. Blatt, B. Cushing, K. Digges, J.H. Siegel, M.C. Flanigan, Automatic crash notification and the urgency algorithm: its history, value, and use, Adv. Emerg. Nursing J. 26 (2) (2004) 143–156.
[3] W.E. Evanco, Impact of rapid incident detection on freeway accident fatalities, 1996.
[4] S.R.E. Datondji, Y. Dupuis, P. Subirats, P. Vasseur, A survey of vision-based traffic monitoring of road intersections, IEEE Trans. Intell. Transp. Syst. 17 (10) (2016) 2681–2698.
[5] S. Haria, S. Anchaliya, V. Gala, T. Maru, Car crash prevention and detection system using sensors and smart poles, in: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2018, pp. 800–804.
[6] L. Patrono, P. Rametta, L. Stefanizzi, An iot-aware remote monitoring system for emergencies in rallying, in: 2017 2nd International Multidisciplinary Conference on Computer and Energy Science (SpliTech), IEEE, 2017, pp. 1–5.
[7] A. Murty, K. Pavani, T.D. Kalyani, Automobile sos system using mem sensor, Indian J. Sci. Technol. 9 (2016) 17.
[8] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A review of video surveillance systems, J. Vis. Commun. Image Represent. 77 (2021) 103116.
[9] R. Balia, S. Barra, S. Carta, G. Fenu, A.S. Podda, N. Sansoni, A deep learning solution for integrated traffic control through automatic license plate recognition, in:

