informatics

MDPI

Article

Improving Smart Cities Safety Using Sound Events Detection Based on Deep Neural Network Algorithms

Giuseppe Ciaburro * and Gino Iannace

Dipartimento di Architettura e Disegno Industriale, Università degli Studi della Campania Luigi Vanvitelli, 81031 Aversa, Italy; gino.iannace@unicampania.it

* Correspondence: giuseppe.ciaburro@unicampania.it

Received: 30 May 2020; Accepted: 16 July 2020; Published: 20 July 2020

Abstract: In recent years, security in urban areas has gradually assumed a central position, focusing increasing attention on citizens, institutions and political forces. Security problems have a different nature—to name a few, we can think of the problems deriving from citizens' mobility, then move on to microcrime, and end up with the ever-present risk of terrorism. Equipping a smart city with an infrastructure of sensors capable of alerting security managers about a possible risk becomes crucial for the safety of citizens. The use of unmanned aerial vehicles (UAVs) to manage citizens' needs is now widespread, to highlight the possible risks to public safety. These risks were then increased using these devices to carry out terrorist attacks in various places around the world. Detecting the presence of drones is not a simple procedure given the small size and the presence of only rotating parts. This study presents the results of studies carried out on the detection of the presence of UAVs in outdoor/indoor urban sound environments. For the detection of UAVs, sensors capable of measuring the sound emitted by UAVs and algorithms based on deep neural networks capable of identifying their spectral signature that were used. The results obtained suggest the adoption of this methodology for improving the safety of smart cities.

Keywords: deep learning; machine learning; sounds classification; convolutional neural networks; pattern recognition; audio events detection; quadcopter UAV; acoustic measurements; acoustics features

1. Introduction

A smart city is based on an urban development model that integrates modern technologies in information and communication (ICT) for the management of a city's heritage as well as all its components [1]. The city represents the place of complexity and it is the result of a constant evolutionary process, in which a relevant place is assumed by the experimentation of new digital technology. The aim is to improve the quality of citizens' life with the support of the most modern technologies, considering the social, cultural, environmental and physical needs of a society [2].

Data management and visualization are key elements for achieving smartness and currently represent one of the main challenges for researchers. All countries are investing in the search for new methods to simulate data in real time and therefore in analyzing the impacts related to what-if scenarios [3].

A smart city can be designed through three main types of approach: urban planning, technological characteristics, marketing. From a strictly technological point of view, it provides for the use of sensors/actuators, with the capacity for self-management, configuration and autonomous optimization. A smart city also provides the ability to access a terminal from anywhere and always have a mobile Internet connection: All of this can be summarized as a Smart Ecosystem [4].

Smart cities represent a revolution with respect to the traditional approach to designing life places, but the evolution in the way of thinking about new environments cannot ignore the need to guarantee citizens' safety [5]. New technologies offer smart devices that make citizens' lives increasingly simplified. Smart locks guarantee us quick access to our resources, be it banking services, access to hotel rooms or much more simply the return home. These facilities introduce a security risk as hackers and experienced criminals can breach an infinite variety of systems. An integrated cyber defense system that guarantees the citizen a high degree of security in the management of the available services becomes therefore crucial for a smart city [6].

Then there are the security problems related to the operation of the numerous intelligent systems with which smart cities are equipped: robots for collecting garbage or for home deliveries, traffic lights with Internet of Things (IoT) traffic sensors, vehicles without driver. A smart city is managed by a myriad of sensors that produce a large amount of data, for the provision of advanced services intended for the management of increasingly efficient urban centers [7].

This technology introduces new levels of innovation that bring with them still unexplored security threats. Traditional management of these problems is destined to fail: For the security of smart cities, equally sophisticated technology is needed, which must come from artificial intelligence [8].

The advantage of using artificial intelligence and machine learning to support IT security lies in the unlimited amount of data that this technology can leverage and analyze. Thanks to the almost infinite amount of data related to harmful behaviors, machine learning can constantly feed its models to identify new dangers, thus protecting IT systems also from zero-day threats and variants of attacks. Expert systems based on artificial intelligence are not limited to identifying possible threats within the city network but are proposed as a tool to prevent any unauthorized intrusion, providing support for the protection of sensitive data. This result is obtained by identifying the behavior of the service-users: adoption of passwords and information exchanged. This creates a scenario of the situation and decides the necessary measures to ensure safety [9].

Machine learning based algorithms can extract knowledge from a database in a similar way to what is done by a human operator. This procedure occurs through the recognition of patterns that are difficult to identify even for an expert eye. In this way the data collected by the sensors can be automatically processed in search of potentially harmful or even illegal content. Machine learning can extract knowledge from data allowing not only to identify specific contents, but also to identify new trends that allow us to identify new scenarios and foresee potentially dangerous situations [10].

Related Work

Machine learning-based technologies have been used to address various sound problems [11–18]. Lostanlen et al. [19] used an algorithm based on convolutional neural networks to detect night bird avian flight calls. The use of bioacoustics sensors allowed the recording of the sounds generated by the wildlife causing the least disturbance to the species. The purpose of monitoring is the quantification of the population, but this objective requires the classification of the species using the detected acoustic signals. The classification procedure is made difficult by the complexity and variability of the environmental noise. To overcome this difficulty, the authors used per-channel energy normalization, then a context-adaptive neural network layer as the last layer of the network. These technologies have improved the classification process, reducing on one hand the temporal oversizing between audio segments, at sunrise and sunset, and on the other the spatial over-adaptation between the positions of the sensors.

Lim et al. [20] have proposed a procedure for the recognition of audio events using weakly labeled datasets. The authors used an algorithm based on convolutional and recurrent neural networks. To address the shortage of labeled data they applied the data augmentation algorithm. The results obtained with the application of this technology suggest the adoption for the detection of weakly labeled semi-supervised sound events.

Kong et al. [21] used a generative algorithm based on a convolutional neural network to develop a procedure for detecting anomalous sound events in public spaces. The modeling of sound events over time was performed thanks to the WaveNet [22], WaveNet is a convolutional network used in the processing of the acoustic signal, in multilanguage vocal synthesis. The authors used WaveNet to make predictions of the complex acoustic signals that cause classification errors. The system was tested by recording audio in a subway station and comparing the results with the results obtained by adopting other technologies such as long short-term memory network and auto-encoder. The results of the proposed methodology perform better.

Ozer et al. [23] used convolutional neural networks for the classification of sounds. Recognizing sound in complex sound environments is a very difficult task especially if it is performed automatically. The need for the development of automatic audio event recognition systems is reiterated by various sectors, but the complexity of the activity depends on the variability of the background noise levels that reduce its performance. The convolutional neural networks are effectively performing the recognition of images, so the authors extracted the spectrograms of the audio signals and used them as input of an algorithm for the classification of sound events based on this technology. The results returned very high performances in the classification of spectrograms.

Jung et al. [24] used a model based on convolutional neural networks for the classification of human activity. The authors have shown that this technology is able to recognize the activities performed daily by people in indoor environments. To do this, they used sound recordings indoors, taking advantage of the propagation sound characteristic, overcoming physical obstacles. In this way it was possible to classify not only the sounds generated directly by people, but also those generated by the objects they manipulate. The results suggest the adoption of this technology for the iconography of sounds through convolutional neural networks (CNN)-based models.

