applied sciences

MDPI

Article

Ambient Sound Recognition of Daily Events by Means of Convolutional Neural Networks and Fuzzy Temporal Restrictions

Aurora Polo-Rodriguez [1,*], Jose Manuel Vilchez Chiachio [1], Cristiano Paggetti [2] and Javier Medina-Quero [1]

[1] Department of Computer Science, Campus Las Lagunillas, 23071 Jaén, Spain; jmvc0010@red.ujaen.es (J.M.V.C.); jmquero@red.ujaen.es (J.M.-Q.)
[2] I + Srl, Piazza G.Puccini, 26, 50144 Firenze, Italy; c.paggetti@i-piu.it
* Correspondence: apolo@ujaen.es; Tel.: +34-953-21-2802

Abstract: The use of multimodal sensors to describe activities of daily living in a noninvasive way is a promising research field in continuous development. In this work, we propose the use of ambient audio sensors to recognise events which are generated from the activities of daily living carried out by the inhabitants of a home. An edge–fog computing approach is proposed to integrate the recognition of audio events with smart boards where the data are collected. To this end, we compiled a balanced dataset which was collected and labelled in controlled conditions. A spectral representation of sounds was computed using convolutional network inputs to recognise ambient sounds with encouraging results. Next, fuzzy processing of audio event streams was included in the IoT boards by means of temporal restrictions defined by protoforms to filter the raw audio event recognition, which are key in removing false positives in real-time event recognition.

Keywords: activity recognition; audio recognition; fuzzy protoforms

check for updates

Citation: Polo-Rodriguez, A.; Vilchez Chiachio, J.M.; Paggetti, C.; Medina-Quero, J. Ambient Sound Recognition of Daily Events by Means of Convolutional Neural Networks and Fuzzy Temporal Restrictions. *Appl. Sci.* **2021**, *11*, 6978. http://doi.org/10.3390/app11156978

Academic Editors: Sang-Hyun Lee and Seongsoo Cho

Received: 30 April 2021
Accepted: 23 July 2021
Published: 29 July 2021

1. Introduction

Activity recognition (AR) has become an active research topic [1] focused on detecting human behaviours in smart environments [2]. Sensing human activity has been adopted in smart homes [3] with the aim of improving quality of life, allowing people to stay independent in their own homes for as long as possible [4].

In initial approaches, there was a predominance of binary sensors used to describe daily human activities within smart environments in a noninvasive manner. Next, a new generation of devices emerged to integrate a richer perspective in sensing smart objects and people's activities. Among them, the following types of sensors stand out: (i) wearable devices, which have been used to analyse activities and gestures[5]; (ii) location devices, which at present reach extremely high accuracy in indoor contexts [6]; (iii) vision sensors (visible-light or thermal-infrared sensors) in video and image sequences [7]; (iv) audio sensors [8] that recognise events based on audio information. This has been followed by a new trend of multimodal sensors that has enabled the use of general-purpose sensing technologies to monitor activities.

AR approaches are mainly grouped into two categories: knowledge-driven approaches [9] and data-driven approaches [10]. A number of previous AR studies have focused on classifying activities where the beginning and end of the activities, and therefore, the key features are known beforehand, which is referred to as explicit segmentation [11] or offline evaluation, as they do not provide real-time capabilities in AR. However, including real-time capabilities is a key requirement in AR in order to provide responses to real-world conditions [12], enabling *adequate assistance services*. In real-time AR, where the beginning and end of the events are unknown, approaches based on sliding windows to segment the data stream are required [11]. In addition, in the context of multimodal sensors, the use of deep learning models has shown promising performance in processing multimedia data [12].

In this work, we focus on the recognition of daily events by means of ambient sound devices and using deep learning models integrated with smart boards. The contribution of this work can be summarised as follows:

- Collecting a dataset of audio samples of events related to activities of daily living which are generated within indoor spaces;
- Integrating a fog–edge architecture with the IoT boards where the audio samples are collected to provide real-time recognition of audio events;
- Evaluating the performance of deep learning models for offline and real-time recognition of ambient daily living events in naturalistic conditions;
- A straightforward fuzzy processing of audio event streams is described by means of temporal restrictions which are modeled on linguistic protoforms to improve the audio recognition.

The remainder of the paper is organised as follows: In Section 1, we review related works and methodologies; in Section 2, we describe the devices, architecture, and methodology of the approach; in Section 3, we present the results of a case study of event recognition; in Section 4, we detail our conclusions and ongoing work.

Related Works

The integration of technology into smart environments to support our daily lives in an immersive and ubiquitous way was introduced by ubiquitous computing as *the age of calm technology, when technology recedes into the background of our lives* [13]. From this visionary perspective at the beginning of the 1990s to our present Internet of Things, two key characteristics have been exploited over the last 30 years: (i) immersiveness or low invasiveness of integrated devices (both on our bodies and in our environment) and (ii) smart connected devices which provide interpretable outcomes from the information collected by sensors.

As described above, ambient binary sensors have been proposed to describe daily activities in indoor spaces [14] with the goal of deploying immersive sensors, providing encouraging results with accurately labelled datasets [15] under data-driven approaches [10]. Nowadays, the burgeoning growth of devices is promoting multimodal sensors which typically integrate video, audio, and wearable sensors [16], and other IoT devices with increasing high-capacity computing. The new trends are converging toward synthetic sensors [17], which are deployed to sense everything in a given room, enabling the use of general-purpose sensing technologies in order to monitor activities by means of sensor fusion. In this context, audio processing by smart microphones for the labelling of audible events is opening up a promising research field within AR [8].

On the architecture of components for learning and communication of devices, the paradigms of edge computing [18] or fog computing [19] have located the data and services within the devices where sensors are integrated, *providing virtualised resources and engaged location-based services at the edge of the mobile networks* under a new perspective of the Internet of Things (IoT) to develop collaborative smart objects which *interact with each other and cooperate with their neighbours to reach common goals* [20].

In the machine learning models for AR, describing sensor information under data-driven approaches has depended on the type of sensors, whether inertial [5] or binary sensors [15], where integrated methodologies to exploit spatial–temporal features have been proposed [21]. Additionally, deep learning (DL) has also been shown as a suitable approach in AR to discover and extract features from sensors [22]. DL is related to multimodal sensor recognition, such as vision and audio, where obtaining hierarchical features to reduce complexity is key. Regarding vision sensors, the use of thermal vision is proposed to guarantee privacy while preventing dangers such as falling by means of convolutional neural networks (CNNs) [23].

In the field of audio recognition, the combination of CNNs [24] with the use of spectrogram for sound representation [25] has been proven to generate encouraging results in sound recognition, which can be used for environmental sound classification [26–28] and music signal analysis [29,30]. Specifically, both the use of log-Mel spectrogram (LM) and Mel-frequency cepstral coefficient (MFCC) has been proposed for robust representation in sound classification [31].