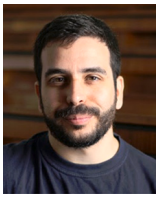
Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Proceedings, Part III 21, Cagliari, Italy, September 13–16, 2021, Springer, 2021, pp. 211–226.
[10] A. Atzori, S. Barra, S. Carta, G. Fenu, A.S. Podda, Heimdall: an ai-based infrastructure for traffic monitoring and anomalies detection, in: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops), IEEE, 2021, pp. 154–159.
[11] J. Li, H. Cheng, H. Guo, S. Qiu, Survey on artificial intelligence for vehicles, Automot. Innov. 1 (2018) 2–14.
[12] R. Abduljabbar, H. Dia, S. Liyanage, S.A. Bagloee, Applications of artificial intelligence in transport: an overview, Sustainability 11 (1) (2019) 189.
[13] N. Akaishi, K. Yatabe, Y. Oikawa, Harmonic and percussive sound separation based on mixed partial derivative of phase spectrogram, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 301–305.
[14] M. Huang, M. Wang, X. Liu, R. Kan, H. Qiu, Environmental sound classification framework based on l-mhp features and se-resnet50 network model, Symmetry 15 (5) (2023) 1045.
[15] L. Nanni, G. Maguolo, S. Brahnam, M. Paci, An ensemble of convolutional neural networks for audio classification, Appl. Sci. 11 (13) (2021) 5796.
[16] W. Mu, B. Yin, X. Huang, J. Xu, Z. Du, Environmental sound classification using temporal-frequency attention based convolutional neural network, Sci. Rep. 11 (1) (2021) 21552.
[17] E.K. Fishman, D.R. Ney, D.G. Heath, F.M. Corl, K.M. Horton, P.T. Johnson, Volume rendering versus maximum intensity projection in ct angiography: what works best, when, and why, Radiographics 26 (3) (2006) 905–922.
[18] S.S. Thomas, S. Gupta, V.K. Subramanian, Event detection on roads using perceptual video summarization, IEEE Trans. Intell. Transp. Syst. 19 (9) (2018) 2944–2954, https://doi.org/10.1109/TITS.2017.2769719.
[19] V. Machaca Arceda, E. Laura Riveros, Fast car crash detection in video, in: 2018 XLIV Latin American Computer Conference (CLEI), 2018, pp. 632–637.
[20] C. Wang, Y. Dai, W. Zhou, Y. Geng, A vision-based video crash detection framework for mixed traffic flow environment considering low-visibility condition, J. Adv. Transp. 2020 (2020).
[21] S. Robles-Serrano, G. Sanchez-Torres, J. Branch-Bedoya, Automatic detection of traffic accidents from video using deep learning techniques, Computers 10 (11) (2021), https://doi.org/10.3390/computers10110148, https://www.mdpi.com/2073-431X/10/11/148.
[22] Z. Neili, K. Sundaraj, A comparative study of the spectrogram, scalogram, melspectrogram and gammatonegram time-frequency representations for the classification of lung sounds using the icbhi database based on cnns, Biomed. Eng. (Biomedizinische Technik) 67 (5) (2022) 367–390.
[23] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M.D. Plumbley, Detection and classification of acoustic scenes and events, IEEE Trans. Multimed. 17 (10) (2015) 1733–1746.
[24] S. Rovetta, Z. Mnasri, F. Masulli, Detection of hazardous road events from audio streams: an ensemble outlier detection approach, in: 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), IEEE, 2020, pp. 1–6.
[25] M. Sammarco, M. Detyniecki, Crashzam: sound-based car crash detection, in: VEHITS, 2018, pp. 27–35.
[26] Y. Li, X. Li, Y. Zhang, M. Liu, W. Wang, Anomalous sound detection using deep audio representation and a blstm network for audio surveillance of roads, IEEE Access 6 (2018) 58043–58055.
[27] Y. Arslan, H. Canbolat, Performance of deep neural networks in audio surveillance, in: 2018 6th International Conference on Control Engineering & Information Technology (CEIT), IEEE, 2018, pp. 1–5.
[28] X. Zhang, Y. Chen, M. Liu, C. Huang, Acoustic traffic event detection in long tunnels using fast binary spectral features, Circuits Syst. Signal Process. 39 (2020) 2994–3006.
[29] E. Hidayat, D. Lange, J. Karlovsek, Exploring the interrelationships of variables in Australian road tunnel incidents using Bayesian networks, Prosiding KRTJ HPJI 16 (1) (2023) 1–15.
[30] R.T. Sound, Car accident detection and reconstruction through sound analysis with crashzam, in: Smart Cities, Green Technologies and Intelligent Transport Systems: 7th International Conference, SMARTGREENS, and 4th International Conference, VEHITS 2018, Funchal-Madeira, Portugal, March 16-18, 2018, Revised Selected Papers, vol. 992, Springer, 2019, p. 159.
[31] B. Kumar, A. Basit, M. Kiruba, R. Giridharan, S. Keerthana, Road accident detection using machine learning, in: 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE, 2021, pp. 1–5.
[32] Z. Mnasri, S. Rovetta, F. Masulli, Audio surveillance of roads using deep learning and autoencoder-based sample weight initialization, in: 2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON), IEEE, 2020, pp. 99–103.
[33] N. Strisciuglio, M. Vento, N. Petkov, Learning representations of sound using trainable cope feature extractors, Pattern Recognit. 92 (2019) 25–36.
[34] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, V. Vigilante, Detecting sounds of interest in roads with deep networks, in: International Conference on Image Analysis and Processing, Springer, 2019, pp. 583–592.
[35] A. Greco, N. Petkov, A. Saggese, M. Vento, Aren: a deep learning approach for sound event recognition using a brain inspired representation, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3610–3624.
[36] J. Kreiser, M. Meuschke, G. Mistelbauer, B. Preim, T. Ropinski, A survey of flattening-based medical visualization techniques, Comput. Graph. Forum 37 (3) (2018) 597–624, https://doi.org/10.1111/cgf.13445, https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13445.
[37] S. Jian, H. Kaiming, R. Shaoqing, Z. Xiangyu, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision & Pattern Recognition, 2016, pp. 770–778.
[38] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, Cascade classifiers trained on gammatonegrams for reliably detecting audio events, in: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2014, pp. 50–55.
[39] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, Audio surveillance of roads: a system for detecting anomalous sounds, IEEE Trans. Intell. Transp. Syst. 17 (1) (2015) 279–288.
[40] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, Reliable detection of audio events in highly noisy environments, Pattern Recognit. Lett. 65 (2015) 22–28.
[41] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomed. Signal Process. Control 47 (2019) 312–323, https://doi.org/10.1016/j.bspc.2018.08.035, https://www.sciencedirect.com/science/article/pii/S1746809418302337.
[42] M. Huzaifah, Comparison of time-frequency representations for environmental sound classification using convolutional neural networks, arXiv preprint, arXiv:1706.07156, 2017.
[43] M. Daouad, F.A. Allah, E.W. Dadi, An automatic speech recognition system for isolated amazigh word using 1d & 2d cnn-lstm architecture, Int. J. Speech Technol. 26 (3) (2023) 775–787.
[44] Y. Mansar, Audio classification: a convolutional neural network approach, https://github.com/CVxTz/audio_classification, 2018.
[45] S. Abdoli, P. Cardinal, A.L. Koerich, End-to-end environmental sound classification using a 1d convolutional neural network, Expert Syst. Appl. 136 (2019) 252–263.
[46] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.



