Explaining Human Activities Instances Using Deep Learning Classifiers

Luca Arrotta
Dept. of Computer Science
University of Milan
Milan, Italy
luca.arrotta@unimi.it

Gabriele Civitarese
Dept. of Computer Science
University of Milan
Milan, Italy
gabriele.civitarese@unimi.it

Michele Fiori
Dept. of Computer Science
University of Milan
Milan, Italy
michele.fiori@studenti.unimi.it

Claudio Bettini
Dept. of Computer Science
University of Milan
Milan, Italy
claudio.bettini@unimi.it

***Abstract*—The recognition of human activities in sensorized smart-home environments enables a wide variety of healthcare applications, including the detection of early symptoms of cognitive decline. The most effective Human Activity Recognition (HAR) methods are based on supervised Deep Learning classifiers. Those models are usually considered as black boxes, and the rationale behind their decisions is difficult to understand for human beings. The recent advances in eXplainable Artificial Intelligence (XAI) offer promising tools to make HAR models more transparent. The state-of-the-art explainable HAR methods provide explanations for the output of classifiers that periodically predict the performed activity on short time windows (usually in the range of 15-60 seconds). However, non-technical users may be more interested in investigating explanations associated with complete activity instances (e.g., an instance of the cooking activity may last 30 minutes). Unfortunately, temporally extending the time window harms the recognition rate of HAR classifiers. In this paper, we propose *DeXAR++*: a novel method that generates explanations for human activity instances based on deep learning classifiers. The sensor data time windows used for classification are encoded as images. *DeXAR++* aggregates the explanations generated by a computer-vision XAI approach on each time window to obtain a single explanation for approximated activity instances. Moreover, *DeXAR++* includes a novel visualization approach particularly suitable for non-expert users. We evaluate *DeXAR++* with both automatic and user-based evaluation methodologies on a public dataset of activities performed in smart-home environments, showing that our results outperform the ones obtained by state-of-the-art methods.**

***Index Terms*—eXplainable AI, Human Activity Recognition, Deep Learning**

I. Introduction

The sensor-based recognition of human activities in smart homes is an enabling technology for a wide variety of healthcare applications [1]. For example, clinicians are interested in continuously monitoring the behavior of fragile elderly subjects to detect early symptoms of cognitive decline [2].

Even though Human Activity Recognition (HAR) has been widely studied in the last two decades, several well-known open issues still limit its deployment in real-world scenarios, including annotated data scarcity and lack of personalization.

Supervised Deep Learning (DL) classifiers are very effective in this domain [3]. However, they are usually considered as black boxes, and the rationale behind their decisions is not transparent to the end-users. Providing explanations to DL-based HAR models' decisions is crucial both for data scientists to fine-tune the sensing setup and the model as well as for clinicians/caregivers to trust and understand the output of the system.

The eXplainable AI (XAI) paradigm recently emerged in the general machine learning community to build more transparent machine learning models [4]. XAI methods usually associate each model's output with an explanation that the target users can effectively understand (e.g., indicating the portions of the input that were important for the prediction). The majority of existing XAI approaches have been proposed for the computer vision and Natural Language Processing domains.

In a previous work, we proposed DeXAR [5]: an XAI approach for HAR based on deep learning models. This domain poses new challenges since it implies explaining the spatio-temporal relationships between raw sensor data measurements that were important for classification. A significant limit of DeXAR is that the activity model periodically classifies the performed activity considering relatively short time windows (e.g., in the order of seconds). Hence, explanations are generated on each time window, and they only reflect a small portion of the whole activity instance (e.g., an instance of the *cooking* activity may last 30 minutes). While explanations generated on such time windows may still help data scientists or technicians to refine the system, they are more difficult to interpret considering non-technical users (e.g., clinicians, caregivers, or the monitored subjects themselves), since they are interested in investigating the whole activity execution. Note that increasing the classification time window is not an option since it would have a negative effect on the recognition rate [6].

In this work, we extend DeXAR by proposing *DeXAR++*: a method for explaining human activities instances using deep learning classifiers in smart-home environments. Similarly to DeXAR, *DeXAR++* transforms sensor data into semantic images to take advantage of CNNs and image-based XAI approaches. *DeXAR++* generates explanations about approximated activity instances by aggregating the explanations computed on short time windows. In order to make explanations readable by end-users who are non-expert in sensors and machine learning, *DeXAR++* generates them both as sentences in natural language and graphical images.

We evaluated *DeXAR++* considering a publicly available

dataset of Activities of Daily Living (ADLs) performed in a smart-home environment. First, we adopted a quantitative metric we recently proposed [5] to automatically evaluate the consistency of the explanations with common-sense knowledge about activity execution. For instance, *"the activity is Cooking because Bob is standing in the kitchen in front of the electrical stove which is turned on"* is an explanation consistent with the *Cooking* activity. On the other hand, *"the activity is Eating because Bob was watching the television, sitting at the kitchen table"* is not an explanation consistent with the *Eating* activity.

We also performed a user-based evaluation taking advantage of the Amazon Mechanical Turk framework, involving 121 subjects that evaluated the quality of our explanations by comparing three different image-based XAI methods. Our results indicate that *DeXAR++* outperforms DeXAR, that generates explanations on time-windows instead of activity instances. Moreover, our results confirm that XAI approaches based on Model Prototypes [7] are more effective than saliency and model induction methods.

The contributions of this paper are threefold:

- We propose a XAI method that explains approximated activity instances based on DL models that perform predictions on relatively short time windows.
- We propose a novel approach to visualize explanations for non-technical users.
- A user-based study on a public dataset involving 121 participants shows that *DeXAR++* generates explanations that are better than the ones generated on time windows.

II. Related Work

A. Human Activity Recognition in Smart-Home Environments

Human Activity Recognition (HAR) in smart-home environments has been widely studied in the last decades [1], [8]. While some approaches proposed solutions based on cameras [9], such devices may be perceived as intrusive in the home environment, especially considering elderly subjects. In this work, we focus on unobtrusive sensor-based ADLs recognition. Several sensors can be used in home environments: a) environmental sensors to track the interaction of the subject with the surrounding environment (e.g., magnetic sensors on doors, presence sensors, plug sensors, etcetera), b) wearable sensors to keep track of the subjects' physical movements (e.g., smartwatches), c) micro-localization sensors to keep track of the position of the user in the home (e.g., Bluetooth Low Energy beacons, UWB antennas).