In the given field of environmental sound recognition in indoor spaces, we highlight several approaches. In [32], the recognition of events, such as a bouncing ball or cricket, was carried out by means of spectral representation of sound with frame-level features, which was learned using Markov models. In [33], two classes of sounds (i.e., tapping and washing hands) were recognised using spectral and histogram of sounds by SVM in naturalistic conditions within a geriatric residence. In part of the study by [8], 3D spatial directional microphones allowed high-quality multidirectional audio to be captured to detect events and the location of sounds in an environment. For this purpose, Mel-frequency cepstral coefficients are computed as spatial features which are related to events using Gaussian and hidden Markov models. In [34], 30 events were collected to recognise the 7 rooms or spaces where the inhabitant carried out activities (bathroom, bedroom, house entrance, kitchen, office, outdoor, and workshop). In this work, log-Mel spectrograms were also evaluated for sound event classification, together with a DL model (VGG-16) pretrained with YouTube audios, with encouraging results but where accuracy was demonstrated to differ notably between controlled conditions and real-life contexts.

Moreover, fuzzy logic has been demonstrated to provide suitable sensor representation from the first AR methods [35] to recent works [21]. In addition, fusing and aggregating heterogeneous data from sensors have become key in edge–fog distributed architectures [36]. In concrete terms, the representation of temporal features by means of fuzzy logic has increased performance in several contexts of AR [10,37]. In addition, fuzzy logic has provided an interpretable representation of outcomes for low-level processing of sensor data [38] and has improved accuracy in uncertain and inaccurate sensor data [39]. Protoforms and fuzzy logic were proposed by Zadeh [40] as a useful knowledge model for reasoning [41] and summarisation [42] of data under uncertainty. The use of protoforms [43] and fuzzy rules to infer knowledge has provided suitable representations [44].

Based on the works and the approaches reviewed in this section, in this work, we present a dataset focused on daily living events in indoor environments to enhance AR using smart IoT boards. The proposed audio recognition model was based on spectral information of audio samples, together with learning from CNNs, which provides high-performance recognition with automatic spatial feature extraction. The audio predictions from DL models were filtered using fuzzy protoforms to provide a coherent recognition of daily audio events which define temporal restrictions. In addition, a case scenario in naturalistic conditions was evaluated to analyse the impact of the recognition of daily events in real time.

2. Materials and Methods

In this section, we describe the proposal of devices, architecture, and methods for ambient sound recognition of daily events by means of smart boards and CNNs. First, in Section 2.1, we present the IoT board and audio sensors in an edge–fog architecture for collecting and labelling environmental sounds. Second, in Section 2.2, a DL model for ambient sound recognition of daily events is presented using a Mel-frequency spectrogram and CNNs. Third, in order to filter the raw audio event recognition, fuzzy processing of audio event streams is included in the IoT boards by means of temporal restrictions defined by protoforms, which is detailed in Section 2.2.

2.1. Materials: Devices and Architecture

In this section, we describe the materials and devices proposed for sound recognition of daily living events in smart environments. In the context of the Internet of Things and ubiquitous computing, the integration of devices into the spaces where the data are collected is characterised by immersiveness and low invasiveness. Here, an edge–fog computing approach was implemented.

First, we proposed the use of audio sensors connected to smart boards to collect and recognise sound events. The selected smart board was Raspberry Pi [45], which enables computing capabilities for machine learning, including deep learning models [46]. The audio sensors integrated were low-cost microphones with a USB connector, providing plug-and-play connectivity with Raspberry Pi under Raspbian Operating System. In Figure 1, we show both connected devices deployed in a bathroom.

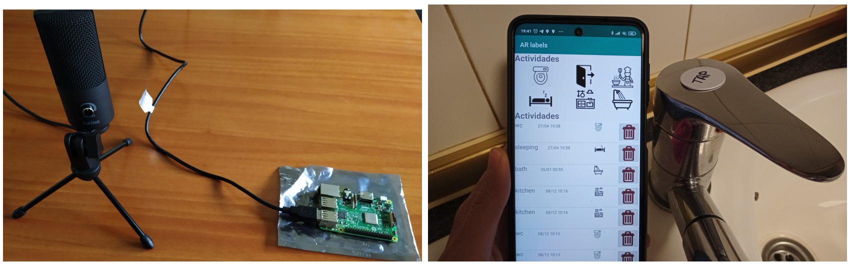

Figure 1. (**Left**) Raspberry Pi B+ with USB microphone which sets up the IoT device for collecting and recognizing ambient sound events; (**Right**) mobile application for labelling of events together with NFC tag to facilitate data collection and labelling.

The aim of integrating audio sensors into smart boards for the recognition of daily events was to (i) collect sound samples for training purposes, (ii) train deep learning models from labelled sound samples, and (iii) recognise audio events to evaluate the trained models in a real-time context. The programming language used to code the application embedded into the Raspberry Pi was Python [47], and the deep learning models were implemented on Python with Keras, an open source library for neural networks [48]. The remote services for labelling of data and spreading the recognised output of audio events in real time were developed under MQTT, which provided a publish/subscribe protocol for wireless sensor networks [49]. This approach was inspired by the paradigms of fog and edge computing [50].

Second, for the purpose of collecting and labelling sound samples from smart environments, the Raspberry Pi collected sound samples of a given duration in the smart board in real time. In addition, the Raspberry Pi board was subscribed to an MQTT topic, where the start and end of each event were published to label a given sound event from a mobile application. Between the start and end of the time interval, the board stored the sound samples, associating each instance with a label. The mobile application for labelling sound samples was developed in Android [51], providing a mobile tool to label the events in a handheld device. In order to facilitate the task of labelling while the daily tasks are performed, NFC tags were placed on the objects and furniture involved in the events, such as doors or taps. The NFC tags automatically activated labelling in the mobile application when touched by the user, sending the start and end of a sound label under MQTT. In Figure 1, we show the NFC tags and the mobile application for labelling sound events.

Third, the recognition model of sound events was trained with the labelled data, computing real-time recognition of ambient sounds. For this purpose, the deep learning model for sound recognition, which is described in Section 2.2, had been previously trained and stored in the Raspberry Pi. The model received the segments of audio samples from the audio sensor as input and classified them according to the target labels. The prediction for each target was published by MQTT in real time to be reachable by other smart devices or AR models.

Fourth, fuzzy filtering of raw audio event prediction was carried out by means of temporal restrictions using linguistic protoforms in an interpretable way. This enabled us to filter predictions which did not match with protoforms defined by fuzzy temporal windows and fuzzy quantifiers. The architecture of components of the proposed approach is described in Figure 2.