Cakır et al. [25] studied new classification techniques based on learning the characteristics for detecting sound events in living environments. The authors have shown that CNNs can learn the shift invariant filters that are essential for modeling sound events. The authors also applied recurrent neural networks (RNN) to model the long-term temporal characteristics of sound events. Finally, they combined the convolutional and recurrent levels in a single classifier. This classifier takes full advantage of the characteristics of both technologies, giving excellent results.

Unmanned aerial vehicles (UAVs) are a technology that has attracted much interest in recent years and whose use has grown, not only in the military, but also in the civilian field. In fact, these aircraft are much more versatile than traditional aircraft as they have lower operating and management costs, they can be equipped with various types of sensors, they can fly at low altitudes and, without requiring the presence of the human pilot on board, they can be used in operations that require reaching inaccessible or inaccessible areas or in all those missions characterized by a high danger to human life. Their versatility means that they are used to carry out numerous applications in very different contexts: for example, UAVs are increasingly used for remote sensing and for missions to support search and rescue activities in areas affected by natural disasters [26].

To date, UAVs are piloted via a radio control or via a control station: The first solution is used almost exclusively in amateur settings, while the second is used to carry out more complex applications. Through the control station it is possible to program the flight plan to be followed by the aircraft by specifying a list of waypoints—or a series of geographic coordinates—that the drone must reach in sequence [27].

UAVs represent a great opportunity for the evolution of smart cities. This technology lends itself to various applications ranging from infrastructure monitoring to home delivery of goods. However, the services that UAVs can provide to citizens are not limited to these examples. A crucial topic is covered using UAVs for citizens' safety: Several states in the world already use UAVs to monitor crowds during mass events. Unfortunately, the many useful uses are also accompanied by possible malicious activities that can be supported using UAVs. A UAV can travel on specific radio channels avoiding countermeasures based on electromagnetic disturbance, it can carry demanding payloads, such as

the transport of drugs and mobile phones even in prisons. With an appropriate sizing of the engines, UAVs can travel at considerable distances by providing a powerful air/ground communication system at the service of criminal associations. A further use for criminal activities then comes from terrorist organizations: The attack foiled the Venezuelan premier first and the attack on the oil extraction plant in Saudi Arabia are just examples [28].

Timely identification of the presence of UAVs in living environments becomes crucial for the safety of citizens. Unfortunately, the small size and the exclusive equipment of rotating parts make UAVs in fact difficult to trace, at least with traditional systems. A possible method of tracking the presence of UAVs makes use of the acoustic signal that they emit when they are in motion [29].

In this study, a new method is proposed for the detection of UAV in outdoor environments characterized by anthropogenic background noise. Data recorded by acoustic sensors are used to train a model based on convolutional neural networks. Outdoor recordings were made relating to two scenarios: anthropogenic noise and anthropogenic noise in the presence of a switched-on UAV. The collected data were divided into two sets: a first set was used for training the model and the next set was used for the evaluation of its performance. The model based on convolutional neural networks performed well, suggesting the use of this methodology to improve the safety of smart cities. The study is organized as follows: In Section 2, the materials and methods used are described in detail, first the instruments used for the recordings and the techniques used are presented, then a description of the UAVs used for the measurements is provided, then we analyze the feature extraction techniques used and finally convolutional neural networks are explored. In Section 3, the results obtained in this study are described providing an adequate discussion of the results achieved. In Section 4, conclusions are provided with some possible examples of use in real life and possible evolutions of the research.

2. Materials and Methods

The goal of this work is to develop a procedure to detect the presence of a UAV in a complex urban acoustic scenario. This procedure is proposed as a safety tool in a smart city using acoustic sensors that detect the acoustic signals of outdoor environments in real time. The detected acoustic signal is used as input by a system based on the use of convolutional neural networks which identifies the presence of UAVs. The detection of an event activates an alert which notifies the security forces of a possible case of danger and identifies its location. In Figure 1 flowchart with indications of all the steps of the elaborated methodology is proposed.

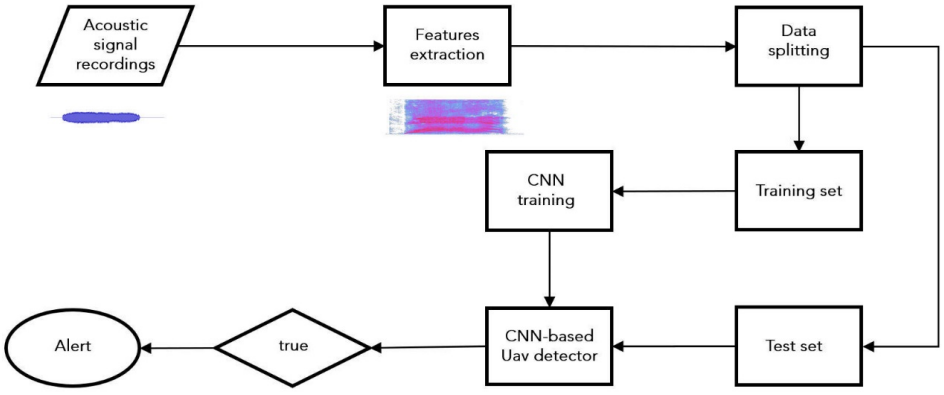

Figure 1. Flowchart of the unmanned aerial vehicle (UAV)-detection procedure in complex urban sound environments. The procedure involves the detection of signals from the environment with acoustic sensors, the extraction of features, the training of a model based on convolutional neural networks and the development of a convolutional neural networks (CNN)-based UAV detector. Subsequently the detected acoustic signals will be sent directly to the system which in case of detection of the UAV activates an alert to the security forces.

2.1. Audio Data Registrations

For the model training, audio recordings were made in front of an elementary school during the period when children leave school. At this stage there are many parents who await the exit of the children to bring them home. It is a scenario where people's safety is put to the test. Possible attackers could cause panic and create a situation of extreme danger also considering the presence of children. In this scenario, a control system based on audio recordings could significantly improve the security of smart cities.

The recordings relating to two scenarios were collected: ambient noise from the crowd and ambient noise from the crowd with the UAV turned on. To make measurements without disturbing the people present pending the end of the lessons, the UAV was anchored to a tripod. Furthermore, to simulate the movement of the UAV, its position with respect to the recording point has been changed: The UAV anchored to the tripod has been positioned in three points, respectively at 5 m, 10 m and 15 m with respect to the registration station (Figure 2).

