Explainable Activity Recognition for Smart Home Systems

DEVLEENA DAS, Georgia Institute of Technology, USA
YASUTAKA NISHIMURA, KDDI Research, Japan
RAJAN P. VIVEK, Georgia Institute of Technology, USA
NAOTO TAKEDA, KDDI Research, Japan
SEAN T. FISH, Georgia Institute of Technology, USA
THOMAS PLÖTZ, Georgia Institute of Technology, USA
SONIA CHERNOVA, Georgia Institute of Technology, USA

Smart home environments are designed to provide services that help improve the quality of life for the occupant via a variety of sensors and actuators installed throughout the space. Many automated actions taken by a smart home are governed by the output of an underlying activity recognition system. However, activity recognition systems may not be perfectly accurate and therefore inconsistencies in smart home operations can lead a user to wonder "why did the smart home do that?" In this work, we build on insights from Explainable Artificial Intelligence (XAI) techniques to contribute computational methods for explainable activity recognition. Specifically, we generate explanations for smart home activity recognition systems that explain what about an activity led to the given classification. To do so, we introduce four computational techniques for generating natural language explanations of smart home data and compare their effectiveness at generating meaningful explanations. Through a study with everyday users, we evaluate user preferences towards the four explanation types. Our results show that the leading approach, SHAP, has a 92% success rate in generating accurate explanations. Moreover, 84% of sampled scenarios users preferred natural language explanations over a simple activity label, underscoring the need for explainable activity recognition systems. Finally, we show that explanations generated by some XAI methods can lead users to lose confidence in the accuracy of the underlying activity recognition model, while others lead users to gain confidence. Taking all studied factors into consideration, we make a recommendation regarding which existing XAI method leads to the best performance in the domain of smart home automation, and discuss a range of topics for future work in this area.

Additional Key Words and Phrases: Smart Home, Activity Recognition, Explainable AI, Human-AI-Interaction

1 INTRODUCTION

Smart homes are residential environments that are augmented with sensors, actuators and computational reasoning systems designed to provide services that improve quality of life for the occupant [5]. Research in, and deployment of, smart home environments has expanded rapidly in recent years due to the availability of low cost and low power sensors, complemented with advances in wireless technologies and generalizable Machine Learning systems [18, 22, 31]. While some smart home systems simply allow users to remotely control devices (e.g., change the thermostat while away from the house) [20], of greater long-term interest and impact are smart homes that provide intelligent automated assistance. Example capabilities range from automating the lights to complement the user's activity [34], to turning off appliances if the user leaves the house [29], monitoring activities of daily living [18, 39, 40], alerting caregivers of anomalies in a user's behavior [7, 23], and providing aid to users who require assistance in independent living [37, 42]. Through the development of such capabilities, smart homes have the potential for significant societal impact in supporting healthcare and independent living.

In each of the above examples, automated actions taken by the smart home are governed by the output of an underlying Activity Recognition (AR) system. For example, detecting that the user is waking up may trigger the lights to turn on, leaving the house may cause the stove to turn off, and repeated visits to the medicine cabinet

Manuscript is under review. Please write to ddas41@gatech.edu for up-to-date information

may result in a verbal reminder that medications were already taken that day. However, activity recognition systems are not perfectly accurate. For example, in [4], 13% of Work activities were misclassified as Relaxing, in [56] 6% of Brushing Teeth activities were misclassified as Sleeping, and in [24] 2% of Sleeping activities were misclassified as Housekeeping. Although performance of activity recognition algorithms will continue to improve, expectations of perfectly accurate systems are impractical, particularly in complex real-world settings in which occupant behavior or the number of occupants may change over time.

Inconsistencies in activity recognition lead to smart home operations that are inappropriate or surprising to the user, such as turning off the lights even though the user is still awake. In such scenarios, it is natural for the user to ask "why did the smart home do that?" To date, no research has considered the *explainability* of smart home operations, and activity recognition in particular. Do users find it sufficient to just view the label of the recognized activity? Or do they prefer a more detailed explanation of why a particular activity was detected? If the latter, how can such explanations be generated, and what information should they contain in order to be interpretable by users who are not experts in Artificial Intelligence (AI)?

Prior work in human-computer-interaction (HCI) has demonstrated that the explainability of AI systems is important for their success [1, 14, 63]. The field of Explainable AI (XAI) has emerged specifically for research on the development of interpretable machine learning algorithms that can increase the transparency of black-box models [3, 10, 13, 47, 65]. While the majority of XAI techniques focus on expert systems designed for machine learning developers [50, 51, 62, 64], a growing body of work at the intersection of XAI and HCI has developed explainable methodologies targeting non-technical users [11, 16].

In this work, we build on insights from leading Explainable AI techniques (LIME [47], SHAP [33], and Anchors [48]) to contribute computational methods for *explainable activity recognition*. Specifically, we introduce activity recognition techniques that not only generate a label for the observed activity but also an accompanying explanation about which temporal sensory observations led to the given classification. We study these techniques in the context of activity monitoring for smart home systems, focusing on a healthcare scenario in which a remote caregiver seeks to monitor the activities of an older adult living alone. Specifically, our work makes the following contributions:

(1) Integrating leading approaches in activity recognition and XAI research, we introduce four computational techniques for generating natural language explanations of temporal, multimodal activity recognition data and compare their effectiveness at generating meaningful and accurate explanations.

(2) We apply each of the above techniques to smart home activity recognition. Through a study with everyday users, we evaluate user preferences with respect to the type of explanations users find most effective.

Our results show that: *i)* the leading approach, SHAP, has a 92% success rate in generating accurate explanations, although significant tradeoffs exist between explanation accuracy and computational efficiency across the XAI methods studied; *ii)* in 84% of sampled scenarios users preferred natural language explanations over a simple activity label; and *iii)* explanations generated by some XAI methods can lead users to lose confidence in the accuracy of the underlying activity recognition model, while others lead users to gain confidence. Taking all studied factors into consideration, we find SHAP-based explanations to be the most effective when considering multiple factors such as XAI model accuracy, computational efficiency and user preference; however, we also observe that user preferences among XAI explanation types vary and that no one XAI explanation type is consistently favored in all scenarios.