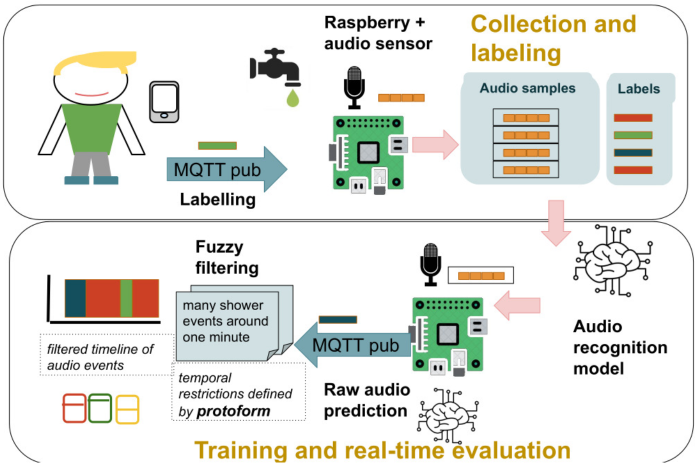

Figure 2. Architecture of components for ambient sound recognition of daily living events. The real-time prediction of sound events was carried out in smart boards under an edge–fog computing approach.

2.2. Deep Learning Model for Ambient Sound Recognition of Daily Events

In this section, we describe a classifier model for ambient sound recognition of daily events based on spectral representation and DL models. First, as detailed previously, the translation from unidimensional digital audio samples to bidimensional spatial representation based on spectrogram features (a picture of sound) provides encouraging results in ambient audio classification [25].

In this work, a window size of 3 s was defined to segment and collect the ambient audio samples, as it provides a suitable time interval for audio recognition [26]. The collection frequency of the of the ambient audio sensor was set to 44.1 kHz.

Next, we extracted two representations of the spectrum of each sound, which were evaluated as input by different CNNs:

- Log-mel spectrogram (LM) was calculated for time–frequency representation of audio signals using a log power spectrum on a nonlinear Mel scale of frequency. When defining the length of the fast Fourier transform window to 2048, it produces images sized 128 × 130.
- Log-scaled Mel-Frequency cepstral coefficients (MFCCs) with 13 components from the raw audio signals, which computes the spectrum of sound using a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency [52]. As traditional MFCCs use between 8 and 13 cepstral coefficients [53], we proposed 13 features to provide the most representative information of audio samples. Based on this configuration, the resulting MFCC spectrogram of positive frequencies developed images sized 13 × 130.

In Figure 3, we provide MFCCs of the audio samples collected from daily living events, which were used as inputs subsequently to be classified with the corresponding sound labels using a CNN.

CNNs are described as feature extractors and classifiers with encouraging results in image recognition [54]. The use of different CNN models with several layers of feature extraction [26,31] has been proposed for ambient audio recognition purposes according to the representation of the spectrum of the sounds. Therefore, in this work, two CNN models were evaluated: (i) a CNN model with five convolutional layers for MFCC processing, where a unique average pooling is included after convolutions due to reduced input space 13 × 130 × 1 and (ii) a CNN model with 5 convolutional layers and a max pooling reduction whose configurations are shown in Table 1.

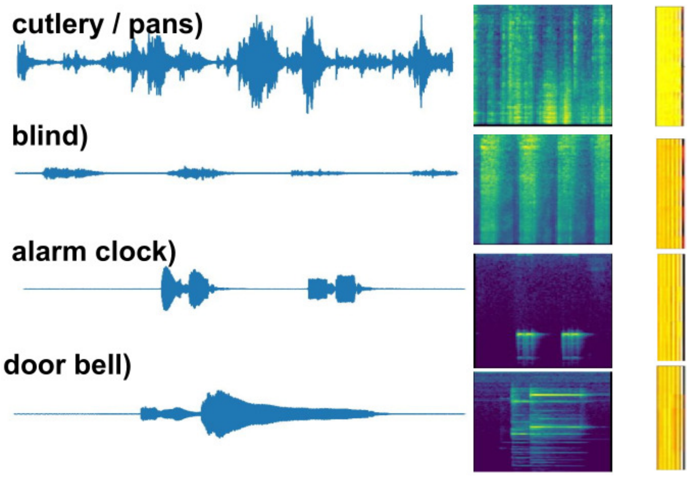

Figure 3. Example of raw audio signals at 44.1 kHz, log-Mel spectrogram (LM), and Mel-frequency cepstral coefficient (MFCC) of the ambient audio events: cutlery, blind, alarm clock, and door bell.

Table 1. Network Architecture from models CNN + MFCC and CNN + LM.

Network Architecture from Model CNN + MFCC	
Input	13 × 130 × 1
Conv(3 × 3)	11 × 128 × 16
Conv(3 × 3)	9 × 126 × 16
Conv(3 × 3)	7 × 124 × 32
Conv(3 × 3)	5 × 122 × 64
Conv(3 × 3)	3 × 120 × 128
Conv(3 × 3)	1 × 118 × 256
GlobalAvgPool2D	256
Dense	1024
Dense	15
Network Architecture from Model CNN + LM	
Input	128 × 130 × 1
Conv(2 × 2)	127 × 129 × 16
Max-Pool(2 × 2)	63 × 64 × 16
Conv(2 × 2)	62 × 63 × 32
Max-Pool(2 × 2)	31 × 31 × 32
Conv(2 × 2)	30 × 30 × 64
Max-Pool(2 × 2)	15 × 15 × 64
Conv(2 × 2)	14 × 14 × 128
Conv(2 × 2)	13 × 13 × 128
Flatten	21,632
Dense	1024
Dense	1024
Dense	15

The models were implemented with Keras under Python to enable integration with Raspberry Pi in real time, using an edge-computing approach which publishes the events detected without exposing sensitive audio sensor data from homes, guaranteeing the privacy of the inhabitant.

Fuzzy Protoforms to Describe Daily Events from Audio Recognition Streams

In this section, we describe the formal representation of audio streams computed under a linguistic representation [36]. The aim of fuzzy processing is to include a filtering process of audio classification in real-time conditions in order to provide temporal restrictions and criteria to identify a given event.

The stream of audio recognition from a smart audio sensor s^j is composed of a set of predictions $s^j = \{m_i^j\}$, where each prediction is represented by $m_i^j = \{v_i^j, t_i^j\}$, where v_i^j represents a given audio from a recognised event, and t_i^j the time-stamp for the sensor j in a given time t_i.