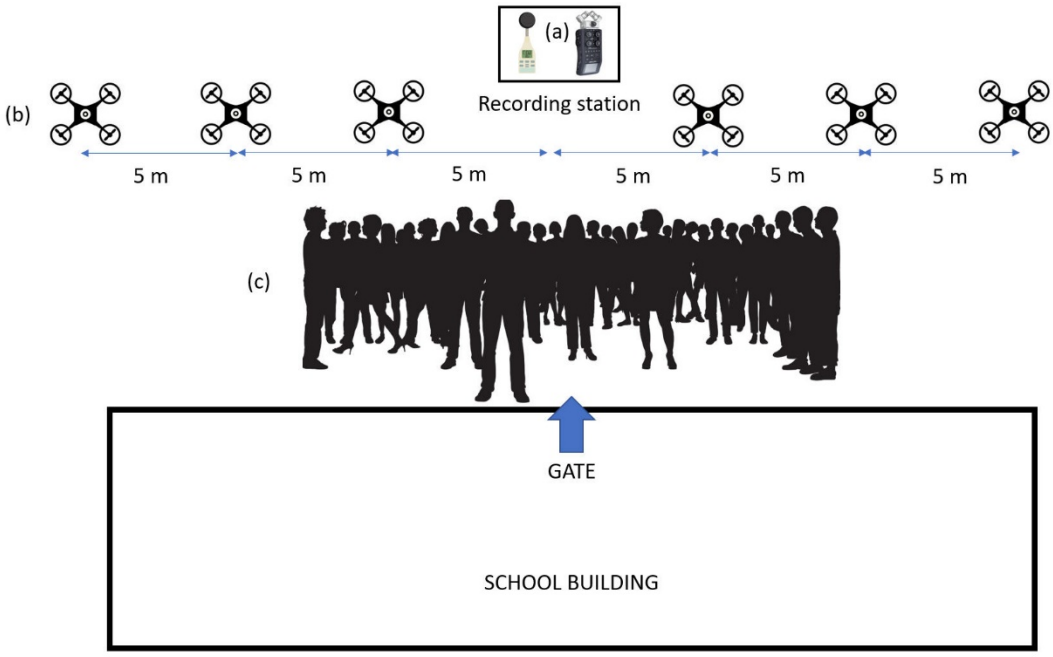

Figure 2. Recording scenario with (**a**) indication of the positions of the recording station, (**b**) the UAV anchored to the tripod and (**c**) the crowd waiting for the end of the day of lessons. The recordings relating to three scenarios were collected: ambient noise from the crowd, ambient noise with the UAV turned on, ambient noise from the crowd with the UAV turned on. To simulate the movement of the UAV, the UAV anchored to the tripod was positioned in three points, respectively at 5 m, 10 m and 15 m. These positions were identified on both sides with respect to the recording station. The registration station was located about 10 m from the edges of the crowd.

The recordings were made with a high-quality portable recorder Zoom H6 Handy Recorder (Zoom Corporation, Tokyo, Japan) with a X-Y microphone (Table 1).

Table 1. Recorder and microphone specifications.

Type	Specifications
Microphone	Sensitivity: −41 dB, 1 kHz at 1 Pa Input gain: −∞ to 46.5 dB Maximum sound pressure input: 136 dB SPL
Recording device	Sampling rate: 48 KHz Bit depth: 24 bit

The recordings will be used first to train the model based on convolutional neural networks, and then to evaluate the model's performance. For the characterization of the noise, measurements were performed simultaneously with a Class 1 Solo 01 dB sound level meter. In this way we will be able to analyze the differences between the sounds present in the environment, to identify features capable of discriminating between the different signals.

2.2. UAV Characterization

For the development of the UAV presence detection model, we used a radio-controlled quadcopter with a frequency of 2.4 GHz and a control distance of 80 m, made of acrylonitrile butadiene styrene (ABS). The UAV is equipped with a 6-axis gyroscope that manages its balance in flight. The UAv has the following dimensions: 310 × 310 × 120 mm. For propulsion, the UAV is equipped with four propellers with two blades. The four propellers move in the opposite direction alternately: two in a clockwise direction and the other two in an anticlockwise direction. The propeller represents the noise source of the device, therefore its shape and rotation speed as well as ensuring the flight performance of the UAV also determine its sound emission. The propeller is composed of a hub and blade: the hub takes on the load-bearing characteristic of the structure and transmits the rotation to the blade, in contrast to the blade determines the thrust that allows the UAV to fly.

To characterize the sound emissions of the UAV blade, measurements were carried out in an anechoic chamber measuring 4.40 m × 4.40 m × 4.50 m. The anechoic chamber is an acoustically absorbent room whose walls are made of fiberglass wedges. The purpose of these measurements is to identify the characteristic frequencies of the signal emitted by the UAV to highlight the same frequencies in the signal recorded by the acoustic sensors. Again, to simulate the flight conditions inside the anechoic chamber which has a limited volume, the UAV was anchored to a tripod. In this configuration, several measurement sessions were carried out, bringing the UAV to maximum speed conditions [30]. The measurements were performed with a Class 1 Solo 01 dB sound level meter, in accordance with the UNI EN ISO 3745:2012 standard [31].

2.3. Features Extraction

Before proceeding to the elaboration of the simulation model it is necessary to subject the audio recordings to elaboration to extract the features that will be used to identify the presence of the UAV. This procedure will then also be performed once the model is ready as the extracted features will be sent as input to the system.

Ambient noise is characterized by a wide range of frequencies with different levels, which determines its complexity. To analyze the sound content, a time domain analysis of the levels alone does not give us the necessary information. To characterize these complex sounds it is necessary to perform an analysis in the frequency domain [32]. Frequency analysis gives us the energy levels in the different frequency bands. In this way we can obtain the spectrum of the signal, that is, a representation on the Cartesian plane with the frequencies on the abscissa and the energy content of the sound on the ordinate.

In this work, we work two types of feature extractions: A first approach involves the use of a bank of band-pass filters that allows you to cut frequencies outside a certain range. We will use filters in one-third octave bands which return equivalent levels of linear sound pressure, weighted "Lin" at intervals equal to 125 ms [33]. These results will serve to characterize the environmental noise in order to identify specific trends capable of discriminating between the two scenarios that we want to identify: ambient noise from the crowd and ambient noise from the crowd with the UAV turned on.

We then decided to extract a spectrogram of the recorded sounds. A spectrogram is the representation of a sound using a diagram. The need to use a spectrogram is due to the limit of the spectrum which does not highlight the moments in which frequency variations occur. The spectrogram represents a signal that shows the relationship between three variables that characterize any sound: Frequency, Time and Intensity (color scale). Using a spectrogram, it is possible to understand how a

sound is made. In a spectrogram the abscissa represents time, the ordinate represents the frequency and intensity of the sound is represented by colors. In the color representation, dark colors represent low intensity sound, light colors for high intensity [34].

2.4. Convolutional Neural Network-Based Model

Convolutional neural networks are applied for management data characterized by a grid topology. CNN focuses local reports on adjacency structures that appear in the data. These processes of knowledge extraction take place through adaptive learning of patterns that start from the bottom to reach the highest level. CNNs are widely used in machine vision for object recognition. Convolutional neural networks are deep neural networks created to operate on grid inputs characterized by strong spatial dependencies in local regions. An example of grid input is a two-dimensional image: it is nothing more than an array of values between 0 and 255. Adjacent pixels present in an image are interlaced with each other and together they define a pattern, a texture, a contour, etc. CNNs associate these characteristics with values called weights, which will be similar for local regions with similar patterns [35].

The quality that distinguishes them is also the operation that gives them their name, that is, convolution. It is nothing more than the scalar product between matrices, specifically between a grid structure of weights and a similar structure from the input. Convolutional networks are deep networks in which the convolutional layer appears at least once, although the majority use far more than one. The input of a 2D CNN is an image, that is, an array of values of each single pixel occupying a precise position within the image. RGB (RED, GREEN, BLUE) images are described by a set of matrices of the intensities of the primary colors; the dimensions of an image are therefore not limited to height and width, but there is a third: depth, depth. The dimensions of the image, including the depth, will be given to the input layer, the very first layer of a CNN. The subsequent activation maps, that is, the inputs of the subsequent layers, also have a multidimensional structure and will be congruent in number with the independent properties relevant to the classification [36].