Three main categories of HAR methods have been proposed to classify the performed activities from the continuous stream of sensor data: *data-driven*, *knowledge-based*, and *hybrid*. Data-driven approaches are mainly based on supervised machine learning techniques. Deep learning is becoming predominant in this area since it generally leads to high recognition rates [3]. However, training data is generally hard to acquire in smart-home environments due to privacy problems, intrusiveness, and costs. Knowledge-based methods do not require labeled data since they rely on logic-based formalisms (e.g., ontologies) that define semantic relationships between sensor events and ADLs [10]. The main drawback of knowledge-based methods is their lack of flexibility. Indeed, human activities exhibit high variability in execution modalities, dynamic context information, and the intrinsic uncertainty of sensor data. Finally, hybrid solutions are emerging to combine the strengths of both worlds [11]. Knowledge-based methods are directly interpretable and hence easily explainable. This paper focuses on the more challenging task of explaining supervised approaches based on deep learning.

Recently, some research groups proposed to tackle sensor-based activity recognition as an image classification problem to take advantage of Convolutional Neural Networks (CNNs) [12]–[16]. However, the images generated by those approaches can not be directly interpreted without experience in activity recognition and sensing.

B. Explainable AI for Human Activity Recognition

Most of the XAI approaches in the literature for HAR consider methods based on cameras and computer vision [17]–[19]. Deriving explanations for HAR approaches based on sensor data poses new challenges.

The majority of existing approaches in this area considered inherently interpretable models. For instance, the work proposed in [20] is based on feature importance derived from the model parameters of Random Forests. The authors in [21] proposed a rule-based classifier in charge of automatically learning a set of human-readable rules encoding correlations between sensor events and activities. This model reaches recognition rates similar to well-known interpretable classifiers (e.g., Decision Tree, JRip) while generating simpler explanations. In [22], the authors proposed a fuzzy rule-based model.

Besides using models that are not accurate as the ones based on deep learning, the above-mentioned works do not generate explanations dedicated to non-expert users. On the contrary, HealthXAI [23] uses an inherently interpretable model to provide explanations in natural language targeted to clinicians. However, that work focuses on detecting high-level abnormal behaviors of elderly subjects. Since HealthXAI is also based on HAR (that is not explainable), we believe that *DeXAR++* tackles an orthogonal problem.

To the best of our knowledge, the only existing work that tackles the problem of explaining deep learning HAR models is DeXAR [5]. However, its explanations for non-expert users are generated on relatively short time windows that do not reflect the actual activity instances performed by the subjects.

III. Methodology

DeXAR++ aims at explaining the output of a deep learning classifier in charge of recognizing human activities based on a continuous stream of smart-home sensor data. The generated explanations are targeted at humans with no experience in machine learning and sensing (e.g., caregivers, clinicians) that are interested in monitoring the resident's behavior. However, in order to maximize the recognition rate, activity classifiers

periodically provide an output considering relatively short sensor data time windows. Usually, such time windows do not reflect the duration of the whole activity instances.

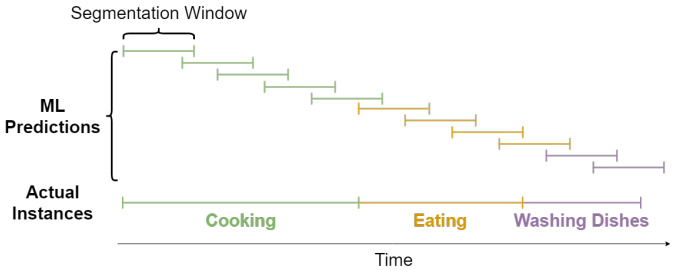


Fig. 1. The time span of classification time windows compared to the time span of the corresponding activity instances

Example 1: Figure 1 compares the time span of the classification time windows with the time span of the actual activity instances performed by the subject. As it is possible to note, as common in HAR methods, time windows have fixed-length and overlap. Considering the *Cooking* activity, a single instance is covered by a relatively high number of time windows.

Example 2: Alice is an elderly woman living alone in her smart apartment, which is instrumented with several sensors. Based on the stream of sensor data, a deep learning classifier outputs every 30 seconds the activities performed by Alice. At 11:00 AM, Alice starts cooking homemade lasagna, and she ends the preparation at 12:30 PM. Assuming that the classifier always outputs the correct activity, during this time interval, it predicts 180 times that the current activity performed by Alice is *preparing a hot meal*.

By looking at Example 2, it is evident how each activity instance may potentially be associated with a high number of time windows. Existing XAI approaches are designed to provide an explanation for each classified time window. While this approach may be useful to technicians that need to tune the classifier and the sensing setup, this approach may not be effective considering non-expert end-users interested in monitoring the overall activity instances (e.g., considering the example above, that Alice is cooking from 11:00 AM to 12:30 PM). *DeXAR++* aims at mitigating this problem by grouping the classified time windows to approximate activity instances. For each approximated activity instance, the explanations generated on each time window are aggregated to generate a single human-readable explanation.

A. Architecture

Figure 2 depicts the overall architecture of *DeXAR++*.