2 RELATED WORK

Our work aims at exploring the value automated explanations can bring to smart home applications and practical deployments. In what follows we summarize existing work related to two major research directions: *i)* activity recognition in smart homes; and *ii)* methods of explainable AI. Our work draws upon methods from both

categories, thereby contributing to accessible and understandable, and thus potentially more acceptable, smart home technologies.

2.1 Activity Recognition in Smart Homes

Smart home environments are designed to provide services that help improve the quality of life for the occupant via a variety of sensors and actuators installed throughout the space, and through automated inference about an occupant's activities and their relevant contexts [49]. A wide variety of sensors have been utilized to capture relevant contextual factors in smart homes, including (but not limited to) those that measure ambient temperature, record sound, motion, wifi signal characteristics, visual information (through, for example, cameras), and proximity information [2, 7, 61]. Furthermore, body worn sensors are also widely used for direct recording of an occupant's movements [60]. In the context of the smart home domain, activity recognition specifically serves to identify and log the occupant's activities of daily living (ADL) from temporal-ordered sensor data. The automatically inferred activity labels then determine which assistive actions the smart home shall perform, such as turning on a light, closing the garage door, or adjusting a thermostat. Activity recognition in smart home environments remains a challenging problem specifically demanding innovation in Machine Learning (ML) for automated sensor data analysis, mainly due to the diversity of environments, users, sensor configurations and actuators that must be considered, particularly due to the unique nature of each individual household and user.

Prior work in AR for smart home domains has studied a wide variety of Machine Learning classification techniques, including Conditional Random Fields [38], Naive Bayes [45], Hidden Markov Models [43, 59], and Artificial Neural Networks [8, 35]. Recently, with the popularity of deep learning, Liciotti et al. [31] have shown that LSTM based architectures for smart home systems can outperform the above mentioned traditional ML techniques. Specifically Liciotti et al. employed five LSTM architectures (uni-LSTM, Bi-LSTM, Cascading-LSTM, Ensemble-LSTM, and Cascade-Ensemble-LSTM) and compared their performances between each LSTM model as well as against traditional ML techniques. The authors reason that the network's ability to model long term dependencies between sequences of data and to learn non linear feature representations allow for the LSTM-based models to better model ADL patterns in unbalanced datasets.

Several large-scale smart home activity recognition datasets have been developed to aid the research community in robust evaluation and benchmaking. Datasets capturing activities of daily living include CASAS [9], ARAS [6], Placelab [26], the MIT Activity Recognition dataset [54], and the van Kasteren dataset [59]. Each of these datasets covers a varying range of ADL activities, and differs in the number of household occupants, as well as the number and types of sensors utilized for data collection. For example, the ARAS dataset includes multiple household occupants observed over a two month time-span using 20 home sensors, while the CASAS dataset covers a range of both single and multi-person households over 2-8 month time spans using 20-86 sensors.

In this work, our goal is not to improve the state-of-the-art in activity recognition itself. Instead, our work explores the added benefit that model-agnostic XAI methods can bring when paired with leading AR approaches. To do so, we leverage Liciotti et al.'s state-of-the-art LSTM-based AR [31], specifically adapting their uni-LSTM model. We validate our approach using the CASAS dataset, specifically using the Milan household data, which was also used in [31]. We describe our model and the dataset in detail in Section 4.

2.2 Explainable AI (XAI)

The field of Explainable AI focuses on the development of algorithms that provide insights into how an AI system makes decisions and predictions. A number of recent surveys of XAI provide detailed perspectives on various aspects of the XAI problem, including general surveys of the XAI field [3, 10, 13, 47, 65], XAI applied to medical domains [58], use of natural language processing in XAI [44], and user experience in XAI research practices [19]. Due to the breadth of the field, we focus our discussion here on XAI methods designed for classification-based

problems, such as activity recognition as it is targeted by our work, as well as XAI methods addressing non-expert users. We refer readers to the above surveys for all additional topics relating to XAI.

Insights from psychology: Prior work in the field of psychology has provided insights into the desirable qualities of an explanation. Early work by Hilton [25] posed explanations as a social and conversational process, arguing that good explanations must not only be true, but must answer a "why" question (whether one was asked or not). A taxonomy of explanations used in a number of psychology works [12, 32] further categorizes explanations into three types: *i)* mechanistic; *ii)* teleological; or *iii)* formal. Mechanistic explanations are given with regards to how something functions, and teleological explanations appeal to purpose. These types of explanations are helpful for explaining processes. Formal explanations are those which are concerned with categorical definition and help explain why an element is considered as part of a set. As our work seeks to improve non-expert understanding of a classifier's decision, our explanations are most concerned with mechanistic explanations.

Explaining classification problems: Classification-based XAI methodologies can be categorized along several axes. First we categorize XAI techniques by the complexity and generalizability of models they can explain:

model-intrinsic, which refer to models that leverage an inherently interpretable structure, such as a rule list [30] or decision tree [53], that does not require further processing to be explainable; or

model-agnostic, such that they provide explanations for underlying computational models that in themselves are not interpretable [33, 47, 48, 66].

Model-intrinsic techniques do not scale well to complex, multi-dimensional spaces [3]. Model-agnostic techniques are typically more desirable since they are not dependent upon a fixed underlying classification model and can be applied widely to any state-of-the-art technique.

Additionally, XAI techniques can be characterized as:

local, which explain the behavior of the model for a specific singular decision; or

global, which explain the behavior and reasoning of the entire model.

In the context of activity recognition we are most interested in local models that explain the model reasoning that led to the generation of a specific activity label. For example, if a smart home turns off the lights while the user is watching TV, the user may ask "why did you think I was sleeping?" and will require an answer about the classification of that particular activity and not a description of the complete AR model.