From the sensor streams, we defined *protoforms* which integrate an interpretable, rich, and expressive approach that models the expert knowledge in the stream linguistically. The protoform is in the shape of the following:

$$Q_k V_r T_j$$

where are Q_k V_r T_j are identifiers of the following linguistic terms:

- V_r defines a crisp term, whose value is directly related to a recognised event r.
- T_j defines a fuzzy temporal window (FTW) j where the audio event V_r is aggregated. The FTWs are described according to the distance from the current time t^* to a given timestamp t_i as $\Delta t_i = t^* - t_i$ using the membership function $\mu_{T_j(\Delta t_i)}$, which defines a degree of relevance between $[0,1]$ for the time elapsed Δt_i between the point of time t_i current time t^*.
- We defined an aggregation function of V_r over T_j which computes a unique aggregation degree of the occurrence of the event V_r within a temporal window T_j. Therefore, the following t-norm and t-conorm are defined to aggregate a linguistic term and temporal window:

$$V_r \cap T_j(s_i^{\bar{l}}) = V_r(s_i^l) \cap T_j(\Delta t_i) \in [0,1]$$
$$V_r \cup T_j(s_i^{\bar{l}}) = \bigcup_{s_i^{\bar{l}} \in S^l} V_r \cap T_j(s_i^{\bar{l}}) \in [0,1]$$

  where we use Fuzzy weighted average (FWA) [55] to compute the degree of the linguistic term in the temporal window. In this way, the t-norm computes the temporal degree for each point of time of the temporal window, and the co-norm aggregates these computed degrees in the whole temporal window in a unique representative degree.
- Q_k is a fuzzy quantifier k that filters and transforms the aggregation degree of the audio event V_r within the FTW T_j. The set of quantifiers defined in this domain are represented by the fuzzy sets shown [56]. The quantifier applies a transformation $\mu_{Q_K} : [0,1] \to [0,1]$ to the aggregated degree of $\mu_{Q_K}(A_k \cup T_j(S^r))$ [57].

In this work, a given protoform was defined for each event or audio class to be recognised. The protoform defines temporal restrictions using the relevance of the term (quantifier) in the temporal window (FTW) under conditions of relative normality. For example, the phrase *many vacuum cleaner sounds for half a minute* determines a protoform in which the term *many* defines the quantifier, and the term *half a minute* defines the temporal window. The degree of the protoforms, which is computed between 0 and 1, determines the degree of truth of the recognition of the audio event. Applying these temporal restrictions enabled the removal of false positives in AR, which is key in analysing the normality of behaviours in daily life.

3. Results

In this section, we present the results of the approach. First, a collection of ambient sounds from daily living events in the home is presented, together with the evaluation of the proposed methodology in offline and real-time conditions in different case studies. The data were collected in a home with four rooms (living room, bedroom, kitchen, and bathroom) for an inhabitant who lives there as their usual residence.

First, we created a dataset of ambient sounds from daily living events in the home. The selected activities/events to be recognised in the case study are detailed in Table 2. For each label, a balanced dataset of 100 sound samples with a duration of 3 s was collected in naturalistic conditions. For the labelling of events, the mobile application described in Section 2.1 was integrated to determine the start and end of each event.

Table 2. Sound events and descriptions developed in the context of daily activities.

Class	Description
Vaccum cleaner	Audio sample of vacuuming
Tank	Audio sample of flushing toilet
Cutlery + pans	Audio sample of cutlery and pans
Alarm clock	Audio sample of alarm clock sound
Shower	Audio sample of shower
Extractor	Audio sample of an extractor fan
Kitchen tap	Audio sample of a kitchen tap
Bathroom tap	Audio sample of a bathroom tap
Printer	Audio sample of a printer operating
Microwave	Audio sample of a microwave operating
Blind	Audio sample of a window blind being moved
Door	Audio sample of a door being opened or closed
Phone	Audio sample of a phone ringing
Doorbell	Audio sample of a doorbell ringing

In the first evaluation, a cross-validation method was carried out to analyse the capabilities of the audio recognition model in offline conditions over the collected and balanced dataset with an explicit segmentation of the audio samples with a window size of 3 s. Next, the approach was evaluated in real time over four scenarios in which audio samples were collected from ambient microphones while the inhabitant carried out activities of daily living in naturalistic conditions. The case studies have a duration of 2220 s, with a total of 760 samples analysed.

The dataset of audio samples collected in this work and the labels of the scenes are available in the following repository (Repository: https://github.com/AuroraPR/Ambiental-Sound-Recognition (last access 15 July 2021), which includes the implementation of the proposed methods with Python and Keras.

3.1. *Offline Case Study Evaluation*

In this section, we describe the results provided by the deep learning models based on CNN and LM and MFCC representation for ambient sound recognition with the data collected and a public dataset in an offline context using 10-fold cross-validation.

- Ad hoc ambient audio dataset. In this case, the dataset includes audio samples which have been collected in a single home and were labelled with an explicit segmentation of 3 s for events occurred in controlled conditions using the approach described in Section 2.1. All classes described in Table 2 are included in the dataset.
- Audioset dataset (Repository: https://research.google.com/audioset/ (last access 15 July 2021)). This public dataset provides videos from YouTube and labelling in the segment where a given sound occurs. From the categories of the dataset, we selected 12 events related to our classes which are included in the dataset: “Toilet flush”, “Conversation”, “Dishes, pots, and pans”, “Alarm clock”, “Water”, “Water tap”, “Printer”, “Microwave oven”, “Doorbell”, “Door”, “Telephone ringing” and “Silence”. The sounds collected from Audioset correspond to a balanced dataset with 60 files for each class which includes an explicit segmentation of the sound events.

For each dataset, we present a comparison of the confusion matrices for each fold in the cross-validation that was computed. First, in Figure 4 we present the performance of the DL models in ambient sound recognition of daily events for the ad hoc dataset. Second, in Figure 5, we present the performance of the DL models in the Audioset dataset. In

Table 3, we describe the metrics of f1 score, precision, and recall for both DL models, and the evaluated datasets.

As can be observed, the performance of the ad hoc ambient audio dataset has excellent results for both CNN models in controlled conditions, CNN + MFCC showing the best results. However, the performance in sound recognition of daily events with the Audioset dataset is highly deficient. This is due to the fact that the audio samples from YouTube videos include noise overlapping with other sounds and audio generation from heterogeneous sources. For interested readers, the Audioset samples are available in the repository of this work.

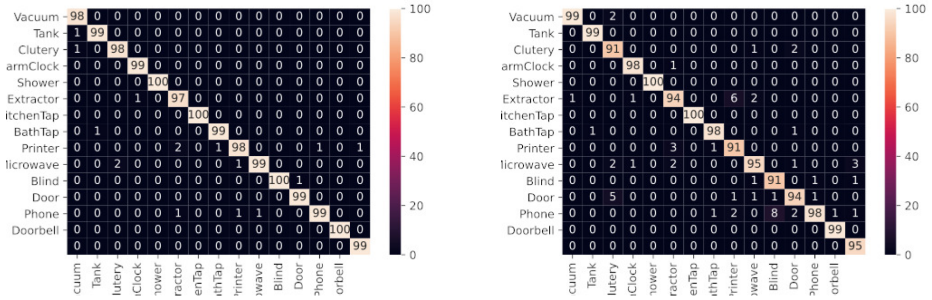

Figure 4. Confusion matrices in ad hoc ambient audio dataset. (**Left**) CNN + MFCC; (**Right**) CNN + LM.

