applied sciences

MDPI

Article

A Scalogram-Based CNN Approach for Audio Classification in Construction Sites

Michele Scarpiniti [1,*], Raffaele Parisi [1] and Yong-Cheol Lee [2]

[1] Department of Information Engineering, Electronics and Telecommunications (DIET), Sapienza University of Rome, via Eudossiana 18, 00184 Rome, Italy; raffaele.parisi@uniroma1.it
[2] Department of Construction Management, Louisiana State University, Baton Rouge, LA 70803, USA; yclee@lsu.edu
* Correspondence: michele.scarpiniti@uniroma1.it; Tel.: +39-06-44585869

Abstract: The automatic monitoring of activities in construction sites through the proper use of acoustic signals is a recent field of research that is currently in continuous evolution. In particular, the use of techniques based on Convolutional Neural Networks (CNNs) working on the spectrogram of the signal or its mel-scale variants was demonstrated to be quite successful. Nevertheless, the spectrogram has some limitations, which are due to the intrinsic trade-off between temporal and spectral resolutions. In order to overcome these limitations, in this paper, we propose employing the scalogramas a proper time–frequency representation of the audio signal. The scalogram is defined as the square modulus of the Continuous Wavelet Transform (CWT) and is known as a powerful tool for analyzing real-world signals. Experimental results, obtained on real-world sounds recorded in construction sites, have demonstrated the effectiveness of the proposed approach, which is able to clearly outperform most state-of-the-art solutions.

Keywords: automatic construction site monitoring (ACSM); environmental sound classification (ESC); deep learning; convolutional neural network (CNN); continuous wavelet transform (CWT); scalogram; audio processing

Citation: Scarpiniti, M.; Parisi, R.; Lee, Y.-C. A Scalogram-Based CNN Approach for Audio Classification in Construction Sites. *Appl. Sci.* **2024**, *14*, 90. https://doi.org/10.3390/app14010090

Academic Editors: Giovanni Costantini and Daniele Casali

Received: 20 November 2023
Revised: 15 December 2023
Accepted: 20 December 2023
Published: 21 December 2023

1. Introduction

In recent years significant research efforts have been made in the field of Environmental Sound Classification (ESC) [1], allowing significant results to be obtained in practical sound classification applications. This initiative has been enabled by the use of Convolutional Neural Networks (CNNs), which allowed a superior performance in image processing problems [2] to be obtained. In order to extend the use of CNNs to the field of audio processing, the audio input signal is usually transformed into suitable bi-dimensional image-like representations, such as spectrograms, mel-scale spectrograms, and other similar methods [3,4].

Recently, the approaches employed in ESC have been transferred to advancing the construction domain by converting vision-based work monitoring and management systems into audio-based ones [5–7]. In fact, audio-based systems not only are more cost-effective than video-based ones, but they also work more effectively in a construction field when sources are far from the light of sight of sensors, making these systems very flexible and appropriate for combining other sensor-based applications or Artificial Intelligence (AI)-based technologies [7]. Furthermore, the amount of memory and data flow needed to handle audio data is much smaller than the one needed for video data. In addition, audio-based systems outperform accelerometer-based ones since there is no need to place sensors onboard, thus promoting 360-degree-based activity detection and surveillance without having an illumination issue [8].

Such audio-based systems can be successfully used as Automatic Construction Site Monitoring (ACSM) tools [7,9–11], which can represent an invaluable instrument for project

managers to promptly identify severe and urgent problems in fieldwork and quickly react to unexpected safety and hazard issues [12–16].

ACSM systems are usually implemented by exploiting both machine learning (ML) and deep learning (DL) techniques [17]. Specifically, several ML approaches, including Support Vector Machines (SVMs), the k-Nearest Neighbors (k-NN) algorithm, the Multilayer Perceptron (MLP), random forests, Echo State Networks (ESN), and others, have already demonstrated their effectiveness in properly performing activity identification and detection in a construction site [5,16]. However, DL approaches generally outperform ML-based solutions providing much improved results [6]. We expect that DL techniques including CNNs, Deep Recurrent Neural Networks (DRNNs) implemented with the Long Short-Term Memory (LSTM) cell, Deep Belief Networks (DBNs), Deep ESNs, and others can produce more suitable and qualified performances than ML ones for robustly managing construction work and safety issues.

Approaches based on CNNs have demonstrated good flexibility and considerably convincing performance in these applications. In fact, CNNs exhibit advanced accuracy in image classification [18]. In order to meet the bi-dimensional format of images, the audio waveform can be transformed into a bi-dimensional representation by a proper time–frequency transformation. The main time–frequency representation used in audio applications is the spectrogram, i.e., the squared magnitude of the Short Time Fourier Transform (STFT) [19,20]. The spectrogram is very rich in peculiar information that can be successfully exploited by CNNs. Instead of using the STFT spectrogram, in audio processing, it is very common to use some well-known variants, such as the constant-Q spectrogram, which uses a log-frequency mapping, and the mel-scale spectrogram, which uses the mel-scale of frequency to better capture the intrinsic characteristic of the human ear. Similarly, the Bark and/or ERB scales can be used, producing other variants of the spectrogram [21].

Although the spectrogram representation and its variants provide an effective way to extract features from audio signals, they entail some limitations due to the unavoidable trade-off between the time and frequency resolutions. Unfortunately, it is hard to provide an adequate resolution in both domains: a shorter time window provides a better time resolution, but it reduces the frequency resolution, while using longer time windows improves the frequency resolution but obtains a worse time resolution. Even if some solutions have been proposed to mitigate such an unwanted effect (such as the time–frequency reassignment and synchrosqueezing approach [22]), the problem can still affect the performance of deep learning methods. Moreover, the issue is also complicated by the fact that sound information is usually available at different time scales that cannot be captured by the STFT.

Motivated by these considerations, in this paper, we propose a new approach for the automatic monitoring of construction sites based on CNNs and scalograms. The scalogram was defined as the squared magnitude of the Continuous Wavelet Transform (CWT) [23]. By overcoming the intrinsic time–frequency trade-off, the scalogram is expected to offer an advanced and robust tool to improve the overall accuracy and performance of ACSM systems. In addition, the wavelet transform allows to it work at different time scales, which is a useful characteristic for the processing of audio data. Hence, the main idea of the paper is to use the scalogram instead of the spectrogram as the input to a CNN-based deep learning model. Although the methodology is not new, the proposed idea has been extensively tested on real data acquired in construction sites and, compared to most popular state-of-the-art methodologies, shows clear and significant improvements.

The rest of this paper is organized as follows. Section 2 shows the related work. Section 3 introduces the CWT, while Section 4 describes the proposed approach. Then, Section 5 explains the adopted experimental setup. Section 6 describes some implementation aspects, while Section 7 shows the obtained numerical results and confirms the effectiveness of the proposed idea. Finally, Section 8 concludes the work and outlines some hints for future research.

2. Related Work

In the digital era, great and increasing attention has been devoted to research on automated methods for real-time monitoring of activities in construction sites [15,24,25]. These modern approaches are able to offer better performance with respect to the most traditional techniques, which are typically based on manual collection of on-site work data and human-based construction project monitoring. In fact, these activities are typically time-consuming, inaccurate, costly, and labor-intensive [13]. In the last years, the literature related to applications of deep learning techniques to the construction industry has been continuously increasing [26,27]. In particular, many works have been published describing proper exploitation of audio data [5,16].

The work of Cao et al. in [28] was one of the first attempts in this direction. They introduced an algorithm based on the processing of acoustic data for the classification of four representative excavators. This approach is based on some acoustic statistical features. Namely, for the first time the short frame energy ratio, concentration of spectrum amplitude ratio, truncated energy range, and interval of pulse (i.e., the time interval between two consecutive peaks) were developed in order to characterize acoustic signals. The obtained results were quite effective for this kind of source; however, no other types of equipment were considered.

Paper [29] proposed the construction of a dataset of four classes of equipment and tested several ML classifiers. The results obtained in this work were aligned to those shown in [5], which compared and assessed the accuracy of 17 classifiers on nine classes of equipment. These two papers work on both temporal and spectral features extracted from audio signals. Similarly, [30] compared some ML approaches on five input classes by using a single in-pocket smartphone, obtaining similar numerical results.