In this work, we focus on XAI techniques that are designed to be model-agnostic and provide local interpretability. State-of-the-art model-agnostic methodologies include LIME [47], SHAP [33], and Anchors [48]. While LIME utilizes perturbation to find surrogate models that fit and explain individual instances, SHAP leverages game theory foundations of Shapley Values [52], which provide the marginal contributions of features towards an input instance being explained. Anchors formulates a multi-armed bandit problem [27] found in reinforcement learning to produce "IF-THEN" rules that provide local interpretability for individual instances. Each of these methodologies outputs an ordered list of the local features most relevant to a classification (e.g., "[M026, M017, T002]"). In this paper, we examine the effectiveness of LIME, SHAP and Anchors when applied to activity recognition for smart home data, and contribute techniques for generating interpretable natural language explanations from their output. We provide additional details of each of these algorithms and our approach in Section 5.

XAI for everyday people: Most prior work in XAI has focused on developing explantion techniques for machine learning experts, developers, or specific domain experts [3, 21, 50, 51, 64]. However, recent efforts at the intersection of XAI and HCI have led to increased development of techniques for interaction with everyday users, or non-experts. For example, in the context of explaining video activity recognition for cooking tasks, Nourani et al. [41] establish that high levels of veracity in explanations can improve human task performance and system understanding. Additionally, recent work has established the importance of contextualized reasoning as well as the importance of natural language explanations for effective understanding. Specifically, Ehsan et al. [16]
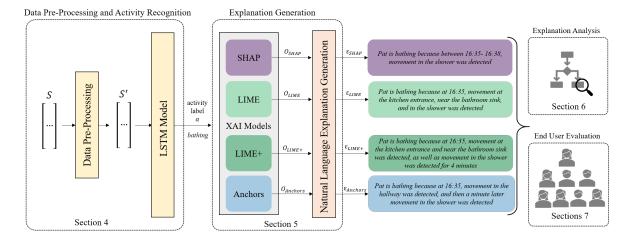
Fig. 1. Overview of the work presented in this paper, outlining the different components within our work. For data pre-processing and activity recognition (Section 4), we first pre-process smart home data S into a multivariate, fixed-interval representation S', suitable for explainable activity recognition. The output activity label, a, as well as the activity recognition model is used by each XAI model to generate top contributing features O. These features are translated into a natural language explanation types $\mathcal{E}$, corresponding to the XAI model (Section 5). We analyze the accuracy of the explanation types (Section 6) as well as their effectiveness in helping end users understand a smart home's activity recognition label (Section 7).

leveraged sequence-to-sequence learning to autonomously generate natural language rationales in the domain of Markovian-based games, such as Frogger, to study user preferences of rationales. Their results demonstrate that users significantly prefer "complete-view" rationales, which utilize the entire state space as context, as opposed to "focused-view", which utilize only a subset of the full state space. Additionally, in the specific context of understanding robot failures yet related to the work presented here, Das et al. [11] extended sequence to sequence learning to autonomously generate context-based natural language explanations in a continuous state space. Their results demonstrated that explanations grounded in environmental as well as temporal and / or historical context, most accurately helped non-experts identify robot failures as well as correct recovery solutions. However, both of the above sequence-to-sequence methodologies require vast amounts of training data in the form of expert-labeled explanations, making them challenging to generalize to new applications. In this work, we take a template-based approach to generating explanations, similar to Elizalde et al. [17] who utilize template-based approach to explain recommendations made by a Markov Decision Process for operating a steam generator.

Evaluation of explanations: A recent survey by Miller [36] provides valuable insights on how computational techniques from explainable AI can build on existing research in social sciences, reviewing relevant papers from philosophy, cognitive psychology/science, and social psychology, and how they relate to XAI. Of particular interest is Miller's investigation of the criteria that researchers have used to evaluate explanations, which include the coherence and simplicity of an explanation [46, 55], as well as its truthfulness [28]. In this work, we evaluate our explanation for truthfulness in Section 6. We then control for truthfulness, and evaluate the coherence of our explanations with real-world users in Section 7.

3 OVERVIEW

Our overarching goal is to adopt explainable AI techniques in order to make activity recognition models interpretable to everyday people through natural language explanations. Toward this goal, we take a leading activity

recognition model (the uni-LSTM from [31]) and integrate it with three leading model-agnostic XAI techniques – LIME [47], SHAP [33] and Anchors [48]. We validate our work within a commonly targeted smart home scenario: assisted living and autonmous health care through a smart home. To do so, we utilize a widely used smart home activity recognition dataset (the CASAS Milan dataset [9]) and contribute a technique for automatically generating natural language explanations of activity labels using the above methods. We evaluate the resulting explanations for accuracy, and also examine the opinions of everyday users by conducting a user study comparing the resulting explanations. Across these efforts, our work seeks to answer the following research questions:

RQ1: Do smart home users prefer natural language explanations over simple activity labels?
RQ2: Do explanations give users more confidence in the activity recognition model?
RQ3: Out of the explanation generation methods explored in this work, which produces the most accurate explanations?
RQ4: Which explanation method is ultimately most effective?

In the subsections below, we present an overview of our approach, including the core components of our framework and our evaluation method.

3.1 Problem Formulation

Let $S = \{(s_1, v_1, k_1), ..., (s_t, v_t, k_t)\}$ represent a sequence of sensor events of length t, where sensor s_i takes on the value v_i at time k_i. Let $A = \{a_1, ..., a_n\}$ represent the set of activity labels. Given this data, our goal is to find model $f(S) \rightarrow \{a, \mathcal{E}\}$, where $a \in A$ represents the activity label for the events described by S, and $\mathcal{E}$ is an explanation that describes which events in S most significantly influenced the selection of a as the label. As in prior work on XAI for non-technical users [11, 16], we seek to represent $\mathcal{E}$ in the form of a natural language phrase.

3.2 Explainable activity recognition framework

Figure 1 gives an overview of our approach for the above problem. The objective of the first component of our framework, described in Section 4, is to first pre-process S into a fixed-interval multivariate data representation, S', that facilitates explainable activity recognition. We then use S' to generate an activity label a using an LSTM model extended from [31].