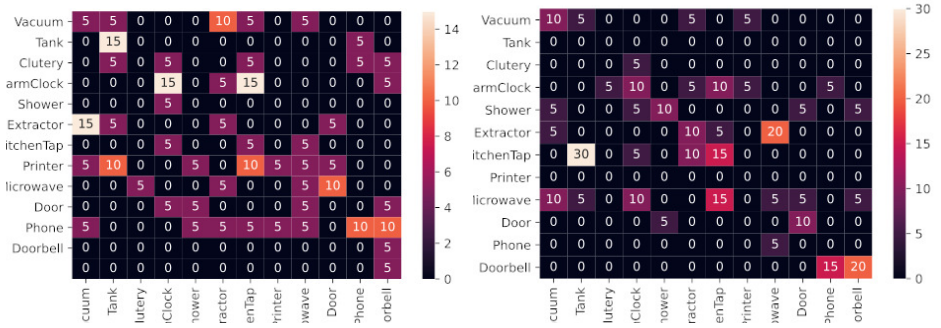

Figure 5. Confusion matrices in Audioset dataset: (**Left**) CNN + MFCC; (**Right**) CNN + LM.

Table 3. Classification metrics from offline case study evaluation.

	Accuracy	Precision	Recall	F1-Score
CNN + MFCC model (ad hoc dataset)	0.99	0.99	0.99	0.99
CNN + LM model (ad hoc dataset)	0.96	0.96	0.96	0.96
CNN + MFCC model (Audioset)	0.23	0.25	0.23	0.23
CNN + LM model (Audioset)	0.29	0.36	0.29	0.32

As we can observe, the collection of an ad hoc ambient audio dataset is strongly recommended given the weak sampling from heterogeneous sources. From the ad hoc ambient audio dataset, we have collected the number of trainable parameters, learning time, millions of instructions (up to 40 epochs) and evaluation time in a Raspberry Pi 3B whose core frequency is 400 MHz, presented in Table 4.

Based on these results, in the next section, we describe the evaluation in real-time conditions using the best configuration with the ad hoc ambient audio dataset and the model based on CNN + MFCC which also requires fewer computational resources for audio recognition learning and evaluation.

Table 4. Trainable parameters, learning time, millions of instructions (MI), and evaluation time.

	Trainable Parameters	Learning Time	Millions of Instructions (MI)	Evaluation Time
Model CNN + MFCC	1.7 M	96 min	230.4×10^3 MI	2.53 s
Model CNN + LM	23.3 M	207 min	496.8×10^3 MI	2.81 s

3.2. Real-Time Case Study Evaluation

Next, we present the results for the evaluation of four scenes at a home in naturalistic conditions using the CNN + MFCC model which performed learning under the ad hoc ambient audio dataset. The six scenes comprised the following sequences of activities:

- (Scene 1) The inhabitant arrived home, went to the kitchen and started talking, then started using cutlery, then turned on the extractor fan for a long while, then turned on the tap, turned on the microwave, and was called on the phone.
- (Scene 2) The inhabitant arrived home, went to the living room and started talking, then started vacuuming, then opened and closed the window blinds and then was called on the phone.
- (Scene 3) The inhabitant arrived home, went to the bedroom and started talking, then started vacuuming, then the alarm clock went off for a long while, then printed some documents, and finally, the individual opened and closed the window blinds.
- (Scene 4) The inhabitant went to the fourth bathroom and started talking, then turned on the tap, then took a shower for a long while, then vacuumed and, finally, flushed the toilet.
- (Scene 5) The inhabitant was talking in the kitchen, then started vacuuming, then talked again and started using cutlery, then opened and closed the window blinds, then turned on the tap and, finally, used the microwave.
- (Scene 6) The inhabitant was in the bathroom vacuuming and started talking, then he took a shower for a long while, then was called on the phone and afterward turned on the tap; finally, the individual left the room closing the door.

In this context, a new label is necessary to recognise *idle* as an event class, which corresponds to the absence of target events, including silence and other ambient sounds produced by the inhabitant. The addition of the *idle* label is key for AR learning in real-time conditions [11,15]. For evaluation purposes, idle activity has been included using a scene cross-validation, where each scene is learned with idle audio samples from other scenes, together with the offline dataset of target events.

In Table 5, we detail the performance of the CNN + MFCC ambient sound recognition model, comparing the ground truth against the inferred classification by means of F1-score, accuracy, precision, and recall for each scene.

Table 5. Classification metrics from real-time case study evaluation for each scene.

	Accuracy	Precision	Recall	F1-Score
Scene 1	0.95	0.97	0.95	0.96
Scene 2	0.99	0.99	0.98	0.98
Scene 3	0.97	0.98	0.97	0.98
Scene 4	0.96	0.97	0.96	0.96
Scene 5	0.91	0.93	0.91	0.92
Scene 6	0.92	0.92	0.90	0.91

3.3. Fuzzy Protoforms and Fuzzy Rules

In this section, we describe the linguistic protoforms which define temporal restrictions from the raw audio prediction in order to provide a coherent recognition of daily audio events. The FTWs and fuzzy quantifiers were defined with membership functions defined by TS and TL functions (listed in Abbreviations). In Tables 6 and 7, we describe the membership functions for quantifiers and FTWs, together with the protoforms for each audio event, which define the temporal restrictions for normality.

Table 6. Membership functions for FTWs and quantifiers defined for the protoform V_r T_j Q_k.

Description in Natural Language	Type	μ_T
some	Q_k	$TR(s_i^l)$ [0.25, 1]
most	Q_k	$TR(s_i^l)$ [0.5, 1]
for a short time	T_j	$TS(\Delta t_i)$ [−6 s, −3 s, 3 s, 6 s]
for a while	T_j	$TS(\Delta t_i)$ [−12 s, −6 s, 6 s, 12 s]

Table 7. Quantifiers and FTWs which define the protoforms corresponding to temporal restrictions for audio recognition.

Event	Quantifier	FTW
Vaccum cleaner	most	for a short time
Tank	most	for a short time
Conversation	some	for a while
Cutlery + pans	most	for a short time
Alarm clock	most	for a short time
Shower	some	for a while
Extractor	some	for a while
Kitchen tap	most	for a short time
Bathroom tap	most	for a short time
Printer	most	for a short time
Microwave	some	for a while
Blind	most	for a short time
Door	most	for a short time
Phone	most	for a short time
Doorbell	most	for a short time
Idle	some	for a while

The impact of filtering the raw audio events from the recognition model was evaluated for the real-time scenarios (offline evaluation was not possible due to not providing a stream of daily events). Beyond the encouraging results described in the previous section, in these scenes, we identified the recognition of scarce audio events which are not related to the correct occurrence of events. In Figure 6, we demonstrate the ground truth and raw audio events predicted in a timeline for the four scenes, including the detection of false-positive events. In Table 8, we describe the false positives and negatives computed from the time interval detection of home events using (i) raw processing and (ii) fuzzy temporal restrictions. Computing the false positives and negatives of time intervals has been described as a relevant metric for detecting events in activity recognition regardless of their duration [10]. The evaluation of these audio events in temporal windows using protoforms, which determine a minimal restriction for recognition, has enabled filtering the most spurious occurrences, as well as defining a degree of adherence between 0 and 1 to the protoform. The use of fuzzy temporal restrictions provides an encouraging method, reducing false positives from the raw audio recognition from 24 occurrences to 2 occurrences while including only 2 false negatives.