The convolutional neural networks operate on grid structures characterized by spatial relationships between pixels, inherited from one layer to the next through values that describe small local regions of the previous layer. The set of matrices of the hidden layers, the result of convolution or other operations, is called a feature map or activation map, the trainable parameters are tensors called filters or kernels [37].

A CNN is composed of sequences of the following layers:

- Convolutional layer;
- Activation layer;
- Pooling layer.

From a mathematical point of view, a CNN can be regarded as a neural network densely connected with the substantial difference that the first layers carry out a convolution operation. We indicate with the following equation the relationship between input and output of an intermediate layer:

$$Y^{[j]} = K^{[j]} * I^{[j-1]} + b^{[j]} \quad (1)$$

Here:

- $K^{[j]}$ is the kernel;
- $I^{[j-1]}$ is the input passed to the convolutional layer;
- $b^{[j]}$ is the bias.

Input passed to the current layer comes from a previous layer through the following equation:

$$I^{[j-1]} = g^{j-1}\left(Y^{[j-1]}\right) \quad (2)$$

Here:

- g^{j-1} is the activation function.

Let us analyze in detail the basic elements of an architecture based on CNN.

2.4.1. Convolutional Layer

Convolution is the fundamental operation of CNN. It places a filter (kernel) in every possible position of the image, covering it entirely and calculates the scalar product between the kernel itself and the corresponding matrix of the input volume, having equal dimensions. It is possible to view the convolution as a kernel overlay on the input image. A kernel is characterized by the following hyperparameters: height, width, depth and number [38].

Usually the kernels have a square shape and depth equal to that of the layer to which they are applied. The number of possible alignments between kernel and image defines the height and width of the next feature map. A distinction should be made between the depth of the kernel and the depth of the hidden layer/activation map: the first, is the same as the layer to which it is applied, the second derives instead from the number of kernels applied. The number of kernels is a hyperparameter defined based on the ability to distinguish even more complex shapes that you want to give to the network. Kernels are therefore the components to which the characteristics of image patterns will be associated.

The kernels of the first layers identify primitive forms, the later ones learn to distinguish ever larger and more complex forms. One property of the convolution is the translation equivalence: translated images are interpreted in the same way and the values of the activation map translate with the input values. This means that shapes generate similar feature maps, regardless of their location in the image.

One of the properties of the convolution is the following. A convolution on the q layer increases the receptive field of a feature from the q layer to the q + 1 layer. In other words, each value of the activation map of the next layer captures a wider spatial region than the previous one. Feature maps of the layers capture characteristic aspects of increasingly larger regions and this is the reason CNNs can be defined as deep: long sequences of blocks of layers are needed to study the whole image.

The convolution operation involves a contraction of the q layer with respect to q + 1 and a consequent loss of information. The problem can be curbed by using the so-called padding, a technique that involves adding pixels to the edges of the activation maps to maintain the spatial footprint. Obviously, in order not to alter the information, the pixels are assigned null values. The result is an increase in the size (height and width) of the input volume by an amount of which is reduced because of the convolution.

Since the product is zero, the external regions subject to padding do not contribute to the result of the scalar product. Instead, what happens is to allow the convolutional filter to override the edges of the layer and calculate the scalar product only for cells of values other than 0. This type of padding is defined as half-padding since about half of the filter goes beyond the edges, when placed at the ends. Half-padding is used to maintain the spatial footprint.

When padding is not used, we simply talk about valid-padding and in practice it does not give good results for the following reason: while with half-padding the cells at the edges contribute to the information, in the case of valid-padding, these they do not see the filter passage and are under-represented. Another form of padding is full padding, with which the filter is left completely out of the layer, occupying cells with only zeros. Doing so increases the spatial footprint of the layer, in the same way that valid padding reduces it.

A convolutional filter computes the scalar product in every single position of the input layer, but it is also possible to limit the computation to a lower number of positions, using the stride S. The convolution is then applied to the positions 1, S + 1, 2S + 1, etc., along both dimensions. It follows that the stride involves a reduction of the size by a factor of about 1/S and of the S^2 area. Generally,

values limited to 1 or 2 are used while with greater stride the aim is to reduce the memory request. Through the stride it is possible to capture complex patterns in large portions of the image, with results like those produced by max pooling.

In general, the dimensions of the input images are reduced to avoid complications in defining the hyperparameters. As far as convolution is concerned, the number of filters is usually a power of 2 to facilitate computation, the stride of 1 or 2, the size of the filter of 3 or 5. Small filters mean deeper and more performing networks.

Finally, each convolutional filter is associated with a bias given a filter p and the layer q, bias is indicated with b (p, q). The bias is a multiplication factor of the activation map and its presence increases the number of parameters of a unit. Like all other parameters, the bias value is defined by backpropagation during the training phase.

2.4.2. ReLU Activation Function

The nonlinear activation operation follows the convolution operation. For each layer, the activation function generates a layer of equal size with values limited by thresholds. As a simple one-to-one mapping of the activation values, the ReLU function does not alter the spatial import of the layer.

Activation takes place through mathematical functions. While in the past the hyperbolic tangent, the sigmoid function, softsign enjoyed widespread diffusion, they are now limited to non-deep networks and now replaced by the ReLU (Rectified Linear Unit) activation function [39]. The main reason is that in deep neural networks, the gradient of these three activation functions is canceled during backpropagation and prevents the algorithm from continuing with training. In addition, the ReLU function is computationally much more efficient.

$$f(x) = \begin{cases} x, \ if \ x \geq 0 \\ 0, \ if \ x < 0 \end{cases} \tag{3}$$

ReLU is in fact an effective and efficient negative value removal function. At the origin, the function does not approximate the identity and instead generates a high gradient preventing its disappearance.

2.4.3. Pooling Layer

The max pooling extracts the maximum value contained in P × P matrices of each activation map and produces another hidden layer of equal depth. In addition, in this case, as for the convolution, stride is used. Compared to convolution, pooling is carried out at the level of each activation map, therefore their number remains altered and the output is a layer of the same depth. A typical configuration is size 2 × 2 and stride 2: in this way there is no overlap between regions. The use of the stride in pooling is important for three reasons. The first is the reduction of the spatial footprint of the activation maps, the second is a certain degree of translation invariance and the third is an increase in the receptive field. It should be noted that only convolutional layers with strides greater than 1 can be used to reduce the spatial footprint. Despite this, it is still preferred to use max-pooling or some other variant given the degree of nonlinearity and invariance to the translation they introduce [40].

2.4.4. Fully Connected Layer

Connect each input feature with the corresponding output feature. It has the structure of a traditional feed-forward network and dramatically increases the number of parameters that can be trained due to the number of connections. For example, if two fully connected layers have 4096 units each, the number of connections will be greater than 16 million.