The second component of our system, presented in Section 5, generates the natural language explanations, $\mathcal{E}$, that describe the features of S' that most significantly contributed to the classifier's decision. Our work evaluates four different techniques for explanation generation based on state-of-the-art techniques in XAI. Since this is the first work to examine natural language explanations for activity recognition, we do not contribute an entirely new XAI algorithm but rather evaluate how current leading methods in the fields of XAI and AR can be brought together to address the unique temporal data challenges presented by the smart home activity recognition domain. Specifically, we extend the following three leading XAI methods from the literature:

Local Interpretable Model-agnostic Explanations (LIME) [47] – a model-agnostic XAI technique that explains the predictions of a black-box classifier by learning an interpretable model locally around the prediction;
Anchors [48] – a model-agnostic XAI technique that explains the behavior of black-box models with high-precision rules found via a multi-arm-bandit search ; and
SHapley Additive exPlanations (SHAP) [33] – a model-agnostic XAI technique that calculates Shapley values [52] to understand the feature importance of all features utilized by a black-box model for a prediction.

Each method is used to generate an explanation $\mathcal{E}_{LIME}, \mathcal{E}_{Anchors}$ and $\mathcal{E}_{SHAP}$, respectively. Additionally, we extend the LIME methodology to introduce **LIME+** and $\mathcal{E}_{LIME+}$ explanations, which improve upon LIME by identifying blocks of time (rather than individual timesteps) during which sensors contribute strongly to an instance's classification, allowing for more temporal and intuitive explanations. We selected these techniques

based on their XAI property of being model agnostic and applicable to any black-box model used for smart home activity recognition, as well as their ability to provide intuitive outputs that can be translated into natural language explanations for end users. Given the resulting four explanation types, we compare their performance along multiple relevant metrics.

3.3 Evaluating XAI-Model Based Explanations

In order to understand the effectiveness of each of the above explanation types, we first evaluate the accuracy of the XAI-model based explanations in explaining smart home activities. To achieve this, we develop a classification rule set that determines whether an explanation is accurate or inaccurate. In Section 6 we compare explanation accuracies across each explanation type, as well as study the computational costs and reported feature distributions of each model.

Second, we analyze the effectiveness of each explanation type in helping end users understand a smart home's activity recognition label. Selecting only accurate explanations, we then conduct a user study to evaluate user preferences toward explanation types and their perceptions of the system's capabilities. Note that we use only accurate explanations in our user study evaluation in order to control for explanation type without confounding results by mixing models with varying levels of accuracy. We present this study in Section 7.

4 ACTIVITY RECOGNITION FOR EXPLAINABLE SMART HOMES

Activity recognition represents the basis for the work presented in this paper. We adopt a state-of-the-art approach [31] and adapt it towards our application scenario. In this section, we introduce the multivariate, fixed interval data representation we use for activity recognition. We then describe the activity recognition model, specifically the adaptations we introduced for our application domain. The AR model is foundational for our work, and as such we evaluate its overall effectiveness for recognizing relevant activities from analyzing sensor data as they are captured in a typical smart home.

4.1 Data Processing

Recall (from Section 3.1) that $S = \{(s_1, v_1, k_1), ..., (s_t, v_t, k_t)\}$ represents the sensor event sequence obtained from a smart home over t timesteps. An example of such a sequence is:

$$(M024, 1, 03:38:28)$$
$$(M024, 0, 03:45:17)$$
$$(D002, 121.4, 03:50:01)$$

where the first value encodes the identity of the sensor, the second value corresponds to the value of the sensor, and the third value encodes the time of the event. The value encoded by each sensor depends on the sensor type; for example, motion sensors return binary values (lines 1 and 2), while distance sensors mounted near the door may provide real-valued output (line 3).

Given the above, we reformat S into a fixed-interval, multivariate representation that is better suited for generating explanations for time-series data. Specifically, instead of recording sensor event changes, we use a representation that encodes the explicit value of each sensor at each timestep. Thus, given M environmental sensors and a duration of T timesteps, we construct a $T \times M$ matrix S' such that $S'[t][m]$ represents the value of sensor m at timestep t. Through this multivariate representation, we are able to preserve both the temporal context (i.e., the duration of time between sensor events) and sensor events for an activity.

Additionally, characterizing smart home data in a representation that is interpretable to humans is crucial for generating human understandable explanations [47]. Binary sensors already possess such interpretability as they only represent a single "on" or "off" state. However, continuous sensor values do not inherently possess interpretability and often require a post-hoc analysis to understand its patterns and meaning. Therefore, to
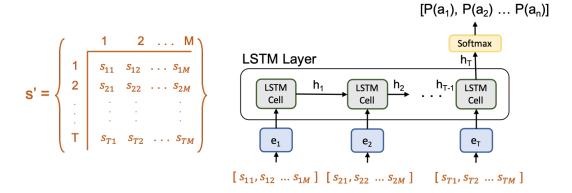
Fig. 2. LSTM model architecture used for activity recognition in smart homes, where S' represents the input data into the LSTM model, and the output is a probability distribution over all candidate activities.

characterize smart home data in an interpretable manner, we discretize continuous sensor values into categorical values that represent the defining pattern observed over an interval (e.g., 1 minute). For example, in the case of the distance sensor above, we categorize its output as "door open". Similarly, if the temperature fluctuated within a one minute interval but increased at the end of the interval, then we categorize the temperature to have been "increased".

4.2 Model Architecture

Figure 2 shows our adapted uni-LSTM model from Liciotti et al. [31]. The LSTM model takes in as input S', where each $S'[t][1:M]$ represents a sequence of sensor events at a particular timestep t. Each sequence $S'[t][1:M]$ is first embedded by an embedding layer e_t before being passed into the LSTM layer. The output of the LSTM layer is a hidden state h_T which is passed through a dense layer with a softmax activation to obtain a probability distribution $\{P(a_1), P(a_1)...P(a_n)\}$ over all activities in A. The most probable activity $a \in A$ is selected as the output of the LSTM model.