For the sake of this work, we assume a resident living alone in her smart apartment. The home environment is instrumented with several environmental sensors (e.g., magnetic sensors, plug sensors, motion sensors, etcetera) to monitor the resident's interactions with the home environment. The resident

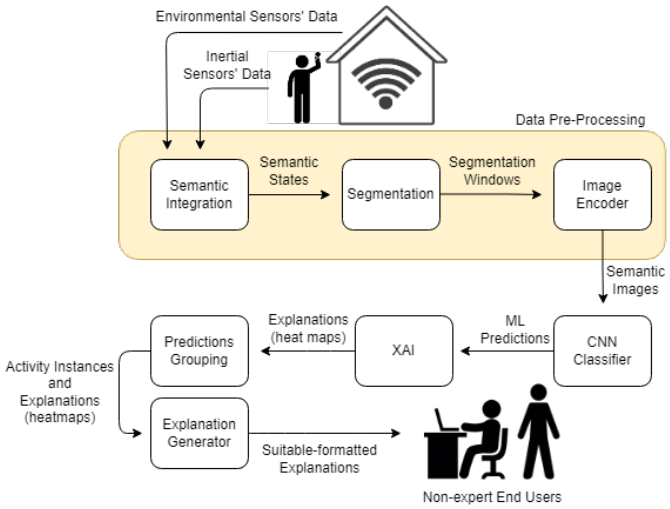


Fig. 2. Overall architecture of *DeXAR++*

also wears a smartwatch equipped with inertial sensors (e.g., accelerometer, gyroscope, magnetometer) that continuously monitor her physical movements. Sensor data are continuously collected by a smart-home gateway and pre-processed by the SEMANTIC INTEGRATION module, which is in charge of inferring semantic states (i.e., high-level information with a clear semantic that describes what happened within a time interval). Then, semantic states are segmented into fixed-length and overlapping time windows by the SEGMENTATION module. In order to take advantage of the effectiveness of CNNs, the IMAGE ENCODER module converts each segmentation window of data to a semantic image representing temporal relationships between semantic states. Each input image is provided to the CNN CLASSIFIER to obtain the most likely activity performed by the resident. Then, the XAI module outputs a heat map as an explanation for each classification output. The PREDICTIONS GROUPING module continuously clusters the classified semantic images and the corresponding explanations to approximately compute activity instances. Moreover, this module is also in charge of generating a single explanation for the approximated activity instances (as a heat map). Finally, since heat maps are not intuitive for non-expert end-users, the EXPLANATIONS GENERATOR module translates them into a suitable format.

B. From Semantic Integration to Image Generation

We take advantage of the method that we originally proposed in DeXAR [5] to transform the stream of sensor data into a sequence of semantic images.

First, we infer from sensor measurements the *semantic states*: high-level events with a clear semantics. Each semantic state $S[t_a, t_b]$ indicates that the high-level event S occurred in the time interval $[t_a, t_b]$.

Example 3: The semantic state $fridge\_door\_open[t1, t2]$ indicates that the fridge door has been opened in the time

interval $[t1, t2]$, and it is generated by observing the activation of the magnetic sensor at the timestamp $t1$ and its deactivation at timestamp $t2$.

Semantic states are then segmented considering fixed-length time windows of w seconds and an overlap factor ov. For each time window, *DeXAR++* generates a semantic image encoding the temporal relationships between the semantic states that occurred during a specific time interval $[t_s, t_e]$. Note that t_s is the start timestamp of the interval, t_e is the last timestamp of the time interval, $t_s < t_e$, and $t_e - t_s = w$.

Besides semantic states, a semantic image generated on the interval $[t_s, t_e]$ also encodes the K activity instances that the system reliably detected before the current one.

Example 4: Consider a semantic image computed considering a window in the interval $[11{:}30{:}00AM, 11{:}30{:}30AM]$. Since the system reliably detected that the resident started and concluded the activities *Watching TV* from $10{:}12{:}00AM$ to $10{:}25{:}00AM$ and *Using PC* from $10{:}26{:}00AM$ to $11{:}28{:}00AM$ (and no other activities) these information are encoded in the semantic image.

Note that *DeXAR++* relies on approximated activity instances (i.e., the same used to generate explanations) in order to compute past activities. The specific algorithm will be presented later in Section III-D.

For the sake of brevity, in this paper, we do not replicate the specific details about the semantic image generation process [5]. In the following, we provide an example of a semantic image generated by *DeXAR++*.

Example 5: Figure 3 depicts an example of a semantic image computed on a time window of 16 seconds corresponding to the *Clearing Table* activity. The semantic image is divided into two main components: semantic states (highlighted in red) and past activities instances (highlighted in green). The white segments in the first component indicate the occurrence of the following semantic states: $DynamicHandManipulations[1, 6]$, $SittingOnDiningChair[1, 6]$ $DynamicHandManipulations[13, 16]$, $InTheDiningRoom[1, 12]$, and $InTheKitchen[13, 16]$. Indeed, the x-axis represents each second of the time window, while the y-axis represents the different semantic states. The white squares in the past activities component indicate that, before the current activity, the resident completed the activities *Setting up table* and *Eating* (in this order).

C. Semantic Image Classification and XAI explanations

Each semantic image is provided to a CNN classifier that outputs the most likely activity A performed by the subject. The specific CNN architectures used in our experimental setting are reported in Section IV-D.

By applying XAI methods for computer vision (e.g., GradCAM [24], LIME [25]), it is possible to obtain a

Fig. 3. Example of image used to feed the activity classifier related to the activity *clearing table*. The semantic states and the past activities sub-matrices are highlighted in red and green, respectively

heat map h depicting the contribution of each pixel to the classification of A by the deep learning classifier. Each pixel in h has a value in the range $[0, 1]$ based on its importance during classification (i.e., relevance).

Example 6: Figure 4 shows the application of the GradCAM algorithm on a semantic image that the CNN model classified as *Clearing Table*. According to this specific XAI approach, the most important features for classification are the past activities performed by the resident (*setting up the table* and *eating*) and the presence of the resident in the kitchen and the dining room.

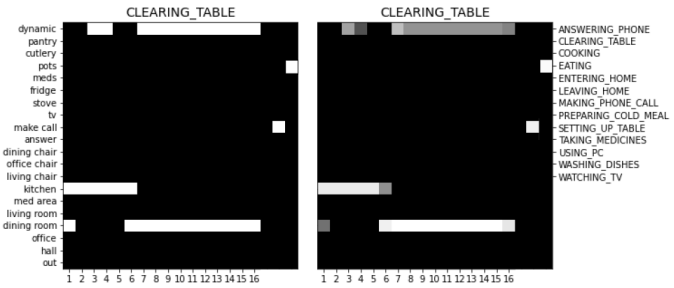


Fig. 4. Left: input image given to the classifier with ground truth *clearing table*. Right: explanation for the predicted class *clearing table* using the XAI method Grad-CAM

D. Predictions grouping

A heat map derived from a semantic image encodes the features that are important to classify a specific time window. In the following, we explain our novel approach to aggregate classified semantic images and the corresponding heat maps to generate explanations for approximated activity instances.