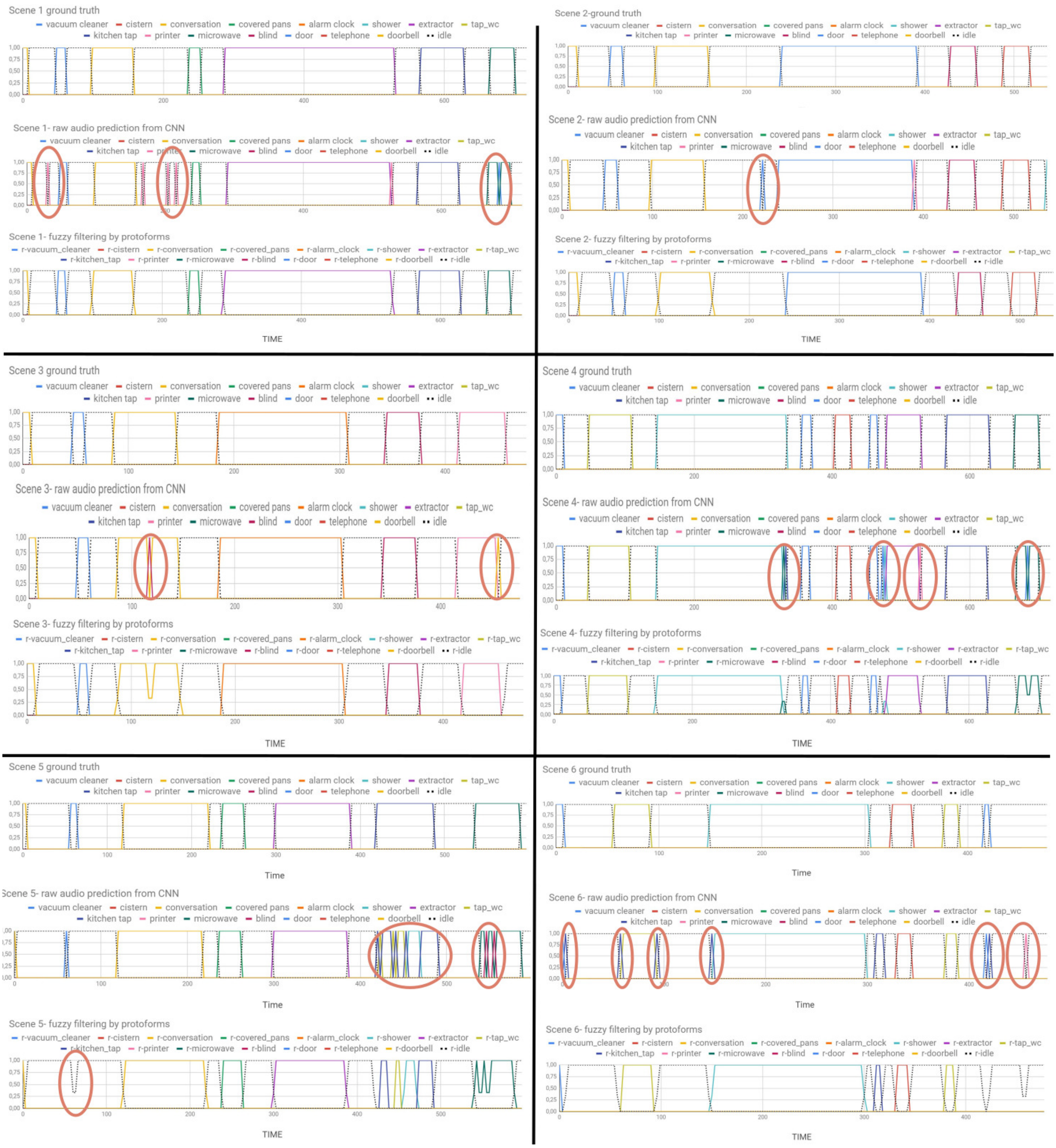

Figure 6. Timeline of the six scenes of the case study: (**Up**) ground truth of the scene; (**Middle**) raw audio recognition from spectral and CNN models; (**Bottom**) fuzzy filtering of audio recognition with protoforms. In red circles are the isolated false positives or false negatives which describe incoherent event recognition.

Table 8. False positives (FP) and negatives (FN) computed from the time interval detection of home events using raw processing and fuzzy temporal restrictions.

	Raw		**Fuzzy**	
	FP	**FN**	**FP**	**FN**
Printer	6	0	0	0
vacuum cleaner	2	0	0	1
blind	3	0	0	0
door bell	1	0	0	0
Kitchen tap	6	0	0	0
microwave	1	0	0	0
shower	2	0	1	0
tap wc	3	0	1	0
door	0	0	0	1
Total	24	0	2	2]

3.4. Limitations of the Work

The activity recognition methods and devices proposed in this work present encouraging performance in offline and real-time recognition of ambient audio events. A balanced dataset with 100 samples per label is sufficient to work in controlled and naturalistic conditions; however, translating the results to deployments "in the wild" [34] would require a larger dataset and additional data preprocessing methods, such as clustering and augmentation. Evaluation with Audioset provided highly deficient results due to noise, overlapping with other sounds, and audio generation from heterogeneous sources. Evaluating audio events in different domains will require extensive datasets and complex processing for domain adaptation methods [58].

4. Conclusions and Ongoing Work

In this work, we evaluated the capabilities of audio recognition models based on spectral information and deep learning to identify ambient events related to the daily activities of inhabitants in a home. To this end, an edge–fog computing approach with smart boards was presented, which enabled the evaluation and recognition of audio samples within the devices while preserving the privacy of the users. Fuzzy processing of audio event streams was included in the IoT boards to filter the raw prediction of audio events by means of temporal restrictions defined by protoforms. The fuzzy processing of audio recognition proved crucial in real-time scenarios to avoid false positives and provide a coherent recognition of daily events detected from protoforms which are directly defined in linguistic terms.

In ongoing research, we aim to integrate a fusion of heterogeneous sensors, such as wearable and binary sensors, to increase the sensing capabilities of audio recognition with other daily activity events. In addition, fuzzy rules could enhance the knowledge-based definition of activities with steady processing from raw data, integrating the data collected from different sensors.