3. Results and Discussion

3.1. Characterization of the Acoustic Scenarios

In this work, we want to identify the presence of a UAV operating in a complex external environment. To start, a measurement of the sound emitted by the UAV in an anechoic chamber was performed [30]. This activity is aimed at characterizing the sound emission of the UAV to highlight the characteristics that can identify its presence in an external environment characterized by different sound sources. The measurements in the anechoic chamber simulate conditions of absence of external noise and in free field, providing an estimate of the emission spectrum of the UAV in the absence of unwanted sound reflections due to the surfaces of the environment and highlighting the tonal components present. Figure 3 shows the trend of the average sound spectrum in 1/3 octave frequency bands.

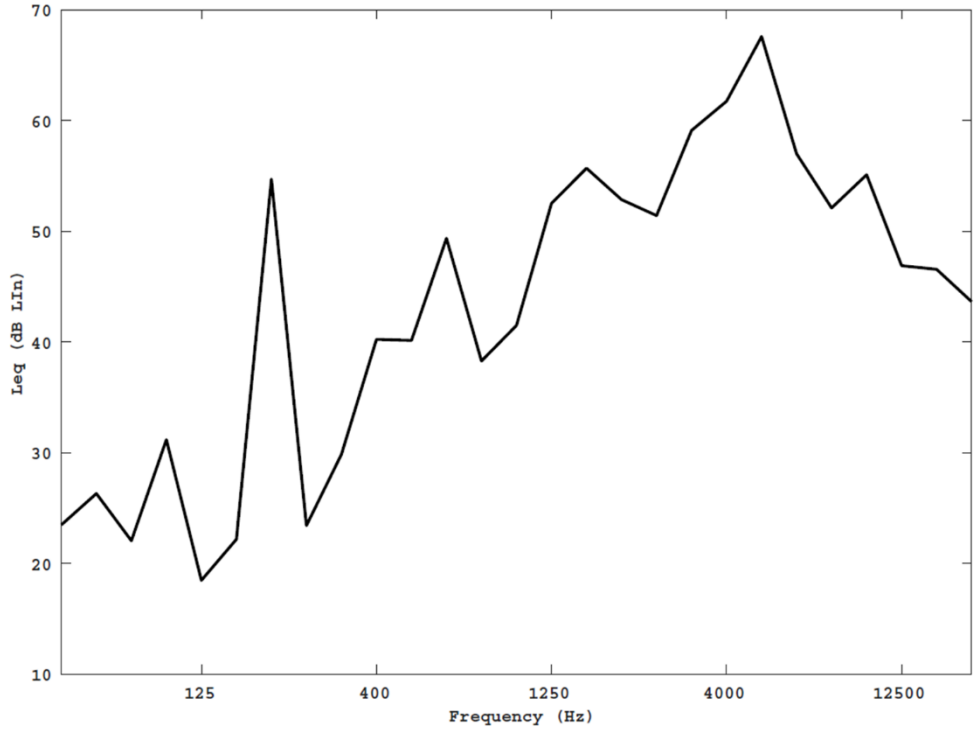

Figure 3. Average spectral levels in the one-third octave band between 50 Hz and 20 kHz (dB Lin) measured in the anechoic chamber. To simulate the flight conditions inside this limited volume chamber, the UAV was anchored to a tripod. In this configuration, several measurement sessions were carried out, bringing the UAV to maximum speed conditions. The measurements were performed with a Class 1 Solo 01-dB sound level meter, in accordance with the UNI EN ISO 3745:2012 standard [31].

The analysis of Figure 3 highlights three tonal components: two at low frequency and one at high frequency. At low frequencies two tonal components can be distinguished, respectively at 100 Hz and 200 Hz. These frequencies are characteristic of the rotation speed of the blades, in fact 100 Hz is the rotation speed of the shaft to which the blade is attached, that is, it represents the speed of rotation of the UAV engine at maximum speed. The 200 Hz frequency, on the other hand, is called blade pass frequency (BPF) and is obtained by multiplying the shaft rate by the number of propeller blades. The high frequency (5000 Hz) tonal component is due to the aerodynamics of the UAV, determined by the interaction between the rotation of the blades and the frame of the UAV which is located a short distance from them [41–44]. This instability produces a broadband noise between 2000 Hz and 20,000 Hz which presents a peak at 5000 Hz. This characteristic is fundamental for the identification of the UAV in scenarios where there are multiple sound sources, as it represents its spectral signature.

To characterize the sounds relating to the different sources, measurements were made for two scenarios: ambient noise from the crowd and ambient noise from the crowd with UAV turned on. The measurements were made using a class 1 sound level meter. The sound level meter was placed at a fixed distance from the edges of the crowd (10 m), while the tripod with the UAV was moved to different positions to simulate its movement. Sound recordings were made in the same location to be used later to train and then tested the UAV detector.

From a comparison between the measures it is possible to note that the trend of the environmental noise emitted during the two scenarios is comparable. It is not possible to notice a difference between the time histories that allow us to discriminate between the two operating conditions through a simple noise threshold. This confirms that environmental noise in such scenarios is so complex that it is not possible to distinguish between the different acoustic sources, at least in the time domain.

To extract further information, it is necessary to perform an analysis in the frequency domain. To do this, the average spectral levels in the 1/3 octave band between 50 Hz and 12,500 Hz (dB Lin) relating to the measurement sessions of the two scenarios were extracted (Figure 4). From the comparison between the ambient noise measured in the presence of the crowd with that in the presence of the crowd and UAV there are differences.

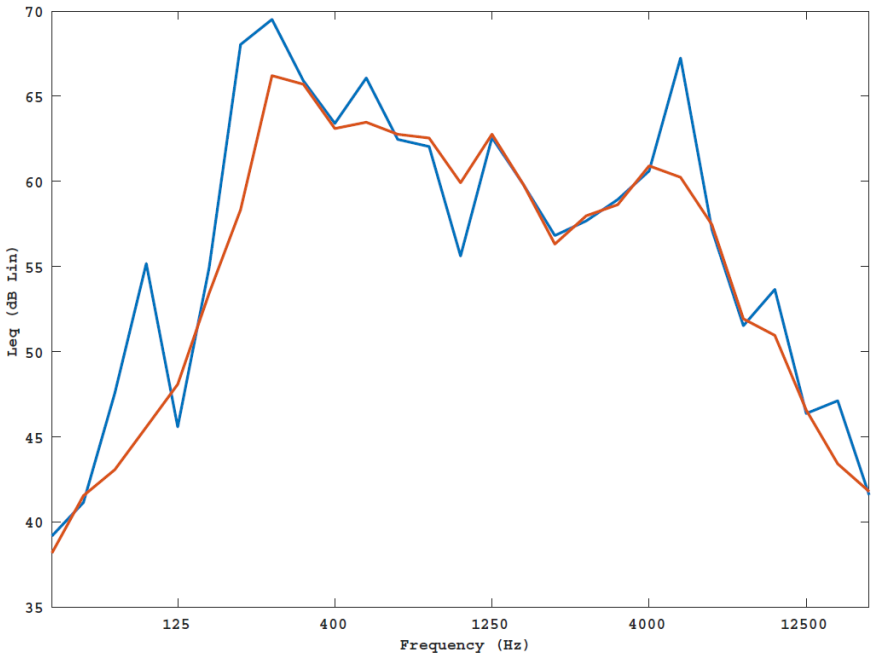

Figure 4. Average spectral levels in one-third octave band between 50 Hz and 12,500 Hz (dB Lin). The red curve represents the average spectral levels of the scenario with the crowd, while the blue curve represents the average spectral levels of the scenario with the crowd and the UAV.