Akbal et al. [14] proposed an SVM classifier. After an iterative neighborhood component analysis selector chooses the most significant features extracted from audio signals, this classifier produces an effective accuracy on two experimental scenarios. Moreover, Kim et al. [7] proposed a sound localization framework for construction site monitoring able to work in both indoor and outdoor scenarios.

Maccagno et al. [31] proposed a deep CNN-based approach for the classification of five pieces of construction site machinery and equipment. This customized CNN is fed by the STFT spectrograms extracted from different-sized audio chunks. Similarly, Sherafat et al. [32] proposed an approach for multiple-equipment activity recognition using CNNs, tested on both synthetic and real-world equipment sound mixtures. Different from [31], this work implements a data augmentation method to enlarge the used dataset. Moreover, this model uses a moving mode function to find the most frequent labels in a period ranging from 0.5 to 2 s, which generates an acceptable output accuracy. The idea to join different output labels inside a short time period was also exploited in [33,34], which implement a Deep Belief Network (DBN) classifier and an Echo State Network (ESN), respectively.

Kim et al. in [35] applied CNNs and RNNs to spectrograms for monitoring concrete pouring work in construction sites, while Xiong et al. in [6] used a convolutional RNN (CRNN) for activity monitoring. Moreover, Peng et al. in [36] used a similar DL approach for a denoising application in construction sites. On the other hand, Akbal et al. [37] proposed an approach, called DesPatNet25, which extracts 25 feature vectors from audio signals by using the data encryption standard cipher and adopts a k-NN and an SVM classifier to identify seven classes.

Additionally, some other approaches also fused information from two different modalities. For example, the work in [38] used an SVM classifier by combining both auditory and kinematics features, showing an improvement of about 5% when compared to the use of only individual sources of data. Similarly, [39] exploited visual and kinematic features, while [40] utilized location data from a GPS and a vision-based model to detect construction equipment. Finally, a multimodal audio–video approach was presented in [41], based on the use of different correlations of visual and auditory features, which has shown an overall improvement in detection performance.

In addition, Elelu et al. in [42] exploited CNN architectures to automatically detect collision hazards between construction equipment. Similarly, the work in [43] presented a critical review of recent DL approaches for fully embracing construction workers' awareness of hazardous situations in construction sites by the employment of auditory systems.

Most of the DL approaches described in this section work on the spectrogram extracted from audio signals or some variants, such as the mel-scaled spectrogram. However, the idea of exploiting different time scales (which is an intrinsic property of audio signals) can be used to improve the overall accuracy of such methodologies. For this purpose, the use of scalograms can be recommended. In fact, while spectrograms are suitable for the analysis of stationary signals providing a uniform resolution, the scalogram is able to localize transients in non-stationary signals. Recently, in fact, [44] introduced a wavelet filter bank for the audio scene modeling task. A deep CNN fed by the scalogram of data outperformed the results provided by the mel spectrogram. However, differently from our approach, the work in [44] considers a scalogram of smaller size and a simpler CNN architecture. The work in [45] adopted scalograms for removing background noise in the fault diagnosis of rotating machinery, obtaining excellent experimental results. However, differently from our approach, given the specific nature of the considered sounds, the authors used a low sampling frequency and frame size, resulting in a very small scalogram size (64×64 pixels). Interestingly enough, [45] considers three different methods to obtain the scalograms, including the CWT. No significant statistical differences have been observed between such methods. In addition, a couple of papers used scalograms also for audio scene classification purposes [46,47]. Both of these works showed very good results when compared to previous solutions. As a matter of fact, the use of the scalogram results in a general improvement in performance as highlighted in all these works. Specifically, the work in [46] exploits a pre-trained CNN to extract, at a specific architecture-dependent layer, useful features to be used by a subsequent linear SVM classifier for the identification of ten environmental categories. This work also uses AlexNet but, differently from our approach, it does not train the CNN layers and does not adopt fully connected layers as a classifier. The work in [47] again uses a pre-trained AlexNet or VGG16/19 nets to extract meaningful features, but, differently, it exploits a Bidirectional Gated Recurrent Neural Network followed by a highway layer to classify fifteen classes. Differently from our approach, the authors of [47] adopt an early data fusion technique by feeding the proposed model with a three-channel image composed of a spectrogram, a scalogram extracted with the Bump wavelet, and a scalogram obtained with the Morse wavelet. However, the high computational cost of this approach, compared with the proposed one, makes it not very suitable for working with the construction site sounds, where only a small number of classes are present.

3. The Continuous Wavelet Transform (CWT) and the Scalogram

In order to overcome the trade-off between the time and frequency resolution in STFT, the Continuous Wavelet Transform (CWT) was introduced [23]. The CWT acts as a "mathematical" microscope in the sense that different parts of the signal may be examined by adjusting the focus.

Given a stationary signal $x(t)$, the CWT is defined as the product of $x(t)$ with the following basis function family:

$$\Psi_{\tau,a}(t) = |a|^{-1/2}\,\Psi\left(\frac{t-\tau}{a}\right), \tag{1}$$

where $a \neq 0$ is a scaling factor (also known as dilation parameter) and τ is the time delay, i.e., $\Psi_{\tau,a}(t)$ is a scaled and translated version of the mother wavelet function $\Psi(t)$. Hence, the CWT of signal $x(t)$ is formulated as:

$$W_x(\tau,a) = |a|^{-1/2}\int_{-\infty}^{\infty} x(t)\Psi^*\left(\frac{t-\tau}{a}\right)dt, \tag{2}$$

where * represents the complex conjugation operator. The delay parameter τ provides the time position of the wavelet $\Psi_{\tau,a}(t)$, while the scaling factor a rules its frequency content. For $|a| << 1$, the wavelet $\Psi_{\tau,a}(t)$ is a very concentrated and narrow version of the mother wavelet $\Psi(t)$, with a frequency content mainly condensed at high frequencies. On the other hand, for $|a| >> 1$, the wavelet $\Psi_{\tau,a}(t)$ is much more broadened and concentrated towards low frequencies.

In the wavelet analysis, the similarity between the signal $x(t)$ and the wavelet $\Psi_{\tau,a}(t)$ is measured as τ and a vary. Dilation by a factor $1/a$ results in different enlargements of the signal with distinct resolutions. Specifically, the properties of the time–frequency resolution of the CWT are summarized as follows:

1. The temporal resolution $\Delta\tau$ varies inversely to the carrier frequency ω_0 of the wavelet $\Psi_{\tau,a}(t)$; therefore, it can be made arbitrarily small at high frequencies.
2. The frequency resolution $\Delta\omega$ varies linearly with the carrier frequency ω_0 of the wavelet $\Psi_{\tau,a}(t)$; therefore, it can be made arbitrarily small at low frequencies.

Hence, the CWT is well suited for the analysis of non-stationary signals containing high-frequency transients superimposed on long-lasting low-frequency components [23].

The CWT implements the signal analysis at various time scales. For this reason, the squared absolute value of the CWT is called a scalogram, and it is defined as:

$$\mathcal{S}(\tau,a) \triangleq |W_x(\tau,a)|^2 = \frac{1}{a}\left|\int_{-\infty}^{\infty} x(t)\Psi^*\left(\frac{t-\tau}{a}\right)dt\right|^2. \quad (3)$$

The scalogram $\mathcal{S}(\tau,a)$ provides a bi-dimensional graphical representation of the signal energy at the specific scale parameter a and time location τ.