Author Contributions: Conceptualization, J.M.-Q. and C.P.; machine learning model, J.M.-Q. and A.P.-R.; recording of audio samples J.M.V.C.; validation,formal analysis, investigation, supervision, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This contribution has been supported by the Andalusian Health Service by means of the project PI-0387-2018. Funding for this research is provided by EU Horizon 2020 Pharaon Project 'Pilots for Healthy and Active Ageing', Grant agreement No. 857188. Moreover, this research received funding under the REMIND project Marie Sklodowska-Curie EU Framework for Research and Innovation Horizon 2020, under grant agreement No. 734355.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset of audio samples collected in this work and the labels of the scenes are available in the following repository https://github.com/AuroraPR/Ambiental-Sound-Recognition (last access 15 July 2021) , which includes the implementation of the proposed methods with Python and Keras.

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

Sample Availability: The dataset of audio samples collected in this work and the labels of the scenes are available in the next repository https://research.google.com/audioset/ (last access 15 July 2021) which include the implementation with Python and Keras of the proposed methods.

Abbreviations

The following abbreviations are used in this manuscript:

CCN	convolutional neural network
IoT	Internet of Things
MFCC	Mel-Frequency cepstral coefficient
AR	activity recognition
FWA	$V_r \cup T_k(s^j) = \frac{1}{\sum T_k(\Delta t_i^j)} \sum_{m_i^j \in s^j} V_r(v_i^j) \times T_k(\Delta t_i^j) \in [0,1]$
TS	$TS(x)[l_1, l_2, l_3, l_4] = \begin{cases} 0 & x \leq 0 \\ (x - l_1)/(l_2 - l_1) & l_1 \leq x \leq l_2 \\ 1 & l_2 \leq x \leq l_3 \\ (l_4 - x)/(l_4 - l_3) & l_3 \leq x \leq l_4 \\ 0 & l_4 \leq x \end{cases}$
TR	$TR(x)[l_1, l_2] = \begin{cases} 1 & x \leq l_1 \\ (l_2 - x)/(l_2 - l_1) & l_1 \leq x \leq l_2 \\ 0 & l_2 \leq x \end{cases}$

References

1. Chen, L.; Hoey, J.; Nugent, C.D.; Cook, D.J.; Yu, Z. Sensor-based activity recognition. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2012**, *42*, 790–808. [CrossRef]
2. Espinilla, M.; Martínez, L.; Medina, J.; Nugent, C. The experience of developing the UJAmI Smart lab. *IEEE Access* **2018**, *6*, 34631–34642. [CrossRef]
3. Bravo, J.; Fuentes, L.; de Ipina, D.L. Theme Issue: "Ubiquitous Computing and Ambient Intelligence". *Pers. Ubiquitous Comput.* **2011**, *15*, 315–316. Available online: http://www.jucs.org/jucs_11_9/ubiquitous_computing_in_the/jucs_11_9_1494_1504_jbravo.pdf (accessed on 22 July 2021).
4. Rashidi, P.; Mihailidis, A. A survey on ambient-assisted living tools for older adults. *IEEE J. Biomed. Health Inform.* **2012**, *17*, 579–590. [CrossRef]
5. Ordóñez, F.J.; Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. [CrossRef] [PubMed]
6. Ruiz, A.R.J.; Granja, F.S. Comparing ubisense, bespoon, and decawave uwb location systems: Indoor performance analysis. *IEEE Trans. Instrum. Meas.* **2017**, *66*, 2106–2117. [CrossRef]
7. Xu, X.; Tang, J.; Zhang, X.; Liu, X.; Zhang, H.; Qiu, Y. Exploring techniques for vision based human activity recognition: Methods, systems, and evaluation. *Sensors* **2013**, *13*, 1635–1650. [CrossRef] [PubMed]
8. Cruz-Sandoval, D.; Beltran-Marquez, J.; Garcia-Constantino, M.; Gonzalez-Jasso, L.A.; Favela, J.; Lopez-Nava, I.H.; Cleland, I.; Ennis, A.; Hernandez-Cruz, N.; Rafferty, J.; et al. Semi-automated data labeling for activity recognition in pervasive healthcare. *Sensors* **2019**, *19*, 3035. [CrossRef]
9. López-Medina, M.; Espinilla, M.; Cleland, I.; Nugent, C.; Medina, J. Fuzzy cloud-fog computing approach application for human activity recognition in smart homes. *J. Intell. Fuzzy Syst.* **2020**, *38*, 709–721. [CrossRef]
10. Medina-Quero, J.; Zhang, S.; Nugent, C.; Espinilla, M. Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition. *Expert Syst. Appl.* **2018**, *114*, 441–453. [CrossRef]
11. Krishnan, N.C.; Cook, D.J. Activity recognition on streaming sensor data. *Pervasive Mob. Comput.* **2014**, *10*, 138–154. [CrossRef]
12. Radu, V.; Lane, N.D.; Bhattacharya, S.; Mascolo, C.; Marina, M.K.; Kawsar, F. Towards multimodal deep learning for activity recognition on mobile devices. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, Heidelberg, Germany, 12 September 2016; pp. 185–188.
13. Weiser, M. *The Computer for the Twenty-First Century*; Scientific American: New York, NY, USA, 1991.