4.3 Dataset

The explorations presented in this paper are generic for smart home application scenarios. Yet, for practical considerations our developments and evaluations are based on an existing, established benchmark dataset from the field: the CASAS Milan dataset [9]. The Milan dataset contains sensor data collected from a smart home over a period of 92 days. Figure 3 shows the home layout and locations of the set of 33 sensors in the home. The set of sensors include 3 door sensors, 28 motion sensors and 2 temperature sensors. The Milan dataset includes both the start and end time of an activity and includes data for 15 activities of daily living (ADL): *Bed_to_Toilet, Chores, Desk_Activity, Dining_Rm_Activity, Eve_Meds, Guest_Bathroom, Kitchen_Activity, Leave_Home, Master_Bathroom, Meditate, Watch_TV, Sleep, Read, Morning_Meds, Master_Bedroom_Activity.*

To remain consistent with Liciotti et al. [30], we similarly map the 15 ADL activities onto 10 activites. Particularly, our activity set A = {*Other, Work, Take medicine, Sleep, Relax, Leave home, Eat, Cook, Bed to toilet, Bathing*}. Additionally, while the duration of an activity can vary, we set the number of timesteps representing an activity to be 30 one-minute intervals, after verifying that most activities under the Milan dataset are under 30 minutes, and that utilizing a larger duration for an activity does not affect the activity recognition performance. Thus from

Fig. 3. Home Layout and Locations of Sensors utilized in the CASAS Milan Dataset, where door sensors are designated as D### (solid green boxes), motion sensors as M### (red circles) and temperature sensors as T### (solid yellow boxes). Each red dot represents a motion sensor that can detect motion in a localized area, whereas each radiating red area represents a motion sensor that can detect motion over the indicated area.

our data processing methodology (Section 4.1), our final dataset D, includes 3,298 samples of activities $a \in A$, each represented by S', array of size T timesteps $\times M$ sensors, where $T = 30$ and $M = 33$.

4.4 Model Training

To train our LSTM model, we utilize K-Fold cross validation with 10 stratified folds, where each fold preserves the distribution of activity classes present in our dataset D. Each fold includes a training set d_{train} and testing set d_{test}, where d_{train} includes 2,968 samples of activities and d_{test} includes 330 samples of activities. To evaluate each fold, we utilize a validation set d_{val} that is split from d_{train}, and includes 594 samples. We leverage early-stopping to train our LSTM model. As a result, our model trains for an average of 25 epochs. We train with a batch size of 64 and our LSTM has a hidden state size of 64. We utilize a sparse categorical cross entropy loss via Adam with a 0.001 learning rate.

4.5 Model Evaluation

Figure 4 and Table 1 illustrate the performance of our LSTM model across all folds. On average, the LSTM model has an average recall rate of 0.67, precision of 0.73, and F1 score of 0.69 in classifying the 10 ADL. Specifically, we notice that the 'Bathing' and 'Sleep' activity have the highest recall rate, whereas 'Eat' has the lowest recall rate. We notice that these results correlate with the frequency of each activity class, with the exception of activity class 'Other". While 'Other' is the most frequent activity class in the dataset, it has a lower recall rate compared to 'Bathing' and 'Sleep' which are the next most frequent activities. We suspect this is due the variability in the 'Other' activity itself as it can represent a variety of different activities.

5 XAI FOR ACTIVITY RECOGNITION

Our work utilizes XAI methods, namely LIME, SHAP, and Anchors, each of which are model-agnostic and can explain any black-box classifier. In our explorations, we utilize these XAI models to generate explanations for
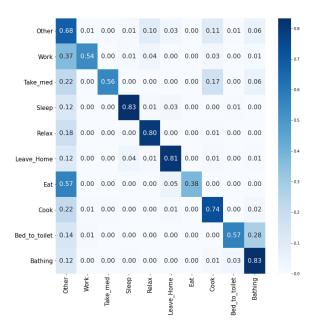
Fig. 4. Confusion matrix our adapted LSTM network architecture, trained on the CASAS Milan dataset, where the y-axis denote the ground truth activities, and the x-axis denote predicted activities.

Table 1. Model performance results, evaluated on CASAS Milan Dataset.

Activity	Precision	Recall	F1-Score
Other	0.71	0.68	0.70
Work	0.71	0.54	0.61
Take medicine	0.74	0.56	0.63
Sleep	0.84	0.83	0.83
Relax	0.72	0.80	0.76
Leave home	0.72	0.81	0.76
Eat	0.67	0.38	0.48
Cook	0.69	0.74	0.71
Bed to toilet	0.64	0.57	0.60
Bathing	0.81	0.83	0.82
Average	0.73	0.67	0.69

the predictions of our LSTM AR model and consequently explain a predicted activity. In order to keep our formulations consistent with the general XAI field, we let f represent a model and $f(x)$ represents the predicted class of an input x. The output of each XAI model, O, represents the set of features that best explain $f(x)$. Note that a feature in our application represents the triple (s_i, v_i, k_i) as defined in Section 4 for the three XAI models under investigation. Additionally, x corresponds to S' as defined in Section 3.1. Table 2 provides example explanations $\mathcal{E}$ for each XAI model.

In this section, we present technical overviews of the three XAI models and how we utilized them to generate explanations for activity recognition in smart home scenarios. We also introduce an extension of LIME, LIME+, through which we explore how the temporal nature of explanations affects user perception of the explanation's value. We utilize the output features of each XAI model to generate associated natural language explanation $\mathcal{E}$ that are understandable by end users.

5.1 LIME

To find a set of features belonging to classification input x that best explains the black-box's prediction $f(x)$, the Local Interpretable Model-agnostic Explanations (LIME) algorithm by Ribeiro et al. [47] trains an interpretable surrogate model g to approximate $f(x)$ in the locality of instance x. Interpretable surrogate models describe the class of models which can be described as inherently interpretable, such as decision trees or linear models [3]. To approximate the decision making of the black-box model for an input x, LIME fits g to a new dataset containing samples $\{z_1...z_n\}$, which are perturbed from x, the instance being explained. The result is a trained surrogate model which can provide a local explanation faithful to $f(x)$, but it is not guaranteed to generalize to other predictions of the same class.

Table 2. Example explanations generated by each of the deployed XAI models.