In general, the mother wavelet $\Psi(t)$ can be any band-pass function [23]. The Haar wavelet is the simplest example of a wavelet, while the Daubechies one is a more sophisticated example. Both of these wavelets have a finite (and compact) support in time. The Daubechies wavelet has a longer length than the Haar wavelet and is therefore less localized than the latter. However, the Daubechies wavelet is continuous and has a better frequency resolution than the Haar one [23]. Other famous wavelet families are the Mexican Hat wavelet (which is proportional to the second derivative function of the Gaussian probability density function), the Bump wavelet, the generalized Morse wavelet, and the Morlet one, also known as the Gabor wavelet. This last wavelet is composed of a complex exponential multiplied by a Gaussian window, and it is very suitable for audio and vision applications since it is closely related to human perception. For this purpose, we remark that it is strongly related to the short-time analysis performed by the peripheral auditory system and to the mechanical spectral analysis performed by the basilar membrane in the human ear [48]. As a matter of fact, the Morlet wavelet is the most widely used wavelet for audio applications [49], and its effectiveness has been shown in analyzing machine sounds [50]. Motivated by these considerations, in the rest of the paper, we use the Morlet wavelet, which is defined as:

$$\Psi(t) = C_\psi\, e^{-\frac{t^2}{2\sigma^2}} e^{j\omega_0 t}, \quad (4)$$

where C_ψ is a normalization factor used to meet the admissibility condition, ω_0 is the central frequency of the mother wavelet (the carrier), and σ^2 is the variance of the Gaussian window equal to: $\sigma = n/\omega_0$. The parameter n, called the number of wavelet cycles and set in this paper to $n = 6$, defines the time–frequency precision trade-off.

4. Proposed Approach

Scalograms obtained from CWT are very rich in information and can improve the results obtained by other approaches, such as the spectrogram or its mel-scale version. The proposed idea consists in extracting the scalograms from the recorded signals after splitting them into chunks of a suitable length (usually 30–50 ms). The extracted scalograms, saved as image files, are fed as input to a CNN architecture. In fact, it is well known that CNNs are very effective for image classification. The literature is rich in state-of-the-art CNNs

that perform very well in image classification. Since the number of classes considered in a construction site is limited, and given the richness of the input representation (the scalogram), in this work, we propose the use of a simple CNN, i.e., the AlexNet one (see Section 5.3). A picture of the proposed idea is shown in Figure 1. A step-by-step flowchart of the proposed methodology is shown in Figure 2.

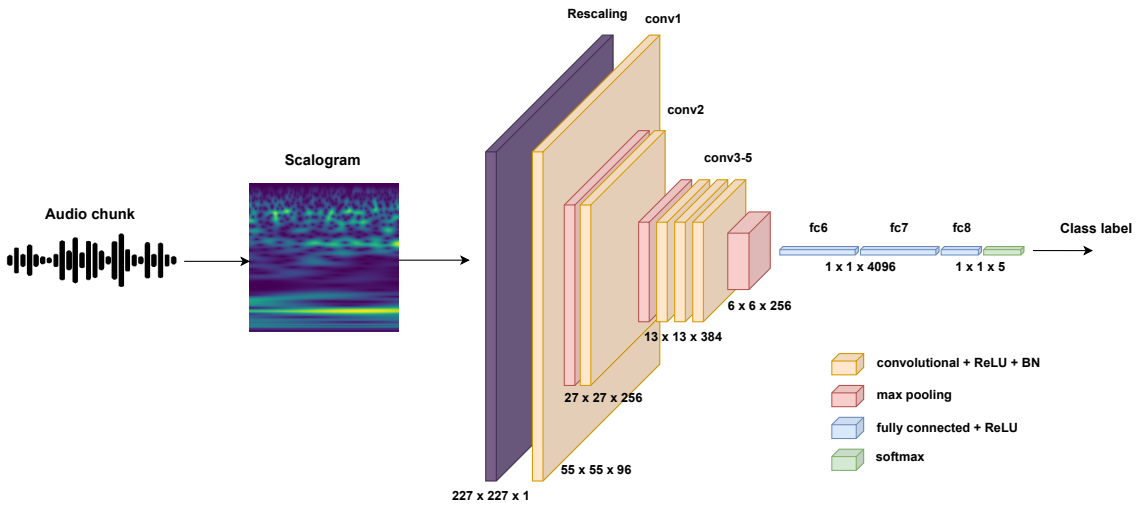


Figure 1. A picture of the proposed idea.

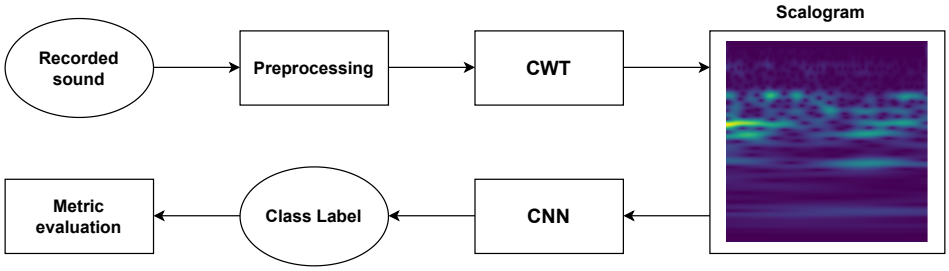


Figure 2. A step-by-step flowchart of the proposed methodology.

5. Experimental Setup

5.1. Dataset

The used dataset consists of a set of recordings related to five machines working in a real-world construction site. Sounds have been recorded with a Zoom H1 digital recorder with a sampling frequency of 44,100 Hz and saved as wave files. The five classes considered in this work are related to three excavators (two compact excavators and a large one), a compactor, and a concrete mixer. For each class, 15 min of recordings are available. The recording of each piece of machinery has been made by placing the recorder about 5–6 m in front of the activities of interest, without any obstacle in the middle. The recorded sounds are related to normal construction site activities (i.e., excavation and concrete mixing work) performed in outdoor scenarios. The subset of sounds considered in this work is related to a single source at a time; segments where more than one piece of equipment is active at the same time have been preventively removed from the dataset. Additional details on the operating scenario can be found in [5].

Each file has been split into chunks of 30 ms each. The entire dataset has been split into a training and a test set, with proportions of 75% and 25%, respectively. In addition, 10% of the training set has been devoted as the validation set to check the convergence performance during the training phase. Details of the used dataset, along with the number of chunks and related training/test splits, are reported in Table 1.

Table 1. Details of the used dataset.

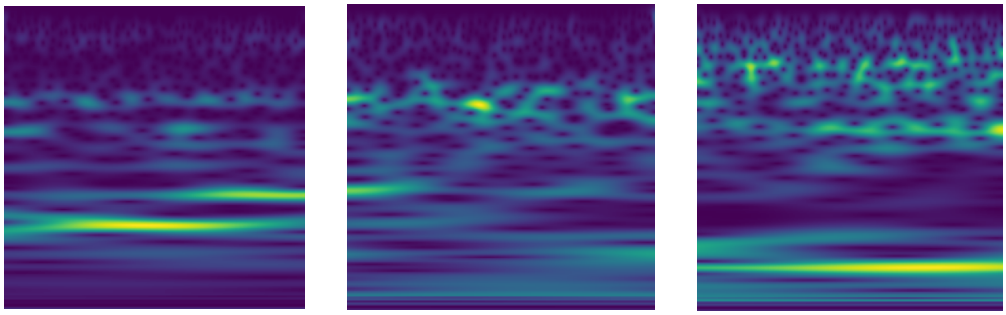
N.	Class	Equipment	Data	Chunks	Split
1	JD50D	Compact excavator John Deere 50D	15:00	30,000	22,500/7500
2	IRCOM	Ingersoll Rand Compactor	15:00	30,003	22,502/7501
3	Mixer	Concrete mixer Mercedes-Benz Actros	14:59	29,999	22,499/7500
4	CAT320E	Hydraulic excavator Caterpillar 320E	15:00	30,001	22,500/7501
5	Hitachi50U	Compact excavator Hitachi ZX50U	14:59	29,999	22,499/7500
		Total	01:14:58	150,002	112,500/37,502

5.2. Preprocessing

After the audio signals have been split into chunks of 30 ms, they have been resampled to 22,050 Hz for memory-saving purposes. This resampling procedure does not affect the quality of the classification, since the energy of audio signals related to construction sites is vanishing at frequencies higher than 10 kHz.

For each resampled chunk, the CWT has been extracted (we used the Python ssqueezepy package, available at: https://github.com/OverLordGoldDragon/ssqueezepy, accessed on 10 November 2023). The Morlet wavelet [23] has been used in this work. The obtained matrix has been then resized to 227×227 in order to be compliant with the input layer of the used AlexNet (see Section 5.3). For simplicity and memory-saving purposes, the obtained resized matrix has been rescaled to the interval $[0, 255]$, converted to integer numbers, and saved as images.