The presence of the UAV adds the tonal components already highlighted in the measurements made in the anechoic chamber. At low frequency these are the components at 100 Hz and 200 Hz, while at high frequencies the component at 5000 Hz. The latter can help us to discriminate between the two acoustic scenarios.

3.2. Features Extraction

To train the algorithm based on the convolutional neural networks, we decided to use the spectrograms of the recordings made for the two scenarios considered. In a spectrogram the abscissa represents time, the ordinate represents the frequency and the intensity of the sound is represented by colors. In the representation of colors, dark colors represent low intensity sound, light colors for high intensity. In this way we can keep track of the time evolution of the frequency content of a signal.

They are essentially images that represent the frequency content of a sound in a time interval. We chose this solution given the predisposition of CNN to the treatment of images that represent the natural form of the inputs of this technology.

To extract the features, we used the sound recordings collected in the measurement sessions. Each recording was divided into pieces of 10 s duration to obtain an adequate number of samples to be used in the training and testing phase. In this way, about 100 samples were extracted equally distributed between the two identification classes (UAV, no-UAV). The images obtained were subsequently processed by making simple transformations without changing their structure. The images were subjected to random rotation, cutting and overturning. In this way we were able to increase the number of images to be used in the subsequent training and testing phases, going from 10 images to 1000. Of these 1000 samples, 70% was used in the training phase and the remaining 30% was reserved for the testing phase.

To understand how a spectrogram can allow the identification of the presence of the UAV in a complex urban scenario, we trace two spectrograms relating to two scenarios. The first scenario will show the spectrogram of the noise produced by the crowd in the presence of a UAV in operation, to do this we use a sample of the recorded sounds lasting just over 30 s. To highlight the representative frequencies of the event, we have restricted the frequencies range to the range 0–10 kHz. The second scenario will show the spectrogram of the noise produced by the crowd (Figure 5).

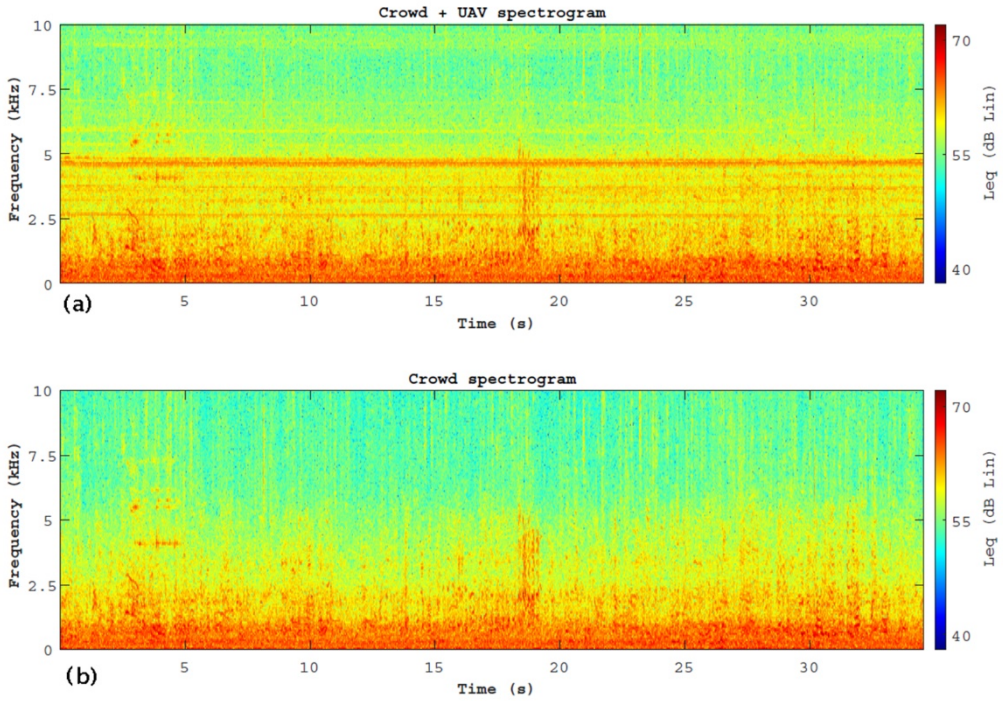

Figure 5. Spectrograms of two scenarios. (**a**) Noise produced by the crowd in the presence of a UAV in operation; (**b**) noise produced by the crowd. The samples refer to about 30 s of recording. To highlight the characteristic frequencies of the event, we have restricted the frequency range on the ordinate to the range 0–10 kHz. Recording in the presence of the UAV was performed by placing the UAV on the tripod at about 15 m from the registration station, as shown in Figure 1.

From the comparison between the spectrograms shown in Figure 5, the differences between the two scenarios are evident. Figure 5a representative of the crowd + UAV scenario shows the high frequency (5 kHz) tonal component which is not present in Figure 5b representative of the crowd scenario. This confirms that the feature chosen for identifying the presence of the UAV is suitable for this task. Furthermore, this high frequency component had already been identified as the spectral signature of the UAV in the measurements carried out in the anechoic chamber and shown in Figure 3. It is a

typical operating characteristic of the analyzed UAV, but other UAVs even if at different frequencies always have a tonal component at high frequencies. This means that the shape of the spectrogram will remain substantially unchanged with a slight shift of the tonal component at different frequencies, even if always localized at high frequencies.

3.3. UAV Detector Based on Convolutional Neural Network

Table 2 shows the convolutional neural network architecture used in this study. The architecture has three convolutional hidden layers with an average pooling layer and a ReLU activation function. Following is a layer of flatten which allows us to flatten the feature map in a single column. A fully connected layer is then inserted to connect each input function with the corresponding output function. Finally, a dense layer is inserted to apply the SoftMax function to return the probability that an image belongs to one of the two classes (UAV, no-UAV) [45].

Table 2. CNN architecture.

Layer Name	Layer Description	Output Shape
Input layer	Input	$(64 \times 64 \times 3)$
First hidden layer	2D convolutional layer	$(31 \times 31 \times 32)$
	Average pooling operation for spatial data	$(15 \times 15 \times 32)$
	Rectified linear unit activation	$(15 \times 15 \times 32)$
Second hidden layer	2D convolutional layer	$(15 \times 15 \times 64)$
	Average pooling operation for spatial data	$(7 \times 7 \times 64)$
	Rectified linear unit activation	$(7 \times 7 \times 64)$
Third hidden layer	2D convolutional layer	$(7 \times 7 \times 64)$
	Average pooling operation for spatial data	$(3 \times 3 \times 64)$
	Rectified linear unit activation	$(3 \times 3 \times 64)$
Flatten layer	Flatten operation	(576)
	dropout	(576)
Fully connected layer	Densely connected NN layer	(64)
	Rectified linear unit activation	(64)
	dropout	(64)
Output layer	Densely connected NN layer	(2)
	Softmax activation	(2)

Each layer processes the input information and outputs it to the next layer, thus allowing increasing complex elaborations. The first layers perform low-level operations on the input data, and the last layers instead operate with semantic features, up to the last layer, whose structure differs according to the problem to be treated. For classification problems, the last layer has several neurons equal to the number of classes to be distinguished, with SoftMax activation function to have the probabilities of belonging to the various classes at the output.