Type	Activity 1: Leaving Home	Activity 2: Cooking	Activity 3: Relaxing
LIME	The activity is 'leaving home' because at 14:01 the front door was open and movement near front door was detected, and then 10 minutes later the thermostat near the kitchen read moderate temperatures.	The activity is 'cooking' because at 17:45 the thermostat near the kitchen read high temperatures, 5 minutes later the thermostat near the bathroom read moderate temperatures and then a minute later the thermostat near the bathroom read moderate temperatures.	The activity is 'relaxing' because at 17:11 the thermostat near the kitchen read moderate temperatures, 10 minutes later the thermostat near the kitchen read high temperatures and then 10 minutes later the thermostat near the bathroom read high temperatures.
LIME+	The activity is 'leaving home' because at 14:01 the thermostat near the kitchen read moderate temperatures for 15 minutes, the front door was open for 6 minutes, and movement near front door was detected for 8 minutes.	The activity is 'cooking' because at 17:44 the thermostat near the kitchen read high temperatures for 7 minutes, the thermostat near the bathroom read moderate temperatures for 7 minutes, and the pantry door was open for 4 minutes.	The activity is 'relaxing' because at 17:03 movement in the TV room was detected for 6 minutes, 18 minutes later the thermostat near the kitchen read high temperatures for 2 minutes and then 10 minutes later the thermostat near the bathroom read high temperatures for 19 minutes.
Anchors	The activity is 'leaving home' because at 14:01 the front door was open, 2 minutes later the coat cabinet door was open, and then 12 minutes later the front door was open.	The activity is 'cooking' because at 17:45 movement near the pantry was detected, 4 minutes later movement near the bathroom sink was detected and then a minute later movement in the living room was not detected.	The activity is 'relaxing' because at 17:08 movement on the TV room couch was detected and 24 minutes later movement in the living room was not detected.
SHAP	The activity is 'leaving home' because at 14:01 the front door was open, 13 minutes later movement in the living room was not detected and then a minute later the front door was open.	The activity is 'cooking' because at 17:33 movement on the TV room couch, 11 minutes later movement at the kitchen entrance was detected and then 6 minutes later movement in the kitchen was detected.	The activity is 'relaxing' because at 17:23 the thermostat near the kitchen read high temperatures, 9 minutes later movement on the TV room couch was detected and movement in the living room was not detected.

Equation 1 defines LIME's objective function and how the output of LIME, O_{LIME}, is derived, where O_{LIME} represents the top b contributing features that can explain $f(x)$:

$$O_{LIME}(x) = argmin_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

Particularly, π_x defines the proximity measure, or how close samples $\{z_1...z_n\}$ are to x. $\mathcal{L}(f, g, \pi_x)$ expresses the mean squared error between $f(x)$ and $g(x)$ weighted by π_x and $\Omega(g)$ defines the complexity of g (e.g., the depth of a decision tree or the number of non-zero weights of a linear model). As such, LIME learns a surrogate model g from a set of surrogate models G that best approximates f in the local region defined by π_x. The outputs $O_{LIME} = \{o_{lime_1}, o_{lime_2}..o_{lime_b}\}$ represent the top b features that explain the prediction $f(x)$. The top b features are found via Lasso [57] with their corresponding contribution weights calculated via Least Squares.

To generate LIME explanations for smart home activity recognition, we restrict the set of surrogate models G to represent linear models such that $g(z') = w_g \cdot z'$. We ensure interpretability of O_{LIME} by minimizing the model complexity hyperparameter, $\Omega(g)$. In our application, $\Omega(g)$ controls the number of top contributing features,
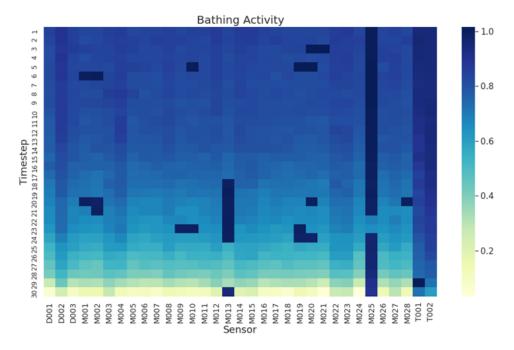
Fig. 5. Distribution of LIME's feature contributions for an instance of 'Bathing'. Note how M013 (bathroom motion sensor) and M025 (closet motion sensor) show strong contribution to the instance's classification over many consecutive timesteps.

b, outputted by O_{LIME}; we let b be set to 3 features. The outputs of LIME, O_{LIME}, are then mapped to natural language explanations $\mathcal{E}_{LIME}$, as described in Section 5.5.

5.2 LIME+

From the outputs of O_{LIME}, we observe that in instances of longer activities $\mathcal{E}_{LIME}$ identifies features corresponding to the same sensor at multiple consecutive timesteps. To further investigate this phenomenon, we analyze the outputs of O_{LIME} across all M features used by the black box model to make a prediction. Figure 5 illustrates the relative contributions of all M features for an instance of the activity 'bathing'. We observe that the activity recognition model recognizes the importance of sensor events *over a period of timesteps* rather than over a single timestep for given activities. Specifically, we observe that M013 (bathroom motion sensor) and M025 (closet motion sensor) contribute strongly to the classification 'bathing' over many consecutive timesteps, which, however, is not captured by O_{LIME}. We posit that it is more unnatural to explain that a sensor event is important towards an activity because of a single timestep rather than a duration of timesteps in the case of longer activities. For example, it is more natural to explain that a resident is sleeping because he/she were present in the bedroom over a period of time rather than only because he/she was present at a single moment in time.

To characterize the duration of a sensor event to an end user, we extend the capabilities of LIME to develop LIME+. The feature outputs of O_{LIME+} are represented by (s_i, v_i, k_i, d_i), which includes an additional duration parameter d_i that describes the duration of a sensor event s_i that best explains instance x. Including the durational importance in a LIME+ explanation is motivated by the desire to provide an increased level of context to end users. Below we describe how we obtain d_i.

We first adapt the original LIME objective function (Equation 1) by removing the model complexity parameter, $\Omega(g)$, from Equation 1:

$$O_{LIME+}(x) = Sort(argmin_{g \in G} \mathcal{L}(f, g, \pi_x)) \tag{2}$$

Removing $\Omega(g)$ allows us to understand the relative contributions of *all M* features used by the black box model f to make a prediction $f(x)$ (similar to Figure 5). From this information, we utilize a *Sort* function to sort the outputs of O_{LIME+} by two criteria: sensor contribution as well as sensor uniqueness.