Some random images related to the extracted CWT from Classes 1, 3, and 4, respectively, are shown in Figure 3. These scalograms clearly capture salient localized events in sound frames, as shown by the horizontal lines or cloud-like points in Figure 3. Spectrograms of the same signals are generally unable to capture salient time/scale characteristics.



(**a**) (**b**) (**c**)

Figure 3. Examples of some scalogram images: (**a**) Class 1, (**b**) Class 3, and (**c**) Class 4.

5.3. Model

The literature is rich in well-performing and famous CNN architectures, as well as customized models for specific applications. Since the problem has been converted into a standard image classification task and the number of classes is limited, in this paper, we consider the well-known AlexNet [51] architecture. Specifically, AlexNet is composed of a cascade of five convolutional layers and three (dense) fully connected ones.

With respect to the original version, we introduce three modifications:

1. The number of channels of the input layer is reduced to only one since the network is fed by the scalogram, which is a single-channel image;
2. We add, after the input layer, a Rescaling layer in order to transform the integer input into floating-point numbers inside the interval $[0, 1]$;
3. The number of output classes has been reduced to five (the original AlexNet works with 1000 classes).

The details of the network organization, layers' shape, and number of parameters of the customized version of AlexNet are summarized in Table 2. Refer also to Figure 1 for a graphical representation.

AlexNet has been trained by minimizing the categorical cross-entropy defined as:

$$\mathcal{L}(y, \widehat{y}, \boldsymbol{\theta}) \triangleq -\frac{1}{B} \sum_{n=1}^{B} \sum_{i=1}^{N_C} y_n^{(i)} \log \widehat{y}_n^{(i)}, \tag{5}$$

where $\boldsymbol{\theta}$ is the vector collecting all of the network parameters, $N_C = 5$ is the number of classes, B is the mini-batch size, $y_n^{(i)}$ is the actual label of the n-th sample and i-th class, and $\widehat{y}_n^{(i)}$ is the corresponding predicted label. The minimization is performed by the gradient descent algorithm:

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \eta \nabla_{\boldsymbol{\theta}_k} \mathcal{L}(y, \widehat{y}, \boldsymbol{\theta}_{k-1}), \tag{6}$$

where η is the learning rate and k is the iteration index; the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\cdot)$ is computed over a mini-batch. In this work, the Adam optimizer, a variant of the gradient descent, has been used [52]. Specifically, the Adam algorithm incorporates an estimate of the first- and second-order moments of the gradient with a bias correction to speed up the convergence process. Details of the Adam algorithm can be found in [52]. The learning rate is set to $\eta = 10^{-4}$ (parameters β_1, β_2, and ε are left at their default values), and a batch size of $B = 32$ is used. The training is run for 10 epochs.

Table 2. Layers and number of parameters of the customized AlexNet.

Layer (Type)	Output Shape	Number of Parameters
Rescaling	(None, 227, 227, 1)	0
Conv2D	(None, 55, 55, 96)	11,712
BatchNormalization	(None, 55, 55, 96)	384
MaxPooling2D	(None, 27, 27, 96)	0
Conv2D	(None, 27, 27, 256)	614,656
BatchNormalization	(None, 27, 27, 256)	1024
MaxPooling2D	(None, 13, 13, 256)	0
Conv2D	(None, 13, 13, 384)	885,120
BatchNormalization	(None, 13, 13, 384)	1536
Conv2D	(None, 13, 13, 384)	1,327,488
BatchNormalization	(None, 13, 13, 384)	1536
Conv2D	(None, 13, 13, 256)	884,992
BatchNormalization	(None, 13, 13, 256)	1024
MaxPooling2D	(None, 6, 6, 256)	0
Flatten	(None, 9216)	0
Dense	(None, 4096)	37,752,832
Dropout	(None, 4096)	0
Dense	(None, 4096)	16,781,312
Dropout	(None, 4096)	0
Dense	(None, 5)	20485
	Total parameters:	58,284,101
	Trainable parameters:	58,281,349
	Non-trainable parameters:	2752

6. Implementation Aspects

In this section, we provide some important remarks about the implementation aspects of the proposed idea.

The computation of the CWT can be memory and computationally demanding. For this reason, we recommend not exceeding the chunk size; 30 ms or 50 ms represents a good compromise between the efficiency and tracking performance of the classifier due to the intrinsic non-stationarity of audio signals.

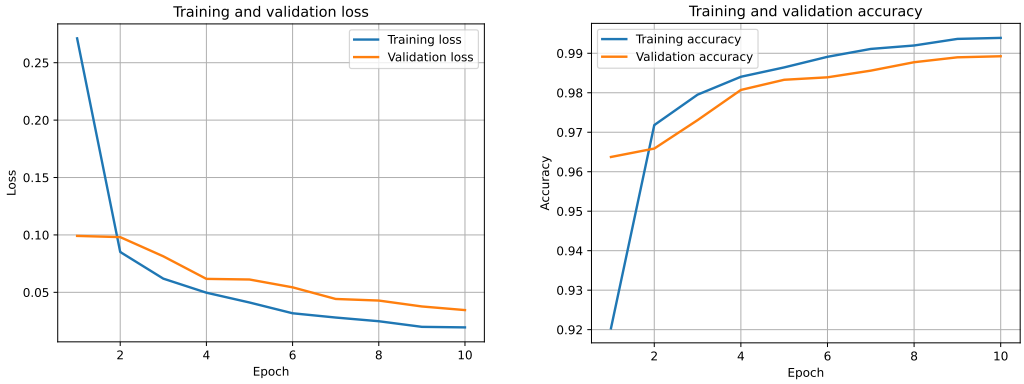
The CWT applied to a 30 ms chunk returns a matrix of the size 230 × 662. In view of using such data as the input to a state-of-the-art CNN, it is convenient to resize the matrix to a commonly used size. Generally, 227 × 227 (for AlexNet) or 224 × 224 (for GoogLeNet, ResNet, and similar architectures) are adequate choices.

However, saving more than 150,000 (see Table 1) floating-point matrices of 227 × 227 entries requires a large amount of disk space and a consistent quantity of RAM memory to load and process the dataset. For this purpose, after the resize, these matrices have been scaled to the interval [0, 255], converted to integer numbers, and saved as images. In this way, it is possible to work with this dataset on a normal office PC while avoiding memory explosion.

Finally, to deal with such data, an additional Rescaling layer has been used in the customized version of AlexNet. This layer converts the integer input data back into the float interval [0, 1].

7. Experimental Results

The proposed model has been trained on the considered dataset for 10 epochs by using 10% of the training set as the validation set (Python 3.10 source code can be downloaded from: https://github.com/mscarpiniti/CS-scalogram, accessed on November 20). The training and validation losses obtained during the training phase are shown in Figure 4a, while Figure 4b shows the corresponding training and validation accuracy. These figures demonstrate the effectiveness of the training, showing that the training procedure is quite stable after about seven epochs. Figure 4b also shows that, at convergence, the training accuracy is about 99.5%, while the validation one is about 99%.



(**a**) Loss. (**b**) Accuracy.

Figure 4. Training and validation loss (**a**) and accuracy (**b**) of the proposed approach.

To evaluate the proposed approach, we have also used the overall accuracy, the per-class precision, the per-class recall, and the per-class F1-score, as well as their weighted averages [53], computed on the test set. Moreover, the confusion matrix is shown in Figure 5. The confusion matrix clearly shows that the proposed approach is able to provide very good results for the classification of real-world signals recorded in construction sites. In fact, most of the instances are in the main diagonal of the matrix. There is a little confusion between the compactor (IRCOM), which has been confused with the JD50D excavator and the concrete mixer, and the CAT320E excavator, which is, again, mainly confused with the JD50D excavator and the concrete mixer. This behavior is due to the fact that all of these pieces of equipment have similar engines.

The results in terms of the precision, recall, and F1-score of the proposed approach are summarized in Table 3. In addition, this table confirms the conclusion drawn from the confusion matrix in Figure 5: the JD50D and Concrete Mixer classes have lower precision, while the compactor (IRCOM) and CAT320E excavator show lower recall. However, the F1-score is quite stable among all classes. The Hitachi 50U excavator performs the best

in between the five considered classes. Despite this slight variability in performance, the weighted averages of the considered metrics are very good and settled at 0.989.