14. Van Kasteren, T.; Englebienne, G.; Kröse, B.J. An activity monitoring system for elderly care using generative and discriminative models. *Pers. Ubiquitous Comput.* **2010**, *14*, 489–498. [CrossRef]
15. Ordóñez, F.; De Toledo, P.; Sanchis, A. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors* **2013**, *13*, 5460–5477. [CrossRef]
16. Ann, O.C.; Theng, L.B. Human activity recognition: A review. In Proceedings of the 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014), Penang, Malaysia, 28–30 November 2014; pp. 389–393.
17. Laput, G.; Zhang, Y.; Harrison, C. Synthetic sensors: Towards general-purpose sensing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6 May 2017; pp. 3986–3999.
18. Shi, W.; Dustdar, S. The promise of edge computing. *Computer* **2016**, *49*, 78–81. [CrossRef]
19. Bonomi, F.; Milito, R.; Zhu, J.; Addepalli, S. Fog computing and its role in the internet of things. In Proceedings of the first edition of the MCC workshop on Mobile cloud Computing, Helsinki, Finland, 17 August 2012; pp. 13–16.
20. Kortuem, G.; Kawsar, F.; Sundramoorthy, V.; Fitton, D. Smart objects as building blocks for the internet of things. *IEEE Internet Comput.* **2009**, *14*, 44–51. [CrossRef]
21. Lopez Medina, M.A.; Espinilla, M.; Paggeti, C.; Medina Quero, J. Activity recognition for iot devices using fuzzy spatio-temporal features as environmental sensor fusion. *Sensors* **2019**, *19*, 3512. [CrossRef]
22. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [CrossRef]
23. Medina-Quero, J.M.; Burns, M.; Razzaq, M.A.; Nugent, C.; Espinilla, M. Detection of falls from non-invasive thermal vision sensors using convolutional neural networks. *Multidiscip. Digit. Publ. Inst. Proc.* **2018**, *2*, 1236.
24. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
25. Wyse, L. Audio spectrogram representations for processing with convolutional neural networks. *arXiv* **2017**, arXiv:1706.09559.
26. Salamon, J.; Bello, J.P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]
27. Kim, J. Urban sound tagging using multi-channel audio feature with convolutional neural networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events, Tokyo, Japan, 2–3 November 2020.
28. Lasseck, M. Acoustic bird detection with deep convolutional neural networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 19–20 November 2018; pp. 143–147.
29. Choi, K.; Fazekas, G.; Sandler, M. Automatic tagging using deep convolutional neural networks. *arXiv* **2016**, arXiv:1606.00298.
30. Pons, J.; Slizovskaia, O.; Gong, R.; Gómez, E.; Serra, X. Timbre analysis of music audio signals with convolutional neural networks. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August–2 September 2017; pp. 2744–2748.
31. Su, Y.; Zhang, K.; Wang, J.; Madani, K. Environment sound classification using a two-stream CNN based on decision-level fusion. *Sensors* **2019**, *19*, 1733. [CrossRef] [PubMed]
32. Beltrán, J.; Chávez, E.; Favela, J. Scalable identification of mixed environmental sounds, recorded from heterogeneous sources. *Pattern Recognit. Lett.* **2015**, *68*, 153–160. [CrossRef]
33. Beltrán, J.; Navarro, R.; Chávez, E.; Favela, J.; Soto-Mendoza, V.; Ibarra, C. Recognition of audible disruptive behavior from people with dementia. *Pers. Ubiquitous Comput.* **2019**, *23*, 145–157. [CrossRef]
34. Laput, G.; Ahuja, K.; Goel, M.; Harrison, C. Ubicoustics: Plug-and-play acoustic activity recognition. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, Berlin, Germany, 14 October 2018; pp. 213–224.
35. Le Yaouanc, J.M.; Poli, J.P. A fuzzy spatio-temporal-based approach for activity recognition. In *International Conference on Conceptual Modeling*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 314–323.
36. Medina-Quero, J.; Martinez, L.; Espinilla, M. Subscribing to fuzzy temporal aggregation of heterogeneous sensor streams in real-time distributed environments. *Int. J. Commun. Syst.* **2017**, *30*, e3238. [CrossRef]
37. Hamad, R.A.; Hidalgo, A.S.; Bouguelia, M.R.; Estevez, M.E.; Medina-Quero, J. Efficient activity recognition in smart homes using delayed fuzzy temporal windows on binary sensors. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 387–395. [CrossRef] [PubMed]
38. Martínez-Cruz, C.; Medina-Quero, J.; Serrano, J.M.; Gramajo, S. Monwatch: A fuzzy application to monitor the user behavior using wearable trackers. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8.
39. Al-Sharman, M.K.; Emran, B.J.; Jaradat, M.A.; Najjaran, H.; Al-Husari, R.; Zweiri, Y. Precision landing using an adaptive fuzzy multi-sensor data fusion architecture. *Appl. Soft Comput.* **2018**, *69*, 149–164. [CrossRef]
40. Zadeh, L.A. Generalized theory of uncertainty: Principal concepts and ideas. In *Fundamental Uncertainty*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 104–150.
41. Zadeh, L.A. A prototype-centered approach to adding deduction capability to search engines-the concept of protoform. In Proceedings of the 2002 Annual Meeting of the North American Fuzzy Information Processing Society Proceedings, NAFIPS-FLINT 2002 (Cat. No. 02TH8622), New Orleans, LA, USA, 27–29 June 2002; pp. 523–525.
42. Kacprzyk, J.; Zadrożny, S. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Inf. Sci.* **2005**, *173*, 281–304. [CrossRef]

43. Peláez-Aguilera, M.D.; Espinilla, M.; Fernández Olmo, M.R.; Medina, J. Fuzzy linguistic protoforms to summarize heart rate streams of patients with ischemic heart disease. *Complexity* **2019**, *2019*, 1–11. [CrossRef]
44. Akhoundi, M.A.A.; Valavi, E. Multi-sensor fuzzy data fusion using sensors with different characteristics. *arXiv* **2010**, arXiv:1010.6096.
45. Upton, E.; Halfacree, G. *Raspberry Pi User Guide*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
46. Monteiro, A.; de Oliveira, M.; de Oliveira, R.; Da Silva, T. Embedded application of convolutional neural networks on Raspberry Pi for SHM. *Electron. Lett.* **2018**, *54*, 680–682. [CrossRef]
47. Monk, S. *Programming the Raspberry Pi: Getting Started with Python*; McGraw-Hill Education: New York, NY, USA, 2016.
48. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
49. Hunkeler, U.; Truong, H.L.; Stanford-Clark, A. MQTT-S—A publish/subscribe protocol for Wireless Sensor Networks. In Proceedings of the 2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08), Bangalore, India, 6–10 January 2008; pp. 791–798.
50. Medina, J.; Espinilla, M.; Zafra, D.; Martínez, L.; Nugent, C. Fuzzy fog computing: A linguistic approach for knowledge inference in wearable devices. In *International Conference on Ubiquitous Computing and Ambient Intelligence*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 473–485.
51. Darwin, I.F. *Android Cookbook: Problems and Solutions for Android Developers*; O'Reilly Media, Inc.: Newton, MA, USA, 2017.
52. Logan, B. Mel frequency cepstral coefficients for music modeling. *Ismir. Citeseer* **2000**, *270*, 1–11.
53. Rao, K.S.; Vuppala, A.K. *Speech Processing in Mobile Environments*; Springer: Berlin/Heidelberg, Germany, 2014.
54. Ciresan, D.C.; Meier, U.; Masci, J.; Gambardella, L.M.; Schmidhuber, J. Flexible, high performance convolutional neural networks for image classification. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.
55. Dong, W.; Wong, F. Fuzzy weighted averages and implementation of the extension principle. *Fuzzy Sets Syst.* **1987**, *21*, 183–199. [CrossRef]
56. Delgado, M.; Ruiz, M.D.; Sánchez, D.; Vila, M.A. Fuzzy quantification: A state of the art. *Fuzzy Sets Syst.* **2014**, *242*, 1–30. [CrossRef]
57. Medina-Quero, J.; Espinilla, M.; Nugent, C. Real-time fuzzy linguistic analysis of anomalies from medical monitoring devices on data streams. In Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare, Cancun, Mexico, 16–19 May 2016; pp. 300–303.
58. Polo-Rodriguez, A.; Cruciani, F.; Nugent, C.D.; Medina, J. Domain Adaptation of Binary Sensors in Smart Environments through Activity Alignment. *IEEE Access* **2020**, *8*, 228804–228817. [CrossRef]