To find d_i for a top sensor s_i at timestep k_i, we first calculate the average contribution of s_i as well as its standard deviation. The number timesteps consecutive to k_i that are within two standard deviations, describe the duration of sensor importance d_i. Additionally, to ensure end user interpretability, we only focus on the top three sorted features in O_{LIME+} to generate E_{LIME+} explanations.

5.3 Anchors

Anchors by Ribeiro et al. [48] explains a prediction $f(x)$ by finding a high-precision rule R called Anchors which represent local, sufficient conditions that explain $f(x)$. A rule is composed of a feature and its value. The algorithm utilizes a perturbation-based approach to find a set of rules. In order to search for a best candidate rule, Anchors formulates a Multi-Arm Bandit problem [27] commonly utilized in reinforcement learning. If there are two or more best candidate Anchors (rules) found, then Anchors outputs the one with the highest coverage.

Formally, Equation 3 and 4 define how the outputs $O_{Anchors}$ are calculated.

$$O_{Anchors} = \max_{R \; s.t. \; P(prec(R) \geq \tau) \geq 1 - \delta} cov(R) \tag{3}$$

$$cov(R) = E_{D_{(z)}}[R(z)] \tag{4}$$

Specifically, $prec(R)$ is the precision of the data satisfying rule R, τ is the precision threshold, δ is used to guarantee that rule R has at least δ percent probability of precision above τ, and $cov(R)$ is the coverage of R. The measure $cov(R)$ represents the percentage of data that satisfied R in a perturbation space D. If $O_{Anchors}$ includes more than three features to explain $f(x)$, we select the top three features, in order of appearance, to generate natural language explanations for end users.

5.4 SHAP

SHapley Additive exPlanations (SHAP) by Lundberg et al. [33] explains a prediction $f(x)$ by calculating Shapley values for all M features used by the black box model. The Shapley Value for a feature j represents its marginal contribution to the overall prediction of $f(x)$ [52]. Formally, Equation 5 defines how the outputs O_{SHAP} are defined. Specifically, M represents the number of input features, x represents the input vector, and ϕ_j represents the Shapley Value for a feature j.

$$O_{SHAP} = \phi_0 + \sum_{j=1}^{M} \phi_j x_j \tag{5}$$

While O_{SHAP} outputs a Shapley value for all M features in order to explain $f(x)$, we utilize only the top three features to generate natural language explanations for end users.

5.5 Explanation Generation

Given the output of the four models above, we produce explanations $\mathcal{E}_{LIME}$, $\mathcal{E}_{LIME+}$, $\mathcal{E}_{Anchors}$ and $\mathcal{E}_{SHAP}$ using a template-based approach similar to [17]. Utilizing domain knowledge, we derive three types of explanation templates ET_1, ET_2 and ET_3 that vary in the amount of included information:
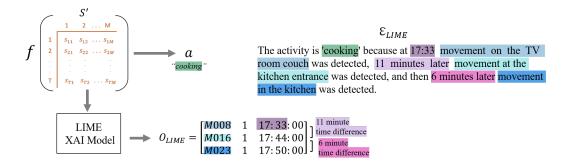
- ET_1: *"The activity is $\langle f(x) \rangle$"*.

Fig. 6. Example process for generating $\mathcal{E}_{LIME}$ from O_{LIME} for an activity instance of 'cooking', where $f(S')$ is used to generate an activity label a, and the LIME XAI model is used to generate O_{LIME}. The template style RT_1 is used to extract the features in O_{LIME} and generate an explanation E_{LIME}.

- ET_2: "The activity is $\langle f(x) \rangle$ because $\langle RT_1 \rangle$"
- ET_3: "The activity is $\langle f(x) \rangle$ because $\langle RT_2 \rangle$"

ET_1 represents a baseline for our work because it simply states the label for the activity, as presented in existing activity recognition systems. ET_2 is used to generate $\mathcal{E}_{LIME}$, $\mathcal{E}_{Anchors}$ and $\mathcal{E}_{SHAP}$ explanations, while ET_3 generates E_{LIME+} explanations via two reasoning templates, RT_1 and RT_2, respectively.

The following summarizes the overall types of information gathered from both the black-box AR model as well as from the XAI models:

- The activity prediction of the black-box classifier, $f(x)$
- Sensor and temporal information (s_i, v_i, k_i) in the features outputs O_{LIME}, O_{LIME+}, O_{SHAP} and $O_{Anchors}$
- Duration of a feature importance, in feature outputs O_{LIME+}

Given this information, we define each explanation reasoning template as follows:

- RT_1 as: {at $\langle k_1 \rangle$ $\langle s_1 \rangle$ $\langle v_1 \rangle$, $\langle k_2 - k_1 \rangle$ minutes later $\langle s_2 \rangle$ $\langle v_2 \rangle$, and $\langle k_3 - k_2 \rangle$ minutes later $\langle s_3 \rangle$ was $\langle v_3 \rangle$}.
- RT_2 as: {at $\langle k_1 \rangle$ $\langle s_1 \rangle$ $\langle v_1 \rangle$ for $\langle d_1 \rangle$, $\langle k_2 - k_1 \rangle$ minutes later $\langle s_2 \rangle$ $\langle v_2 \rangle$ for $\langle d_2 \rangle$, and $\langle k_3 - k_2 \rangle$ minutes later $\langle s_3 \rangle$ was $\langle v_3 \rangle$ for $\langle d_3 \rangle$}.

Figure 6 presents an example of RT_1 being applied in practice to generate $\mathcal{E}_{LIME}$. Note that in some cases we utilize small variations on these templates to improve the readability and sentence structure of explanations.

6 EVALUATION: ANALYSIS OF XAI EXPLANATIONS

In this section, we explore RQ3 and evaluate the accuracy of each XAI-model explanation type for explaining smart home behavior. To do so, we first introduce a classification rule set that describes how to classify an XAI-model explanation as accurate or inaccurate, which we then utilize in our experimental evaluation. We additionally compare and contrast each XAI model's explanations by analyzing the types of sensors events and their values utilized for each explanation type, as well as report the computational cost of running each model. Our analysis is based on the CASAS Milan scenarios [9].