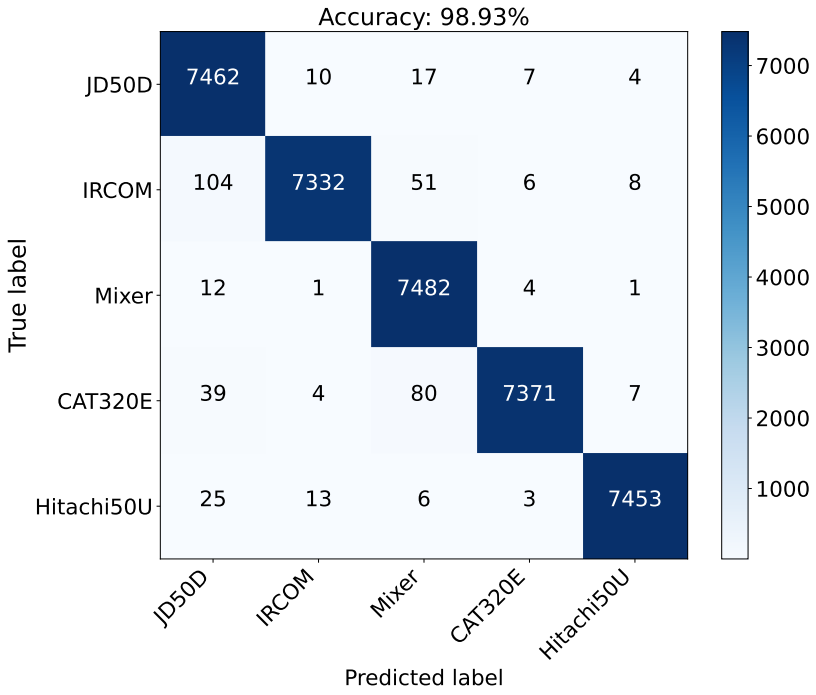


Figure 5. Confusion matrix obtained by the proposed approach.

Table 3. Per-class performance of the proposed approach.

Class	Precision	Recall	F1-Score
JD50D	0.976	0.995	0.986
IRCOM	0.996	0.977	0.987
Mixer	0.979	0.998	0.988
CAT320E	0.997	0.983	0.989
Hitachi50U	0.997	0.994	0.996
All classes	0.989	0.989	0.989

The proposed approach was compared with similar state-of-the-art solutions. Specifically, we compared our approach to the one proposed by Piczak in [4], based on a CNN fed by the spectrograms with corresponding deltas (i.e., the difference of the feature among two consecutive time instants); the approach proposed by Maccagno et al. in [31], based on a custom deep CNN (DCNN) fed by the spectrograms; and the approach proposed by Scarpiniti et al. in [34], based on an ESN working on several spectral features and a majority voting between adjacent chunks. The results obtained by these state-of-the-art approaches in terms of precision, recall, F1-score, and their weighted averages are shown in Tables 4, 5, and 6, respectively. The results presented in these tables confirm that the approach proposed in this paper (see Table 3) performs better than the state of the art for all of the considered metrics. Figure 6 summarizes all of the considered metrics for these compared approaches.

Table 4. Per-class performance of the Piczak approach in [4].

Class	Precision	Recall	F1-Score
JD50D	0.981	0.965	0.973
IRCOM	0.959	0.982	0.970
Mixer	0.942	0.945	0.943
CAT320E	0.894	0.973	0.932
Hitachi50U	0.944	0.795	0.863
All classes	0.944	0.932	0.936

Table 5. Per-class performance of the DCNN-based approach in [31].

Class	Precision	Recall	F1-Score
JD50D	0.955	0.972	0.963
IRCOM	0.957	0.979	0.968
Mixer	0.975	0.985	0.980
CAT320E	0.986	0.973	0.979
Hitachi50U	0.972	0.978	0.975
All classes	0.973	0.973	0.973

Table 6. Per-class performance of the ESN-based approach in [34].

Class	Precision	Recall	F1-Score
JD50D	0.901	0.937	0.919
IRCOM	0.899	0.974	0.935
Mixer	0.837	0.819	0.828
CAT320E	0.769	0.629	0.692
Hitachi50U	0.763	0.823	0.792
All classes	0.834	0.837	0.833

In addition, Tables 7 and 8 show the results of the works proposed by [44,46], which use scalogram-based approaches for acoustic scene classification. We adapt these approaches to work with the scalograms extracted from the construction site sounds. These tables show that, although the works proposed in [44,46] provide good results, the performance is slightly lower than the proposed approach reported in Table 3.

Table 7. Per-class performance of the approach proposed by Chen et al., 2018, in [44].

Class	Precision	Recall	F1-Score
JD50D	0.974	0.975	0.974
IRCOM	0.982	0.979	0.980
Mixer	0.981	0.984	0.982
CAT320E	0.988	0.979	0.984
Hitachi50U	0.975	0.983	0.979
All classes	0.980	0.980	0.980

Table 8. Per-class performance of the approach proposed by Copiaco et al., 2019, in [46].

Class	Precision	Recall	F1-Score
JD50D	0.972	0.973	0.972
IRCOM	0.981	0.977	0.979
Mixer	0.979	0.982	0.981
CAT320E	0.986	0.977	0.982
Hitachi50U	0.972	0.981	0.977
All classes	0.978	0.978	0.978

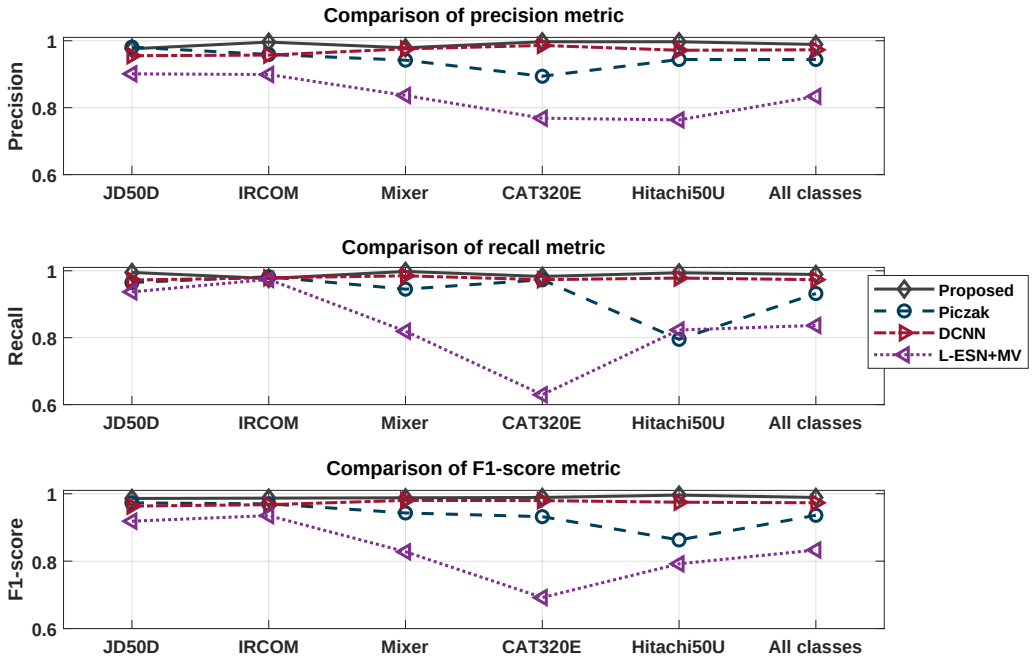


Figure 6. A visual comparison of all the compared approaches for the precision (**top**), recall (**middle**), and F1-score (**bottom**).

Although the Morlet wavelet in (4) is the most used and effective wavelet family in nonstationary audio analysis, we have also tested two other well-known and well-used wavelet families: the generalized Morse and the Bump wavelets [47], respectively. The results in terms of per-class precision, recall, and F1-score, and their related weighted averages, are shown in Table 9. From this table, we can argue that results of the generalized Morse wavelet are quite similar to those obtained by using the Morlet one (see Table 3). On the other hand, the results related to the Bump wavelet are slightly worse, even if they are quite good. The overall accuracies of these two approaches were 98.50% and 97.45%, respectively. These considerations confirm the effectiveness of the Morlet wavelet for analyzing audio signals in general and engine sounds in particular.