Algorithm 1 shows how we continuously process the semantic images to derive activity instances[1]. For each input image img_i, the CNN classifier outputs the most likely activity a_i and

[1]Recall that activity instances are also needed to populate past activities in the semantic images, as we discussed in Section III-B

its confidence c_i. Thanks to an XAI algorithm that processes images, we also compute the heat map h_i, which explains why img_i has been classified as a_i. In Section IV-B, we will describe the specific XAI algorithms that we experimented in this work. We group consecutive semantic images (and their corresponding explanations) classified with the same activity label into an instance $inst$. In order to exclude unreliable predictions, we only consider those instances that include at least α predictions with a confidence higher than σ.

Algorithm 1 Approximating activity instances

```
1: instances ← ∅
2: a_old ← NIL
3: inst ← ∅
4: for img_i in the stream do
5:     (a_i, conf_i) ← classify img_i
6:     h_i ← explanation of a_i using XAI
7:     if a_i == a_old then
8:         inst ← inst ∪ {(a_i, conf_i, h_i}
9:     else
10:        if |{(a, conf, h) ∈ inst|conf > σ}| > α then
11:            instances ← instances ∪ {inst}
12:        end if
13:        inst ← ∅ ∪ {(a_i, conf_i, h_i)}
14:    end if
15:    a_old ← a_i
16: end for
```

E. Generating an explanation for each instance

Given an activity instance $inst = \{(a_1, conf_1, h_1), \ldots, (a_n, conf_n, h_n)\}$ derived by Algorithm 1, *DeXAR++* computes a single explanation h^{inst} by aggregating the heat maps $h_1, \ldots, h_n$ that are associated with the predictions in $inst$. This process computes the two components of h^{inst} (i.e., semantic states and past activities) separately.

1) Semantic states: Each heat map h_i only encodes w seconds of $inst$. Moreover each h_i overlaps (with a factor ov) the heat map h_{i-1}. Hence, the overall duration (in seconds) of the activity instance $inst$ is:

$$l(inst) = w + w \cdot (1 - ov) \cdot (n - 1) \quad (1)$$

Example 7: Consider a segmentation window with size $w = 12s$ with an overlap factor $ov = 75\%$. Suppose that the system recognized an instance of the activity *taking medicines* composed of 10 predictions. Considering Equation 1, the overall duration of this instance is $12s + 12s \cdot 0.25 \cdot 9 = 39s$.

For this reason, we temporally align the semantic states components of the heat maps $h_1, \ldots, h_n$ to capture the actual temporal relationships between all the high-level events that occurred during the activity instance. Indeed, the semantic states component of h^{inst} should represent the most important semantic states that occurred in $l(inst)$ seconds.

Example 8: Figure 5 shows an example of temporal alignment considering an instance of the activity *Entering Home* composed of three time windows. Each column (considering all the heat maps) encodes a second t of the whole activity instance.

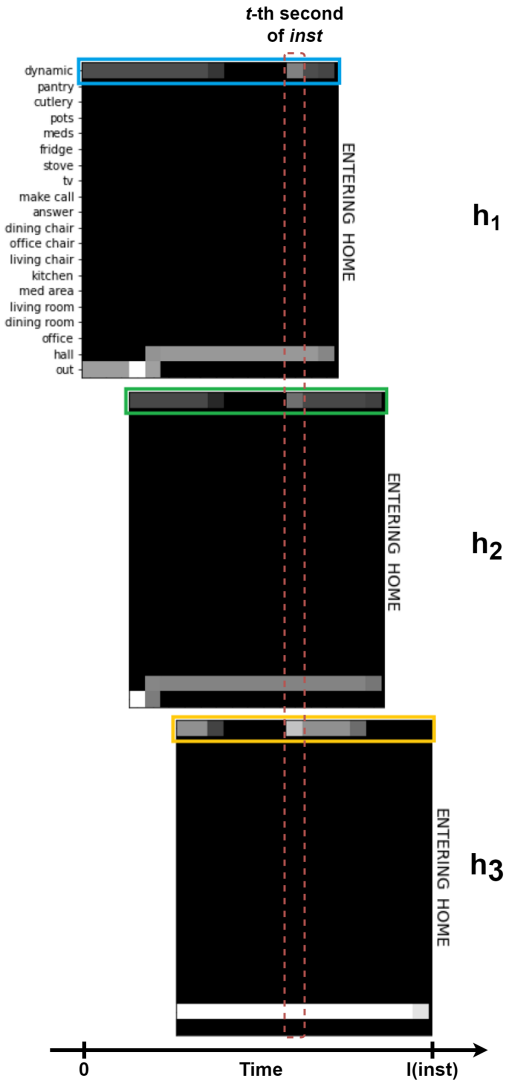


Fig. 5. Temporal alignment of the semantic state components of three consecutive heat maps included in an *Entering Home* activity instance.

Algorithm 2 shows how *DeXAR++* computes the semantic states component of h^{inst} after the temporal alignment of $h_1, \ldots, h_n$.

Algorithm 2 Computing h^{inst} semantic states component

```
for semantic state s do
    for second t = 1, ..., l(inst) do
        maxRel_t^s ← max(rel_t^s(h_1), ..., rel_t^s(h_n))
        if maxRel_t^s ≥ τ then
            rel_t^s(h^inst) ← maxRel_t^s
        else
            rel_t^s(h^inst) ← 0
        end if
    end for
end for
```

From now on, we denote t as the t-th second of $inst$ after

the temporal alignment of the heat maps $h_1, \ldots, h_n$ of $inst$. Let $rel_t^s(h_i)$ be the relevance of the pixel of the row encoding the semantic state s in the heat map h_i at second t. For each semantic state s, *DeXAR++* finds its maximum relevance (i.e., pixel value) $maxRel_t^s$ across all the heat maps of $inst$ at each second t. Specifically, $maxRel_t^s$ is computed as follows:

$$maxRel_t^s = max(rel_t^s(h_1), \ldots, rel_t^s(h_n)) \tag{2}$$

Hence, we compute the relevance $rel_t^s(h^{inst})$ of the pixel in the row encoding the semantic state s in h^{inst} at second t as follows:

$$rel_t^s(h^{inst}) = \begin{cases} maxRel_t^s & \text{if } maxRel_t^s \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

DeXAR++ uses the relevance threshold τ to include in h^{inst} only the semantic states that are highly relevant, excluding noisy pixels.