After training the network with 70% of the available data, the model was tested with the remaining 30% of the data, making sure that the data used in the test phase are data never seen before by the model: In this way problems are avoided overfitting. Accuracy was used to evaluate the model's performance: Accuracy is the measure of how good our model is. It is expected to be closer to one if our model is performing well. The model based on convolutional neural networks returned an accuracy of 0.91, demonstrating the validity of the procedure for identifying a UAV in a complex acoustic scenario.

Accuracy returns the degree of concordance between a measured value and a true value, the more accurate a result the more error-free it is. The accuracy of a forecast therefore indicates how close the expected value of a quantity is to the real value of that quantity. In our case the real value is available as we have measured the accuracy of the model on the test data that has been appropriately labeled. A classification model is a mathematical function that univocally determines the class to which a statistical unit belongs, based on the values observed for the variables of interest. Its predictive accuracy depends on the ability to correctly classify new units, whatever the class they come from.

A result such as the one obtained (0.91) tells us that the model can correctly classify 91 cases out of the 100 that have been presented to it. Furthermore, it should be noted that this performance was obtained on a sample equally represented in the two classes.

To confirm the good performance of the model adopted, a comparison was made between the performance of the algorithm used in this work and that obtained from other state of art works. Hershey et al. [46] used CNN to classify the soundtracks of a video dataset. The authors compared the results obtained with the most common CNN-based architectures. The results show an AUC (Area under the ROC curve) ranging from 0.87 to 0.93. Receiver Operating Characteristic (ROC) curves are graphic patterns for a binary classifier. This curve plots two parameters: True Positive Rate versus False Positive Rate. The performance of the models grows as the number of examples labeled in the training set increases, to confirm the importance of the data in the correct classification of the audio sources. Lee et al. [47] have used CNN for the classification of raw waveform-based audio. The authors elaborated two models of deep convolutional neural networks that use raw waveforms as input and exploit filters with reduced granularity. The first is based on convolution layers and pooling layers. The second is an improved model that also has residual connections, compression modules and multilevel concatenation. The models returned an accuracy ranging from 0.84 to 0.88. Lim et al. [48] used CNN to classify audio events. The authors exploited the features of audio sound as an input image of CNN: They extracted the characteristics of the MEL scale filter bank from each frame and the results obtained from 40 consecutive frames were concatenated and provided as input image. The MEL scale is a scale for perceiving the pitch of a sound. In this way they obtained an accuracy of 0.82 in the classification of audio events.

4. Conclusions

The smart cities security is a topic of crucial importance for the evolution of this architecture. Security does not only concern the data that are collected by the sensors and then processed, but also concerns the development of new technologies which, by exploiting such data, can guarantee new levels of citizens' security. This work proposes a new system for detecting the presence of UAVs in a complex urban scenario. Sounds are detected by acoustic sensors and transmitted to a UAV detector based on artificial intelligence. The expert system uses an algorithm based on convolutional neural networks to recognize the presence of UAV in a complex sound in which there are multiple sound sources. First the spectrograms are extracted from the detected sounds, which are then sent to CNN which recognizes the patterns by returning a response.

From the experimental results, we have the following conclusions:

1. The characterization of the acoustic emission of the UAV highlighted three tonal components: two at low frequency and one at high frequency;
2. From a comparison between the measures of the environmental noise emitted during the two scenarios (UAV, NoUAV) it is comparable. This confirms that the ambient noise in such scenarios is so complex that it is not possible to distinguish between the different acoustic sources, at least in the time domain;
3. The comparison between the frequency spectra of the two scenarios has shown that the 5000 Hz tonal component is a descriptor capable of discriminating between the two scenarios. This result is evident in the spectrograms of the measures;
4. A classification system based on CNN has proven capable of identifying the presence of UAVs with an accuracy of 0.91.

Modern security systems are essentially based on the use of images shot by video cameras for real-time monitoring of the activities taking place in some areas. In this study, the sounds detected by microphones are used by a system based on artificial intelligence to automatically recognize dangerous situations. Sounds that identify dangerous situations are detected to trigger an automatic alert that draws surveillance attention to that area. CNNs require calculations which are expensive since it is an

image processing process. This problem can be overcome with better hardware processing that takes advantage of the graphics processing units (GPU). The recognition system of a dangerous situation proposed in this work requires the intervention of a surveillance operator to verify the alarm signal, therefore the procedure is not fully automated. The latter limit can be overcome by integrating the recognition of the images detected by the video cameras into the proposed classification system. Finally, the procedure used in this study can be extended to other sound sources to identify specific sounds in work or social life contexts, in which the identification of possible risks for users becomes difficult with traditional technologies.

Author Contributions: Conceptualization, G.C. and G.I.; methodology, G.C.; investigation, G.C.; measurements G.C. and G.I.; software G.C.; post processing data, G.C. and G.I.; data curation, G.C. and G.I.; writing—original draft preparation, G.C. and G.I.; writing—review and editing, G.C. and G.I.; visualization, G.C.; supervision, G.C. and G.I.; references study, G.C. and G.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Gaur, A.; Scotney, B.; Parr, G.; McClean, S. Smart City Architecture and its Applications Based on IoT. *Procedia Comput. Sci.* **2015**, *52*, 1089–1094. [CrossRef]
2. Wenge, R.; Zhang, X.; Dave, C.; Chao, L.; Hao, S. Smart city architecture: A technology guide for implementation and design challenges. *China Commun.* **2014**, *11*, 56–69. [CrossRef]
3. Bianchini, D.; Ávila, I. Smart Cities and Their Smart Decisions: Ethical Considerations. *IEEE Technol. Soc. Mag.* **2014**, *33*, 34–40. [CrossRef]
4. Balakrishna, C. Enabling technologies for smart city services and applications. In Proceedings of the Sixth International Conference on Next Generation Mobile Applications, Services and Technologies, Paris, France, 12–14 September 2012; Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2012; pp. 223–227.
5. Elmaghraby, A.; Losavio, M.M. Cyber security challenges in Smart Cities: Safety, security and privacy. *J. Adv. Res.* **2014**, *5*, 491–497. [CrossRef] [PubMed]
6. Jabbar, M.A.; Samreen, S.; Aluvalu, R.; Reddy, K.K. Cyber Physical Systems for Smart Cities Development. *Int. J. Eng. Technol.* **2018**, *7*, 36–38. [CrossRef]
7. Hammoudeh, M.; Arioua, M. Sensors and Actuators in Smart Cities. *J. Sens. Actuator Netw.* **2018**, *7*, 8. [CrossRef]
8. Batty, M. Artificial intelligence and smart cities. *Environ. Plan. B Urban Anal. City Sci.* **2018**, *45*, 3–6. [CrossRef]
9. Ismagilova, E.; Hughes, D.L.; Dwivedi, Y.K.; Raman, K.R. Smart cities: Advances in research—An information systems perspective. *Int. J. Inf. Manag.* **2019**, *47*, 88–100. [CrossRef]
10. Din, I.U.; Guizani, M.; Rodrigues, J.J.P.C.; Hassan, S.; Korotaev, V. Machine learning in the Internet of Things: Designed techniques for smart cities. *Futur. Gener. Comput. Syst.* **2019**, *100*, 826–843. [CrossRef]
11. Gu, G.H.; Noh, J.; Kim, I.; Jung, Y. Machine learning for renewable energy materials. *J. Mater. Chem. A* **2019**, *7*, 17096–17117. [CrossRef]
12. Nichols, J.A.; Chan, H.W.H.; Ab Baker, M. Machine learning: Applications of artificial intelligence to imaging and diagnosis. *Biophys. Rev.* **2018**, *11*, 111–118. [CrossRef] [PubMed]
13. Iannace, G.; Ciaburro, G.; Trematerra, A. Heating, Ventilation, and Air Conditioning (HVAC) Noise Detection in Open-Plan Offices Using Recursive Partitioning. *Buildings* **2018**, *8*, 169. [CrossRef]
14. McCoy, J.; Auret, L. Machine learning applications in minerals processing: A review. *Miner. Eng.* **2019**, *132*, 95–109. [CrossRef]
15. Iannace, G.; Ciaburro, G.; Trematerra, A. Wind Turbine Noise Prediction Using Random Forest Regression. *Machines* **2019**, *7*, 69. [CrossRef]
16. Sun, Y.; Peng, M.; Zhou, Y.; Huang, Y.; Mao, S. Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3072–3108. [CrossRef]