Finally, in Table 10, we summarize the previous results of the proposed approach (last row) and the compared ones by considering some additional machine learning and deep learning approaches. Specifically, Figure 7 shows the accuracy of the compared approaches as a bar plot. Among the machine learning techniques, we considered the results obtained by using a Support Vector Machine (SVM), the k-Nearest Neighbors (k-NN), the Multilayer Perceptron (MLP), and a random forest. All of these approaches provided reasonable results [5], though the results were worse than those provided by deep learning techniques. Among these last methods, we also considered an approach based on a Deep Recurrent Neural Network (DRNN) that exploits different spectral features [54] and one based on a Deep Belief Network (DBN) that works on a statistical ensemble of different spectral features [33]. For the implementation details, we refer to the related references. The results

reported in Table 10 and Figure 7 clearly show once again the effectiveness of the proposed idea, which can be considered an effective and reliable approach for classifying real-world signals recorded in construction sites.

Table 9. Per-class performance of the proposed approach by using the generalized Morse wavelet and the Bump wavelet. Overall accuracy is 98.50% and 97.45%, respectively.

Class	Generalized Morse Wavelet			Bump Wavelet		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
JD50D	0.985	0.977	0.981	0.964	0.983	0.974
IRCOM	0.986	0.983	0.985	0.990	0.967	0.978
Mixer	0.975	0.994	0.984	0.989	0.960	0.974
CAT320E	0.985	0.986	0.986	0.995	0.966	0.980
Hitachi50U	0.995	0.985	0.990	0.938	0.996	0.966
All classes	0.985	0.985	0.985	0.975	0.974	0.975

Table 10. Results of the compared approaches.

Approach	Accuracy	Precision	Recall	F1-Score
SVM [5]	83.66	0.846	0.838	0.842
k-NN [5]	85.28	0.860	0.853	0.857
MLP [5]	91.06	0.913	0.932	0.923
Random Forest [5]	93.16	0.934	0.932	0.933
Piczak [4]	90.03	0.944	0.932	0.936
DRNN [54]	95.32	0.955	0.953	0.954
DCNN [31]	97.08	0.973	0.973	0.973
DBN [33]	97.79	0.978	0.978	0.978
L-ESN+MV [34]	95.26	0.957	0.953	0.952
CNN [44]	97.99	0.980	0.980	0.980
CNN+SVM [46]	97.79	0.978	0.978	0.978
Proposed	98.93	0.989	0.989	0.989

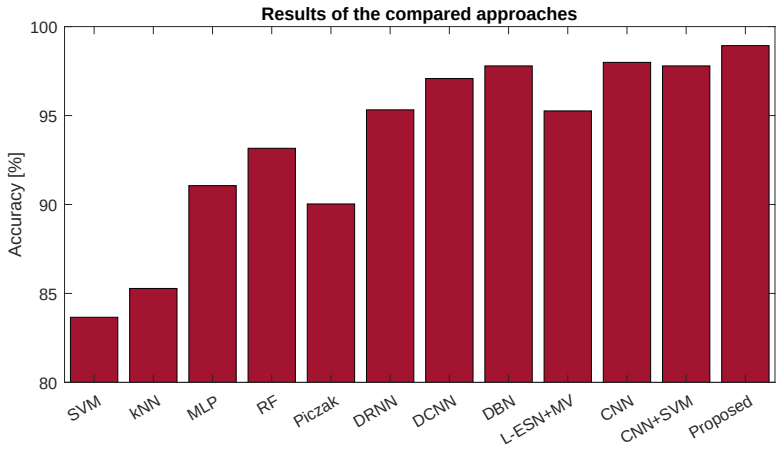


Figure 7. Accuracy of the compared approaches.

As a discussion, we can observe that scalograms generally capture salient localized events in sound frames, as shown by the horizontal lines or cloud-like points in Figure 3. In addition, the convolutional layers of the CNN are able to learn discriminative features, as shown in Figure 8, which, for example, shows the 256 feature maps of the fifth and last convolutional layer of the used architecture. Although single plots in the figure are quite small, it is clear that feature maps in the final layer are more specialized in detecting specific time scales. In fact, the scalogram in Figure 8 shows clear horizontal lines localized at a

specific scale. This kind of scale localization is typical of engines, a fundamental part of the machines considered in our dataset. This behavior justifies the better performance of the proposed approach for the classification of equipment sounds in construction sites.

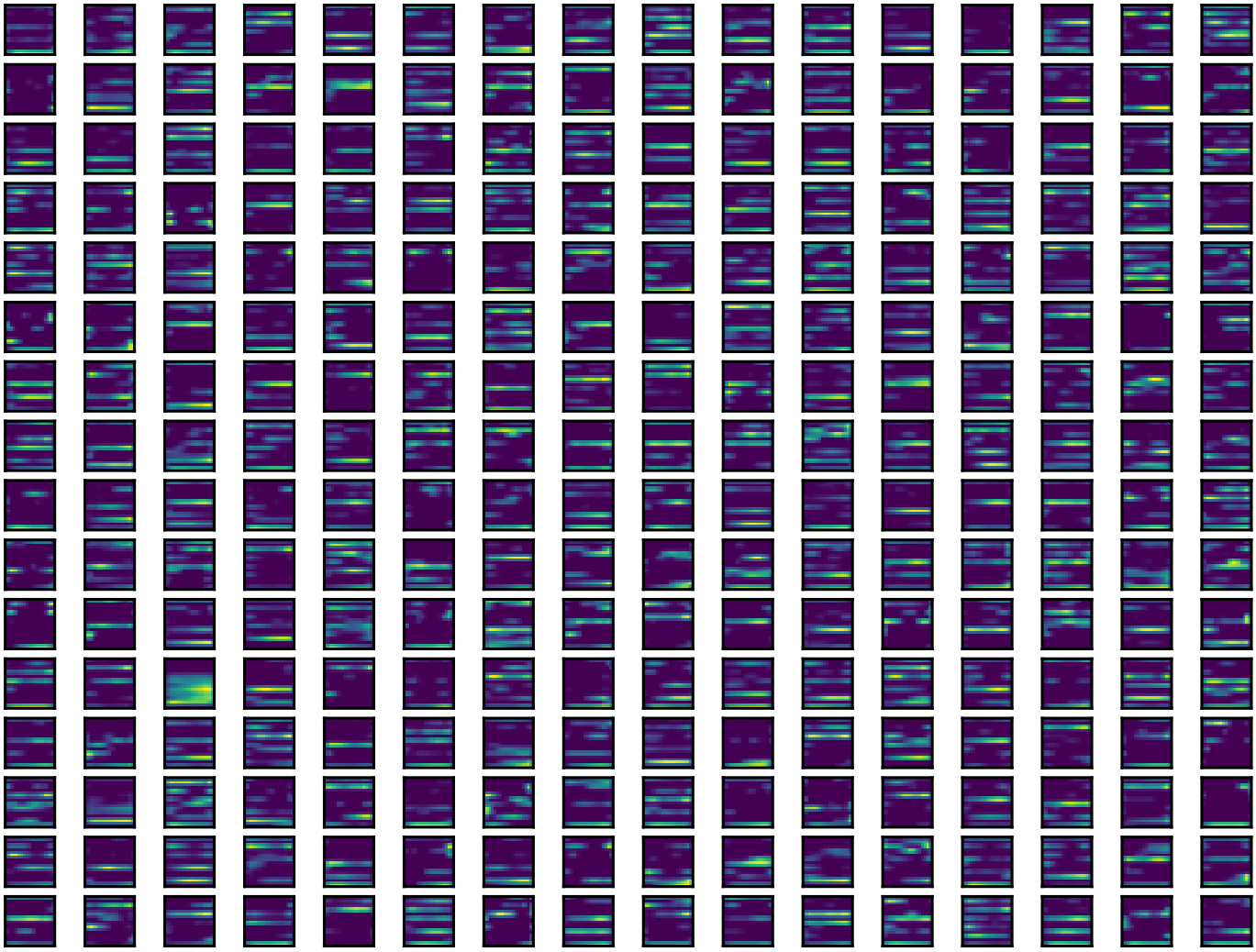


Figure 8. The 256 feature maps of the fifth and last convolutional layer of the trained architecture. Maps have been normalized in [0, 1] for visualization: lighter colors are close to 1, while darker colors are close to 0.