Figure 6 shows the semantic states component of the explanation obtained for the *Entering Home* instance previously introduced as an example in Figure 5.

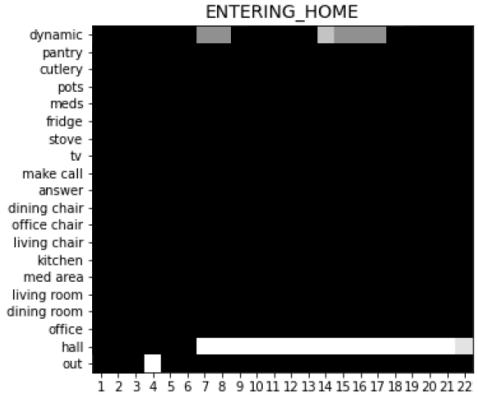


Fig. 6. Semantic states component of the explanation of an *Entering Home* activity instance (Figure 5). Since the instance was composed of 3 temporal windows of 16 seconds with $ov = 0.8$, $l(inst) = 22$.

2) Past activities: Similarly to the semantic states component of h^{inst}, for each past activity p, *DeXAR++* determines its maximum relevance among the heat maps $h_1, \ldots, h_n$, as follows:

$$maxRel^p = max(rel^p(h_1), \ldots, rel^p(h_n)) \tag{4}$$

where $rel^p(h_i)$ is the relevance of the pixel encoding the past activity p in h_i. Thus, the relevance $rel^p(h^{inst})$ of p in h^{inst} is determined by the following equation:

$$rel^p(h^{inst}) = \begin{cases} maxRel^p & \text{if } maxRel^p \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

F. Explanations Generator

Given an instance $inst$, we post-process the heat map h^{inst} to generate an explanation for non-expert users. This explanation comprises two components: a) a sentence in natural language and b) a graphical plot that simplifies the heat map. An example of explanation is provided in Figure 7. The sentence is generated from the heat map by following the NLP approach that we originally proposed in DeXAR [5]. However, while natural language is easy to understand for non-expert users, its intrinsic ambiguity may negatively impact the quality of explanations. This is especially true considering temporal relationships between high-level events. For this reason, *DeXAR++* also provides a novel graphical plot that is a simplified version of the heat map showing the temporal relationships between the semantic states with positive relevance in h^{inst}.

IV. Experimental Evaluation

This section describes the experimental evaluation we carried out to evaluate the effectiveness of the explanations generated by *DeXAR++*.

A. Dataset

For the sake of this study, we consider the *MARBLE* [26] dataset, which includes Activities of Daily Living (ADLs) performed in a smart-home environment. We selected this dataset to compare our results with the ones of DeXAR [5].

This dataset considers a home environment equipped with several environmental sensors: magnetic sensors on some drawers (e.g., the medicine drawer) to detect their opening/-closing events, smart plugs to identify the usage of some home appliances (e.g., the television), and pressure mats to detect which chair the resident is sitting on. Overall, 12 subjects were involved in the data collection campaign. While performing ADLs, each user was also carrying: a) a smartphone in the pocket in charge of detecting incoming and outgoing phone call events and b) a smartwatch on the wrist in charge of collecting inertial sensor data. Overall, the dataset includes 13 different ADLs' types: *Answering Phone*, *Clearing Table*, *Cooking a Hot Meal*, *Eating*, *Entering Home*, *Leaving Home*, *Making a Phone Call*, *Cooking a Cold Meal*, *Setting Up Table*, *Taking Medicines*, *Working*, *Washing Dishes*, and *Watching TV*.

B. Experimental Setup

Since, to the best of our knowledge, DeXAR [5] is the only approach in the literature facing the same problem tackled by *DeXAR++*, it is the only baseline that we use for comparison. Hence, in this work we propose a similar experimental setup. Specifically, we split the dataset into three partitions: 70% for training, 10% for validation, and 20% for test. We consider three different XAI methods:

- *Grad-CAM* [24]: a saliency-based method that analyzes how the last layers of the CNN activate in order to generate the heat map.
- *LIME* [25]: a model induction method that considers the classifier as a black box. By providing several perturbed versions of the input, LIME generates a heat map that highlights the most important pixels for classification.
- *Model Prototypes* [7]: a deep explainable method that requires a specifically designed CNN to automatically

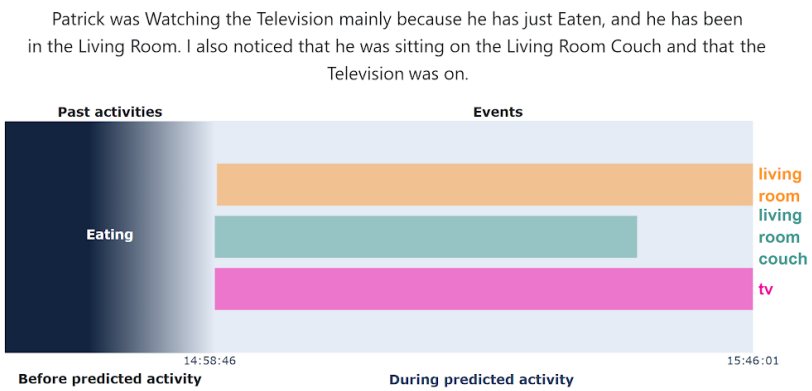


Fig. 7. Explanation for a *Watching TV* activity instance

learn prototypical representations of the output classes. The learned prototypes are used to explain the output of the classifier. For the sake of fairness, we use the method proposed in DeXAR that has been adapted to the HAR domain in order to generate heat maps.

We use two different CNN models depending on the XAI method being used (the ones proposed in DeXAR). The first is called *CNN-GL*, and it is a standard convolutional network that is used to evaluate *Grad-CAM* and *LIME*. The second is called *CNN-MP*, and it is an explainable-by-design model that is required by *Model Prototypes*. Test set images are then given as input to the trained CNN classifier in order to produce the explanations used in our evaluation.