17. Iannace, G.; Ciaburro, G.; Trematerra, A. Modelling sound absorption properties of broom fibers using artificial neural networks. *Appl. Acoust.* **2020**, *163*, 107239. [CrossRef]
18. Rutledge, R.B.; Chekroud, A.M.; Huys, Q.J. Machine learning and big data in psychiatry: Toward clinical applications. *Curr. Opin. Neurobiol.* **2019**, *55*, 152–159. [CrossRef]
19. Lostanlen, V.; Salamon, J.; Farnsworth, A.; Kelling, S.; Bello, J.P. Robust sound event detection in bioacoustic sensor networks. *PLoS ONE* **2019**, *14*, e0214168. [CrossRef]
20. Lim, W.; Suh, S.; Jeong, Y. Weakly Labeled Semi Supervised Sound Event Detection Using CRNN with Inception Module. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, Surrey, UK, 19–20 November 2018; pp. 74–77.
21. Kong, Q.; Xu, Y.; Sobieraj, I.; Wang, W.; Plumbley, M.D. Sound Event Detection and Time–Frequency Segmentation from Weakly Labelled Data. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 777–787. [CrossRef]
22. Oord, A.V.D.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
23. Ozer, Z.; Findik, O.; Ozer, I. Noise robust sound event classification with convolutional neural network. *Neurocomputing* **2018**, *272*, 505–512. [CrossRef]
24. Jung, M.; Chi, S. Human activity classification based on sound recognition and residual convolutional neural network. *Autom. Constr.* **2020**, *114*, 103177. [CrossRef]
25. Cakır, E.; Virtanen, T. Convolutional recurrent neural networks for rare sound event detection. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, New York, NY, USA, 35–26 October 2019.
26. Kille, T.; Bates, P.R.; Lee, S.Y. (Eds.) *Unmanned Aerial Vehicles in Civilian Logistics and Supply Chain Management*; IGI Global: Hersley, PA, USA, 2019.
27. Yu, K. Research on the Improvement of Civil Unmanned Aerial Vehicles Flight Control System. In *Advances in Intelligent Systems and Computing*; Springer Science and Business Media LLC: Singapore, 2018; pp. 375–384.
28. Oh, H. Countermeasure of Uumanned Aerial Vehicle (UAV) against terrorist's attacks in South Korea for the public crowded places. *J. Soc. Disaster Inf.* **2019**, *15*, 49–66.
29. Kartashov, V.M.; Oleynikov, V.N.; Sheyko, S.A.; Babkin, S.I.; Korytsev, I.V.; Zubkov, O.V. PECULIARITIES OF SMALL UNMANNED AERIAL VEHICLES DETECTION AND RECOGNITION. *Telecommun. Radio Eng.* **2019**, *78*. [CrossRef]
30. Iannace, G.; Ciaburro, G.; Trematerra, A. Fault Diagnosis for UAV Blades Using Artificial Neural Network. *Robotics* **2019**, *8*, 59. [CrossRef]
31. International Organization for Standardization (ISO) 3745: 2012. Acoustics.Determination of Sound Power Levels of Noise Sources Using Sound Pressure Precision Methods for Anechoic and Hemi-Anechoic Rooms. Available online: https://www.iso.org/standard/45362.html (accessed on 28 May 2020).
32. Gröchenig, K. *Foundations of Time-Frequency Analysis*; Springer: Berlin/Heidelberg, Germany, 2013.
33. Veggeberg, K. Octave analysis explored: A tutorial. *Eval. Eng.* **2008**, *47*, 40–44.
34. Fulop, S.A. *Speech Spectrum Analysis*; Springer: Berlin/Heidelberg, Germany, 2011.
35. Aghdam, H.H.; Heravi, E.J. *Guide to Convolutional Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10, p. 978.
36. Aggarwal, C.C. *Neural Networks and Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10, p. 978.
37. Kiranyaz, S.; Ince, T.; Abdeljaber, O.; Avci, O.; Gabbouj, M. 1-d Convolutional Neural Networks for Signal Processing Applications. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2019; pp. 8360–8364.
38. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
39. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 807–814.
40. Scherer, D.; Müller, A.; Behnke, S. Evaluation of pooling operations in convolutional architectures for object recognition. In Proceedings of the International Conference on Artificial Neural Networks, Thessaloniki, Greece, 15–18 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 92–101.

41. Kurtz, D.W.; Marte, J.E. *A Review of Aerodynamic Noise from Propellers, Rotors, and Lift Fans*; Jet Propulsion Laboratory, California Institute of Technology: Pasadena, CA, USA, 1970; pp. 32–1462.
42. Zawodny, N.S.; Boyd, D.D. Investigation of Rotor–Airframe Interaction Noise Associated with Small-Scale Rotary-Wing Unmanned Aircraft Systems. *J. Am. Helicopter Soc.* **2020**, *65*, 1–17. [CrossRef]
43. Regier, A.A.; Hubbard, H.H. Status of research on propeller noise and its reduction. *J. Acoust. Soc. Am.* **1953**, *25*, 395–404. [CrossRef]
44. Hubbard, H.H. *Aeroacoustics of Flight Vehicles: Theory and Practice. Volume 1: Noise Sources; No. NASA-L-16926-VOL-1*; National Aeronautics and Space Admin Langley Research Center: Hampton, VA, USA, 1991.
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
46. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017; Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2017; pp. 131–135.
47. Lee, J.; Kim, T.; Park, J.; Nam, J. Raw waveform-based audio classification using sample-level CNN architectures. *arXiv* **2017**, arXiv:1712.00866.
48. Lim, M.; Lee, D.; Park, H.; Kang, Y.; Oh, J.; Park, J.-S.; Jang, G.-J.; Kim, J.-H. Convolutional Neural Network based Audio Event Classification. *KSII Trans. Internet Inf. Syst.* **2018**, *12*. [CrossRef]