8. Conclusions

In this paper, we have investigated the effectiveness of a Convolutional Neural Network (CNN) fed by scalograms in the classification of audio signals acquired in real-world construction sites. Specifically, after splitting the recorded signals into smaller chunks, the scalogram (i.e., the squared magnitude of the Continuous Wavelet Transform) has been computed and used as input to a customized version of the well-known AlexNet. The customization takes into account the single channel of the scalogram input and the reduced number of output classes. Some experimental results and comparisons with other state-of-the-art approaches confirm the effectiveness of the proposed idea, showing an overall accuracy of 98.9%.

In future work, we will investigate the effect of choosing different types of wavelet functions and the idea of early data fusion, i.e., by joining the scalograms with other bi-dimensional representations, such as the spectrogram or similar ones, and providing this augmented representation as the input to a CNN.

Author Contributions: Conceptualization, M.S.; methodology, R.P. and Y.-C.L.; software, M.S. and R.P.; validation, M.S. and Y.-C.L.; formal analysis, Y.-C.L. and M.S.; investigation, M.S., Y.-C.L. and R.P.; data curation, Y.-C.L.; writing—original draft preparation, M.S.; writing—review and editing, Y.-C.L., R.P. and M.S.; visualization, M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Sapienza University of Rome grant numbers RM12117A39E2E9A7 and RM122180FB3CA3F2.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACSM	Automatic Construction Site Monitoring
AI	Artificial Intelligence
CNN	Convolutional Neural Network
CWT	Continuous Wavelet Transform
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
DRNN	Deep Recurrent Neural Network
ESC	Environmental Sound Classification
ESN	Echo State Network
GPS	Global Positioning System
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
RNN	Recurrent Neural Network
STFT	Short Time Fourier Transform
SVM	Support Vector Machine

References

1. Bansal, A.; Garg, N.K. Environmental Sound Classification: A descriptive review of the literature. *Intell. Syst. Appl.* **2022**, *16*, 200115. https://doi.org/10.1016/j.iswa.2022.200115.
2. Zaman, K.; Sah, M.; Direkoglu, C.; Unoki, M. A Survey of Audio Classification Using Deep Learning. *IEEE Access* **2023**, *11*, 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015.
3. Demir, F.; Abdullah, D.A.; Sengur, A. A New Deep CNN Model for Environmental Sound Classification. *IEEE Access* **2020**, *8*, 66529–66537. https://doi.org/10.1109/ACCESS.2020.2984903.
4. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP 2015), Boston, MA, USA, 17–20 September 2015; pp. 1–6. https://doi.org/10.1109/MLSP.2015.7324337.
5. Lee, Y.C.; Scarpiniti, M.; Uncini, A. Advanced Sound Classifiers and Performance Analyses for Accurate Audio-Based Construction Project Monitoring. *ASCE J. Comput. Civ. Eng.* **2020**, *34*, 1–11. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000911.
6. Xiong, W.; Xu, X.; Chen, L.; Yang, J. Sound-Based Construction Activity Monitoring with Deep Learning. *Buildings* **2022**, *12*, 1947. https://doi.org/10.3390/buildings12111947.
7. Kim, I.C.; Kim, Y.J.; Chin, S.Y. Sound Localization Framework for Construction Site Monitoring. *Appl. Sci.* **2022**, *12*, 783. https://doi.org/10.3390/app122110783.
8. Sanhudo, L.; Calvetti, D.; Martins, J.; Ramos, N.; Mêda, P.; Gonçalves, M.; Sousa, H. Activity classification using accelerometers and machine learning for complex construction worker activities. *J. Build. Eng.* **2021**, *35*, 102001. https://doi.org/10.1016/j.jobe.2020.102001.

9. Jungmann, M.; Ungureanu, L.; Hartmann, T.; Posada, H.; Chacon, R. Real-Time Activity Duration Extraction of Crane Works for Data-Driven Discrete Event Simulation. In Proceedings of the 2022 Winter Simulation Conference (WSC 2022), Singapore, 11–14 December 2022; pp. 2365–2376. https://doi.org/10.1109/WSC57314.2022.10015250.
10. Sherafat, B.; Ahn, C.R.; Akhavian, R.; Behzadan, A.H.; Golparvar-Fard, M.; Kim, H.; Lee, Y.C.; Rashidi, A.; Azar, E.R. Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review. *J. Constr. Eng. Manag.* **2020**, *146*, 03120002. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001843.
11. Rao, A.; Radanovic, M.; Liu, Y.; Hu, S.; Fang, Y.; Khoshelham, K.; Palaniswami, M.; Ngo, T. Real-time monitoring of construction sites: Sensors, methods, and applications. *Autom. Constr.* **2022**, *136*, 104099. https://doi.org/10.1016/j.autcon.2021.104099.
12. Zhou, Z.; Wei, L.; Yuan, J.; Cui, J.; Zhang, Z.; Zhuo, W.; Lin, D. Construction safety management in the data-rich era: A hybrid review based upon three perspectives of nature of dataset, machine learning approach, and research topic. *Adv. Eng. Inform.* **2023**, *58*, 102144. https://doi.org/10.1016/j.aei.2023.102144.
13. Navon, R.; Sacks, R. Assessing research issues in Automated Project Performance Control (APPC). *Autom. Constr.* **2007**, *16*, 474–484. https://doi.org/10.1016/j.autcon.2006.08.001.
14. Akbal, E.; Tuncer, T. A learning model for automated construction site monitoring using ambient sounds. *Autom. Constr.* **2022**, *134*, 104094. https://doi.org/10.1016/j.autcon.2021.104094.
15. Meng, Q.; Peng, Q.; Li, Z.; Hu, X. Big Data Technology in Construction Safety Management: Application Status, Trend and Challenge. *Buildings* **2022**, *12*, 533. https://doi.org/10.3390/buildings12050533.
16. Rashid, K.M.; Louis, J. Activity identification in modular construction using audio signals and machine learning. *Autom. Constr.* **2020**, *119*, 103361. https://doi.org/10.1016/j.autcon.2020.103361.
17. Jacobsen, E.; Teizer, J. Deep Learning in Construction: Review of Applications and Potential Avenues. *J. Comput. Civ. Eng.* **2022**, *36*, 1010. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001010.
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 18–22 June 2015; pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.
19. Wyse, L. Audio Spectrogram Representations for Processing with Convolutional Neural Networks. In Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN, Anchorage, AK, USA, 17–18 May 2017; pp. 37–41.
20. Dörfler, M.; Bammer, R.; Grill, T. Inside the spectrogram: Convolutional Neural Networks in audio processing. In Proceedings of the 2017 International Conference on Sampling Theory and Applications (SampTA), Bordeaux, France, 8–12 July 2017; pp. 152–155. https://doi.org/10.1109/SAMPTA.2017.8024472.
21. Traunmüller, H. Analytical expressions for the tonotopic sensory scale. *J. Acoust. Soc. Am.* **1990**, *88*, 97–100. https://doi.org/10.1121/1.399849.
22. Auger, F.; Flandrin, P.; Lin, Y.T.; McLaughlin, S.; Meignen, S.; Oberlin, T.; Wu, H.T. Time-Frequency Reassignment and Synchrosqueezing: An Overview. *IEEE Signal Process. Mag.* **2013**, *30*, 32–41. https://doi.org/10.1109/MSP.2013.2265316.
23. Mallat, S. *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed.; Academic Press: Cambridge, MA, USA, 2009.
24. Sacks, R.; Brilakis, I.; Pikas, E.; Xie, H.; Girolami, M. Construction with digital twin information systems. *Data-Centric Eng.* **2020**, *1*, e14. https://doi.org/10.1017/dce.2020.16.
25. Deng, R.; Li, C. Digital Intelligent Management Platform for High-Rise Building Construction Based on BIM Technology. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 1057–1067. https://doi.org/10.14569/IJACSA.2022.01312121.
26. Mansoor, A.; Liu, S.; Ali, G.; Bouferguene, A.; Al-Hussein, M. Scientometric analysis and critical review on the application of deep learning in the construction industry. *Can. J. Civ. Eng.* **2023**, *50*, 253–269. https://doi.org/10.1139/cjce-2022-0379.
27. Garcia, J.; Villavicencio, G.; Altimiras, F.; Crawford, B.; Soto, R.; Minatogawa, V.; Franco, M.; Martínez-Muñoz, D.; Yepes, V. Machine learning techniques applied to construction: A hybrid bibliometric analysis of advances and future directions. *Autom. Constr.* **2022**, *142*, 104532. https://doi.org/10.1016/j.autcon.2022.104532.
28. Cao, J.; Wang, W.; Wang, J.; Wang, R. Excavation Equipment Recognition Based on Novel Acoustic Statistical Features. *IEEE Trans. Cybern.* **2017**, *47*, 4392–4404. https://doi.org/10.1109/TCYB.2016.2609999.
29. Jeong, G.; Ahn, C.R.; Park, M. Constructing an Audio Dataset of Construction Equipment from Online Sources for Audio-Based Recognition. In Proceedings of the 2022 Winter Simulation Conference (WSC), Singapore, 11–14 December 2022; pp. 2354–2364. https://doi.org/10.1109/WSC57314.2022.10015388.
30. Wang, G.; Yu, Y.; Li, H. Automated activity recognition of construction workers using single in-pocket smartphone and machine learning methods. In *Proceedings of the IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2022; Volume 1101, p. 072008. https://doi.org/10.1088/1755-1315/1101/7/072008.
31. Maccagno, A.; Mastropietro, A.; Mazziotta, U.; Scarpiniti, M.; Lee, Y.C.; Uncini, A. A CNN Approach for Audio Classification in Construction Sites. In *Progresses in Artificial Intelligence and Neural Systems*; Esposito, A.; Faudez-Zanuy, M.; Morabito, F.C.; Pasero, E., Eds.; Springer: Singapore, 2021; Volume 184, pp. 371–381. https://doi.org/10.1007/978-981-15-5093-5_33.
32. Sherafat, B.; Rashidi, A.; Asgari, S. Sound-based multiple-equipment activity recognition using convolutional neural networks. *Autom. Constr.* **2022**, *135*, 104104. https://doi.org/10.1016/j.autcon.2021.104104.
33. Scarpiniti, M.; Colasante, F.; Di Tanna, S.; Ciancia, M.; Lee, Y.C.; Uncini, A. Deep Belief Network based audio classification for construction sites monitoring. *Expert Syst. Appl.* **2021**, *177*, 1–14. https://doi.org/10.1016/j.eswa.2021.114839.