Alessandro Sebastian Podda is an Assistant Professor at the Department of Mathematics and Computer Science of the University of Cagliari. He got a Ph.D. in Mathematics and Computer Science in 2018. Currently, he is Research Unit Coordinator (AI for eHealth and Smart Cities) at the Artificial Intelligence and Big Data Laboratory, as well as member of the Blockchain Laboratory of the University of Cagliari. He is also the former technical director and solution architect of the Doutdes and Sardioin projects and participates in numerous research projects including AlmostAnOracle, Safespotter and Mister. To date, he has been the co-author of more than 30 journal articles and conference papers.



Riccardo Balia is a Ph.D. student at the University of Cagliari, where he previously received his BSc and MSc in Computer Science. He is currently member of the Artificial Intelligence and Big Data Laboratory in the University of Cagliari, and participates in several research projects. His research fields include computer vision, audio analysis, and deep/machine learning applicated to different domains, e.g. computer-aided diagnosis, smart cites, energy.



Livio Pompianu is an Assistant Professor (RTDa) at the Department of Mathematics and Computer Science of the University of Cagliari. Livio started working as Research Fellow at the University of Cagliari during his MSc. Later, he continued his research activity in Cagliari, first as a PhD student (2014) and then as a post-doc (2018). In 2017 he was visiting PhD student at the University of Stirling (UK). He got his PhD in 2018 with a thesis entitled “Analysing blockchains and smart contracts: tools and techniques”. Currently, he is a member of the research groups: Distributed Ledger Technology, Blockchain@Unica, and Trustworthy Computational Societies. His research primarily focuses on information security and artificial intelligence topics. He is co-author of several scientific conference papers and journal articles in these research fields, with more than 300 citations. He also works as technical director and solution architect for research groups and companies, designing and developing software solutions (for instance, in the projects Score, Intellicredit).

Salvatore Mario Carta is Full Professor at the Department of Mathematics and Computer Science of the University of Cagliari. He received a Ph.D. in Electronics and Computer Science from the University of Cagliari in 2003 and in 2005 joined Department of Mathematics and Computer Science of the University of Cagliari as Assistant Professor. In 2006 and 2007 he joined the Swiss Federal Institute of Technology as Invited Researcher. He is author of more than 130 conference and journal papers in the research fields of Artificial Intelligence, Recommendation and Computer Vision, with more than 2000 citations. He is also member of the ACM and of the IEEE and he founded three hi-tech companies, spin-offs of the University of Cagliari, currently leading one of them.

Gianni Fenu is Full Professor of Computer Science at the Department of Mathematics and Computer Science at the University of Cagliari (Italy). He is the Vice-Rector of University of Cagliari and also the Director of the E-learning For Didactic Innovation Center. He received the MSc degree in Engineering at University of Cagliari in 1985, and joined the University of Cagliari in 1988. He teaches courses of Computer Networks at the BSc in Computer Science, and Digital Transformation at the MSc courses in Computer Science. His research interests include complex networks, recommender systems, e-learning and biometrics. He is author and co-author of more than 130 papers published in scientific journals or proceedings of refereed conferences. Many research projects were managed by Gianni Fenu as principal investigator.

Roberto Saia is a postdoctoral researcher at the Department of Mathematics and Computer Science of the University of Cagliari. He got a Master Degree and a PhD in Computer Science at the same University. His research is currently focused on several domains, such as those of the Recommender Systems, Data Mining, Artificial Intelligence, and Security. He is the author or co-author of tens scientific journals, articles, and books.