We selected the hyper-parameters of *DeXAR++* with a grid search approach. The length of the time-windows is $w = 16s$, with an overlap factor $ov = 0.8$. Figures 8 and 9 show how the values of α and σ affect the accuracy in deriving activity instances when using *CNN-MP* and *CNN-GL*, respectively[2]. By increasing α, *DeXAR++* generates better activity instances. However, this also increases the number of discarded time windows. The same behavior is observed when σ is increased. Considering the trade-off between accuracy in generating activity instances and the number of discarded time windows, we chose $\alpha = 2$ and $\sigma = 0.75$ for *CNN-MP*: the 88.7% of the activity instances generated by *DeXAR++* are correct[3], with a 12.5% of discarded predictions. Note that the 53.5% of those discarded time windows were actually miss-classified. On the other hand, we selected $\alpha = 1$ and $\sigma = 0.75$ for *CNN-GL*. This allows generating an 87.3% of correct activity instances, with a discarding rate of 12.2%. In this case, the 60.1% of discarded predictions were actually miss-classified. Figure 10 shows the Empirical Cumulative Distribution Function of the number of time windows for each instance predicted by *DeXAR++*, considering both CNN-MP and CNN-GL. As it is possible to observe, the distribution of the two models is similar. We observed that most activity instances are characterized by a high number of time windows.

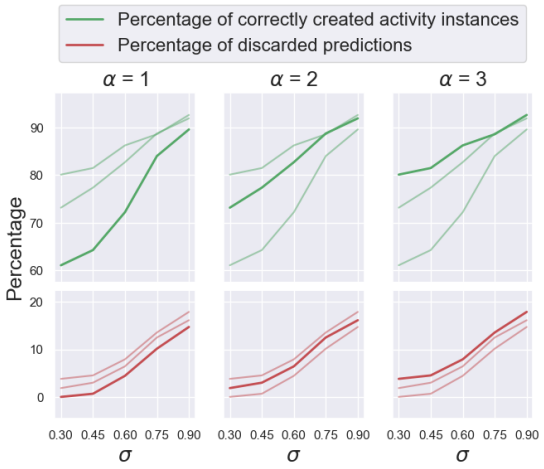


Fig. 8. Impact of α and σ on CNN-MP.

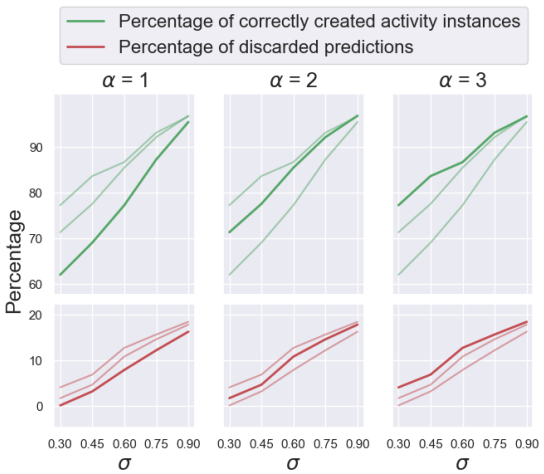
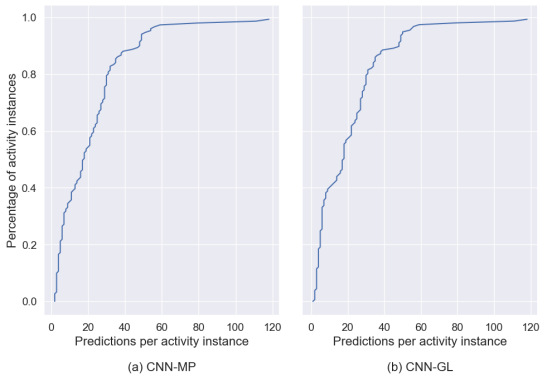


Fig. 9. Impact of α and σ on CNN-GL.

For the sake of fairness, we report in Table I and II the recognition rate that we obtain on both CNNs (at the time-window granularity). These results are consistent with the ones previously presented in DeXAR [5].

[2]Recall that we only consider those instances containing at least α time-windows predicted with a confidence greater than σ.

[3]We consider an activity instance *correct* when the 90% of its time windows are correctly classified considering the ground truth.



(a) CNN-MP (b) CNN-GL

Fig. 10. Distribution of the number of segmentation windows included in the each activity obtained with (a) CNN-MP and (b) CNN-GL

TABLE I
RECOGNITION RATE OF THE CNN USED FOR GRADCAM AND LIME

	precision	recall	f1-score
macro avg	0.80	0.83	0.80
weighted avg	0.90	0.89	0.90

C. Evaluation Methodologies for Explanations

1) Explanation Score: We adopt an evaluation metric, called *Explanation Score*, that we originally proposed to evaluate DeXAR. This metric provides an automatic evaluation of the explanations' quality. In particular, the *Explanation Score* estimates the degree to which explanations are consistent compared to common-sense knowledge about the relationships between activities and semantic states/past activities. The common-sense knowledge consists of high-level properties of the HAR domain that people acquire during their everyday life and on which there is a general agreement. For instance, *washing dishes* is an activity that is commonly performed in the kitchen after *eating* while the subject is standing.

According to common-sense knowledge, a semantic state or a past activity f partially explains an activity A if f explains (even if partially) A.

Example 9: The semantic state *television ON* partially explains the *watching tv* activity, while it does not partially explain *preparing a cold meal*, even if it could actually occur while preparing a cold meal.

The semantic model proposed in DeXAR (based on an ontology) is used to quantitatively evaluate the explanations automatically generated by *DeXAR++*. Given a heat map h^{inst}, the corresponding activity A, and $F^\star$ as the semantic states and past activities with positive relevance in h^{inst}, we compute the *common-sense relevance* $cr()$ $\forall f \in F^\star$:

$$cr(f,A) = \begin{cases} \max_t rel_t^f(h^{inst}) & \textit{if } f \textit{ partially explains } A \\ -\max_t rel_t^f(h^{inst}) & \textit{otherwise} \end{cases} \quad (6)$$

The maximum pixel value for a semantic state/past activity f in h^{inst} is associated when f partially explains the activity A, while the corresponding negative value is associated when f does not partially explain A.