34. Scarpiniti, M.; Bini, E.; Ferraro, M.; Giannetti, A.; Comminiello, D.; Lee, Y.C.; Uncini, A. Leaky Echo State Network for Audio Classification in Construction Sites. In *Applications of Artificial Intelligence and Neural Systems to Data Science*; Esposito, A.; Faudez-Zanuy, M.; Morabito, F.C.; Pasero, E., Eds.; Springer: Singapore, 2023; Volume 360. https://doi.org/10.1007/978-981-99-3592-5_18.
35. Kim, I.; Kim, Y.; Chin, S. Deep-Learning-Based Sound Classification Model for Concrete Pouring Work Monitoring at a Construction Site. *Appl. Sci.* **2023**, *13*, 4789. https://doi.org/10.3390/app13084789.
36. Peng, Z.; Kong, Q.; Yuan, C.; Li, R.; Chi, H.L. Development of acoustic denoising learning network for communication enhancement in construction sites. *Adv. Eng. Inform.* **2023**, *56*, 101981. https://doi.org/10.1016/j.aei.2023.101981.
37. Akbal, E.; Barua, P.D.; Dogan, S.; Tuncer, T.; Acharya, U.R. DesPatNet25: Data encryption standard cipher model for accurate automated construction site monitoring with sound signals. *Expert Syst. Appl.* **2022**, *193*, 116447. https://doi.org/10.1016/j.eswa.2021.116447.
38. Sherafat, B.; Rashidi, A.; Lee, Y.C.; Ahn, C.R. A Hybrid Kinematic-Acoustic System for Automated Activity Detection of Construction Equipment. *Sensors* **2019**, *19*, 4286. https://doi.org/10.3390/s19194286.
39. Kim, J.; Chi, S. Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. *Autom. Constr.* **2019**, *104*, 255–264. https://doi.org/10.1016/j.autcon.2019.03.025.
40. Soltani, M.M.; Zhu, Z.; Hammad, A. Framework for Location Data Fusion and Pose Estimation of Excavators Using Stereo Vision. *J. Comput. Civ. Eng.* **2018**, *32*, 04018045. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000783.
41. Jung, S.; Jeoung, J.; Lee, D.E.; Jang, H.; Hong, T. Visual–auditory learning network for construction equipment action detection. *Comput. Aided Civ. Infrastruct. Eng.* **2023**, *38*, 1916–1934. https://doi.org/10.1111/mice.12983.
42. Elelu, K.; Le, T.; Le, C. Collision Hazard Detection for Construction Worker Safety Using Audio Surveillance. *J. Constr. Eng. Manag.* **2023**, *149*. https://doi.org/10.1061/JCEMD4.COENG-12561.
43. Dang, K.; Elelu, K.; Le, T.; Le, C. Augmented Hearing of Auditory Safety Cues for Construction Workers: A Systematic Literature Review. *Sensors* **2022**, *22*, 9135. https://doi.org/10.3390/s22239135.
44. Chen, H.; Zhang, P.; Bai, H.; Yuan, Q.; Bao, X.; Yan, Y. Deep Convolutional Neural Network with Scalogram for Audio Scene Modeling. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3304–3308. https://doi.org/10.21437/Interspeech.2018-1524.
45. Faysal, A.; Ngui, W.K.; Lim, M.H.; Leong, M.S. Noise Eliminated Ensemble Empirical Mode Decomposition Scalogram Analysis for Rotating Machinery Fault Diagnosis. *Sensors* **2021**, *21*, 8114. https://doi.org/10.3390/s21238114.
46. Copiaco, A.; Ritz, C.; Fasciani, S.; Abdulaziz, N. Scalogram Neural Network Activations with Machine Learning for Domestic Multi-channel Audio Classification. In Proceedings of the 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Ajman, United Arab Emirates, 10–12 December 2019; pp. 1–6. https://doi.org/10.1109/ISSPIT47144.2019.9001814.
47. Ren, Z.; Qian, K.; Zhang, Z.; Pandit, V.; Baird, A.; Schuller, B. Deep Scalogram Representations for Acoustic Scene Classification. *IEEE/CAA J. Autom. Sin.* **2018**, *5*, 662–669. https://doi.org/10.1109/JAS.2018.7511066.
48. Flanagan, J.L. *Speech Analysis, Synthesis and Perception*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1972. https://doi.org/10.1007/978-3-662-00849-2.
49. Gupta, P.; Chodingala, P.K.; Patil, H.A. Morlet Wavelet-Based Voice Liveness Detection using Convolutional Neural Network. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022, pp. 100–104. https://doi.org/10.23919/EUSIPCO55093.2022.9909835.
50. Lin, J. Feature extraction of machine sound using wavelet and its application in fault diagnosis. *NDT E Int.* **2001**, *34*, 25–30. https://doi.org/10.1016/S0963-8695(00)00025-6.
51. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Stateline, NV, USA, 3–8 December 2012, pp. 1097–1105. https://doi.org/10.1145/3065386.
52. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015), San Diego, USA, 7–9 May 2015; pp. 1–15.
53. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63. https://doi.org/10.13140/RG.2.2.22227.99364/1.
54. Scarpiniti, M.; Comminiello, D.; Uncini, A.; Lee, Y.C. Deep recurrent neural networks for audio classification in construction sites. In Proceedings of the 28th European Signal Processing Conference (EUSIPCO 2020), Amsterdam, The Netherlands, 24–28 August 2020; pp. 810–814. https://doi.org/10.23919/Eusipco47968.2020.9287802.