Finally, we compute the *Explanation Score* based on the common-sense relevance:

$$ExplanationScore(F^\star, A) = \begin{cases} \frac{\sum_{f\in F^\star} cr(f,A)}{\sum_{f\in F^\star} |cr(f,A)|} & \textit{if } F^\star \neq \emptyset \\ -1 & \textit{otherwise} \end{cases} \quad (7)$$

Common-sense relevances of semantic states and past activities (whose values are included in the range $[-1,1]$) impact the explanation score. Note that the score is -1 if there are no semantic states or past activities in $F^\star$ that partially explain A, while it is equal to 1 if every semantic state or past activity partially explains A.

2) User-based evaluation: In order to estimate how non-expert users would perceive the explanations, we performed a user-based study by recruiting ≈ 600 users on the Amazon Mechanical Turk platform. Unfortunately, many of them were discarded from our analysis since we classified them as unreliable. Indeed, we introduced some attention questions to exclude bots and users providing random answers just to obtain the reward. For instance, Figure 11 shows the answers' distribution to a question that assessed whether users clearly understood the aim of the study. Hence, we could only select

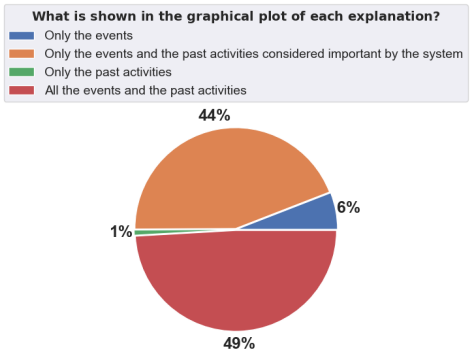


Fig. 11. Answers' distribution to an attention question in our survey. Only 44% of the users provided the correct answer.

TABLE II
RECOGNITION RATE OF THE MODEL PROTOTYPES MODEL

	precision	recall	f1-score
macro avg	0.79	0.81	0.78
weighted avg	0.90	0.90	0.90

121 reliable users for our user-based evaluation. Overall, the 37% of those selected users were females, while the remaining 63% were males. The 29% of the selected users were aged between 18 and 30, 58% were aged between 31 and 50, while the remaining 13% were 50 years-old users. Considering the education level, 26% of the selected users had a high-school diploma, 58% a Bachelor's degree, 11% a Master's degree, and the remaining 5% a doctoral level. Each user

TABLE III
EXPLANATION SCORE OF *DeXAR++* COMPARED WITH THE ONES OF DEXAR [5]. NOTE THAT, CONSIDERING DEXAR, THE SCORE IS COMPUTED ON EACH TIME WINDOW, WHILE ON *DeXAR++* IT IS COMPUTED AT THE INSTANCE LEVEL. FOR EACH XAI APPROACH, WE ALSO SHOW THE SCORE ON THE CORRECT PREDICTIONS AND THE WRONG ONES

	Model Prototypes			**LIME**			**Grad-CAM**		
	All	Correct	Wrong	All	Correct	Wrong	All	Correct	Wrong
DeXAR [5]	0.94	0.95	**0.87**	0.80	0.82	0.70	0.65	0.64	**0.67**
DeXAR++	**0.96**	**0.97**	0.91	**0.85**	**0.88**	**0.60**	**0.80**	**0.81**	0.78

was required to complete a survey to evaluate *DeXAR++*'s explanations. In particular, we generated different sets of activity instances detected by *DeXAR++* randomly sampled from the test set. Each set includes 13 activity instances (one for each activity class). Each instance is associated with three different explanations obtained by each considered XAI method. Each user is first informed about the aim of the survey, the home environment, and the considered activities. Then, the system randomly assigns one of the generated sets of activity instances to each user. For the sake of this work, we only considered explanations associated with correctly classified activity instances. We asked the participants to vote for each explanation with a grade from 1 (absolutely not satisfying) to 5 (completely satisfying). The users were not aware of the method that generated each explanation.

D. Results

1) Explainability evaluation based on common-sense knowledge: Table III compares the results of the common-sense knowledge evaluation obtained through *DeXAR++* with the ones obtained by *DeXAR*. For clarity, the Explanation Score is normalized in the range [0,1].

TABLE IV
IMPACT OF τ ON THE EXPLANATION SCORE OBTAINED BY THE ACTIVITY INSTANCES GENERATED FROM THE PREDICTIONS OF THE MODEL PROTOTYPES MODEL

τ Value	Explanation Score		
	All	Correct	Wrong
0.0	0.81	0.80	0.84
0.1	0.87	0.87	0.89
0.2	0.91	0.91	0.90
0.3	0.94	0.94	0.91
0.4	0.96	0.96	0.91
0.5	**0.96**	**0.97**	**0.91**
0.6	0.94	0.94	0.91
0.7	0.91	0.91	0.91
0.8	0.89	0.89	0.91
0.9	0.83	0.82	0.91

The overall Explanation Score of *DeXAR++* is computed by averaging the score assigned to every activity instance detected on the test set. On the other hand, considering *DeXAR*, we compute the average of the scores assigned to each time window on the test set.

Overall, *DeXAR++* outperforms *DeXAR* regardless of the XAI method used to generate the explanations. Hence, besides providing a single explanation for the whole activity instances, *DeXAR++* also provides explanations that are more consistent with the HAR common-sense knowledge encoded in our semantic model. Consistently with DeXAR, *Model Prototypes* outperforms LIME and Grad-CAM. This means that *Model Prototypes* gives more importance to those semantic states and past activities consistent with the common-sense knowledge encoded in our semantic model. On the other hand, Grad-CAM is the worst in terms of Explanation Score. Indeed, it is known in the literature that such saliency-based approaches often consider features that are not completely relevant for classification [27]. Moreover, considering all the XAI methods, the explanation score on wrongly classified activities is lower than the score computed on the correctly classified ones. These results confirm the quality of the considered metric.

Table IV shows the impact on the explanation score of the parameter τ, which is the relevance threshold to choose the semantic states/past activities that should be included in h^{inst}. We observed that a low τ value leads to including noisy information in the final explanation, negatively impacting its consistency with common-sense knowledge. On the other hand, a too high τ score is too restrictive, leading to an instance's heat map with insufficient information to adequately explain the activity.

2) User-based explainability evaluation: Figure 12 shows the distribution of the grades provided by the recruited users considering the different XAI approaches. We compare these results with those of a similar user-based study performed on DeXAR [5]. Model Prototypes is confirmed as the most accurate XAI approach, while Grad-CAM is the one that generally received the worst grades. Even though the results of *DeXAR++* seem to be better than the ones of DeXAR, this result still has to be confirmed since the users involved in the two studies were different. These results show that the explanation score evaluation correctly estimated the actual user perception of the explanations generated by *DeXAR++*.

V. CONCLUSION

In this paper, we presented *DeXAR++*: a deep explainable method for sensor-based HAR. Compared to existing works, *DeXAR++* aims at providing explanations to whole activity instances. The results of *DeXAR++* show that the quality of the generated explanations outperforms existing approaches that provide explanations to time windows, providing a more practical framework for non-expert users. In future work, we will extend *DeXAR++* in order to capture more complex human behaviors, including anomalies in performing activities. Indeed, among other applications, such a system

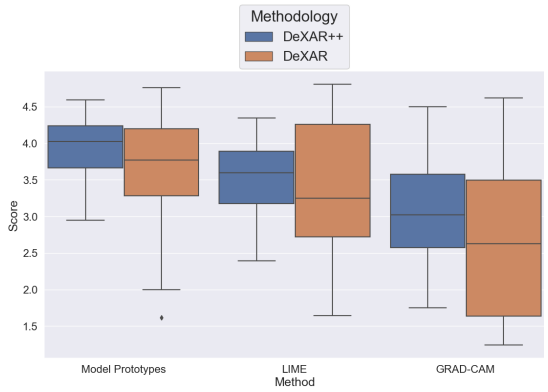


Fig. 12. Boxplot that describes the grade distribution of the user-based study. We compare our results with the ones obtained in DeXAR.

would provide relevant indicators for clinicians non-experts in machine learning interested in monitoring early symptoms of cognitive decline. Hence, we will perform a user-based evaluation considering clinicians (e.g., neurologists). Finally, we will investigate how to mitigate the over-reliance problem that is well-known in XAI systems. Indeed, non-expert users may become over-confident in the system even when it is wrong, just because it provides explanations.

ACKNOWLEDGMENT

The authors want to thank Martina Baga for her excellent work on software implementation.

REFERENCES

[1] U. Bakar, H. Ghayvat, S. Hasanm, and S. C. Mukhopadhyay, "Activity and anomaly detection in smart home: A survey," *Next Generation Sensors and Systems*, pp. 191–220, 2016.

[2] L. Liu, E. Stroulia, I. Nikolaidis, A. Miguel-Cruz, and A. R. Rincon, "Smart homes and home health monitoring technologies for older adults: A systematic review," *International journal of medical informatics*, vol. 91, pp. 44–59, 2016.

[3] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[4] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215, IEEE, 2018.

[5] L. Arrotta, G. Civitarese, and C. Bettini, "Dexar: Deep explainable sensor-based activity recognition in smart-home environments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–30, 2022.

[6] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014.

[7] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[8] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, 2012.

[9] A. Jalal, M. Z. Uddin, and T.-S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 863–871, 2012.

[10] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 961–974, 2011.

[11] A. S. A. Sukor, A. Zakaria, N. A. Rahim, L. M. Kamarudin, R. Setchi, and H. Nishizaki, "A hybrid approach of knowledge-driven and data-driven reasoning for activity recognition in smart homes," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4177–4188, 2019.

[12] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 381–388, IEEE, 2016.

[13] M. S. Singh, V. Pondenkandath, B. Zhou, P. Lukowicz, and M. Liwickit, "Transforming sensor data to the image domain for deep learning—an application to footstep detection," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2665–2672, IEEE, 2017.

[14] M. Gochoo, T.-H. Tan, S.-C. Huang, S.-H. Liu, and F. S. Alnajjar, "Dcnn-based elderly activity recognition using binary sensors," in *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 1–5, IEEE, 2017.

[15] T.-H. Tan, M. Gochoo, S.-C. Huang, Y.-H. Liu, S.-H. Liu, and Y.-F. Huang, "Multi-resident activity recognition in a smart home using rgb activity image and dcnn," *IEEE Sensors Journal*, vol. 18, no. 23, pp. 9718–9727, 2018.

[16] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "Deep learning for human activity recognition: A resource efficient implementation on low-power devices," in *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*, pp. 71–76, IEEE, 2016.

[17] C. Roy, M. Shanbhag, M. Nourani, T. Rahman, S. Kabir, V. Gogate, N. Ruozzi, and E. D. Ragan, "Explainable activity recognition in videos.," in *IUI Workshops*, 2019.

[18] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Explainable video action reasoning via prior knowledge and state transitions," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 521–529, 2019.

[19] S. Suzuki, Y. Amemiya, and M. Sato, "Skeleton-based explainable human activity recognition for child gross-motor assessment," in *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*, pp. 4015–4022, IEEE, 2020.

[20] C. Bettini, G. Civitarese, and M. Fiori, "Explainable activity recognition over interpretable models," in *2021 IEEE International Conference on Pervasive Computing and Communications Workshops*, pp. 32–37, IEEE, 2021.

[21] M. Atzmueller, N. Hayat, M. Trojahn, and D. Kroll, "Explicative human activity recognition using adaptive association rule-based classification," in *2018 IEEE International Conference on Future IoT Technologies (Future IoT)*, pp. 1–6, IEEE, 2018.

[22] H. W. Guesgen, "Using rough sets to improve activity recognition based on sensor data," *Sensors*, vol. 20, no. 6, p. 1779, 2020.

[23] E. Khodabandehloo, D. Riboni, and A. Alimohammadi, "Healthxai: Collaborative and explainable ai for supporting early diagnosis of cognitive decline," *Future Generation Computer Systems*, vol. 116, pp. 168–189, 2021.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[26] L. Arrotta, C. Bettini, and G. Civitarese, "The marble dataset: Multi-inhabitant activities of daily living combining wearable and environmental sensors data," in *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pp. 451–468, Springer, 2021.

[27] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un) reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280, Springer, 2019.