6.1 Methodology: Determining Accurate vs Inaccurate Explanations

We use a set of classification rules to classify each explanation as "accurate" or "inaccurate". Note that this is not a perfect classification, and our definition of accuracy does not in itself pertain to whether the explanation fully captures the inner workings of a complex black-box classifier. Rather we define accuracy in this context as whether an explanation is *sensible* or *credible*. For example, when explaining the classification of a 'Bathing'

Table 3. Example explanations classified as accurate and inaccurate using the classification rules derived for CASAS Milan.

Type	Accurate Explanations	Inaccurate Explanations
$\mathcal{E}_{LIME}$	The activity is 'bathing' because at 15:24 movement in the hallway was detected and then a minute later movement in the shower area was detected.	The activity is 'sleep' because between 01:40 - 01:42 the thermostat near the kitchen read low temperatures, and then 21 minutes later the thermostat near the bathroom read low temperatures
$\mathcal{E}_{LIME+}$	The activity is 'bathing' because at 15:24 movement in the shower area was detected for 2 minutes, movement in the hallway was detected for 2 minutes, and movement near the bathroom was detected for 2 minutes	The activity is 'sleep' because at 01:35 the thermostat near the kitchen read low temperatures, and the coat cabinet door was closed for 27 minutes and then minutes later the thermostat near the bathroom read low temperatures for 26 minutes
$\mathcal{E}_{Anchors}$	The activity is 'bathing' because at 15:24 movement in the shower area was detected, and movement near the front door was not detected	The activity is 'sleep' because at 1:35 movement in living room was not detected, 6 minutes later movement near pantry was not detected and 4 minutes later movement in hallway was not detected
$\mathcal{E}_{SHAP}$	The activity is 'bathing' because at 15:24 movement in the shower area was detected, and movement near the bathroom sink was detected	The activity is 'sleep' because at 01:35 the pantry door was closed, and 37 minutes later movement near the fridge was not detected and the thermostat near the kitchen read low temperatures

activity, we would find it credible if sensors near or in the bathroom were referenced. We would not find it credible if only sensors in the kitchen were referenced. This mapping between activities and sensors that we find credible constitutes an approximation that we use to measure explanation accuracy for our activity set.

To determine the mapping, we hand-selected sensors whose location in the smart home is proximal to the region of the activity. To classify a particular explanation $\mathcal{E}$ as accurate or not, we consider the b features that make up the output O of an XAI algorithm. If *at least* one of the features in O is in the list of sensors for the model-selected label $a \in A$, then we consider the explanation accurate. If none of the features in O are in the validation set of sensors, then it is labeled as inaccurate.

Figure 7 presents the mapping from activities to sensors that we use for the Milan dataset. While this approach constitutes a heuristic, we found it to work very well in practice. Table 3 presents examples of both accurate and inaccurate explanations from all four XAI algorithms.

Activity	Sensors
Bathing	M013, M017, M018
Bed To Toilet	M021, M020, M028, M019, M024, M013, M017
Take Medicine	M012, M016, M022, M014, D003, M023, M015, M003
Leave Home	M001, M002, D001
Work	M007
Sleep	M020, M028, M019, M024, M021
Cook	M023, M015, D003, M014, M022, M012, M016
Relax	M026, M006, M004, M027, M005

Fig. 7. Classification process for determining inaccurate and accurate explanations using the Milan CASAS dataset.

6.2 Results

To analyze explanation accuracy of each model, we measure the percentage of explanations that are classified accurately (*EAcc%*) using the classification rule set introduced above. Note that a total of 167 total explanations were classified, and Figure 8 illustrates the (*EAcc%*) for all XAI models (blue) as well as the computation time required for each model to generate an explanation for a single activity instance (yellow). For our analysis, we utilize a one-way Analysis of Variance (ANOVA) with a Tukey post-hoc test since our data follows a normal distribution
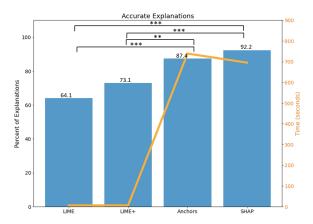
Fig. 8. Blue: percentages of explanations classified as accurate via our classification rules for the CASAS Milan Dataset. Yellow: computational times required to generate an explanation for one activity instance.
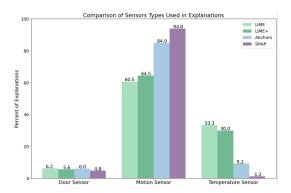

Fig. 9. Percentages of explanations, for each XAI model, that utilize the door sensors, motion sensors and temperature sensors from the CASAS Milan dataset.
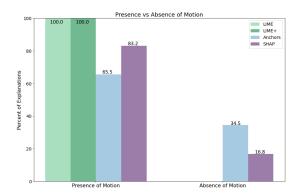

Fig. 10. Percentages of explanations, for each XAI model, that utilize the absence of motion and presence of motion.

(Shapiro-Wilk's Test, p<0.001), and includes homoscedasticity of error variances (Brown-Forsythe Levene's Test, p<0.001). We additionally examine the distribution of sensor types and their values across explanations from each XAI model to understand how these XAI models differ in explanation content, as well as the computational efficiency of each XAI model.

6.2.1 Analyzing Accuracy & Computational Efficiency. The ANOVA determined a significant difference between the percent of XAI-based explanation accuracies (F(3,664)=18.31, p<0.001). We see that $\mathcal{E}_{SHAP}$ have the highest *EAcc%*, while $\mathcal{E}_{LIME}$ have the lowest *EAcc%*. Specifically, we observe that $\mathcal{E}_{SHAP}$ (t(664)=-6.57, p<0.001) and $\mathcal{E}_{Anchors}$ (t(664)=5.45, p<0.001) have a statistically greater *EAcc%* than $\mathcal{E}_{LIME}$. Similarly, $\mathcal{E}_{SHAP}$ (t(664)=-4.48, p<0.001) and $\mathcal{E}_{Anchors}$ (t(664)=3.36, p<0.01) have a statistically greater *EAcc%* than $\mathcal{E}_{LIME+}$. These results show that

both SHAP and Anchors generate significantly higher percentages of accurate explanations compared to the LIME and LIME+ XAI models, with the SHAP XAI model generating the most percentages of accurate explanations.

We also report the computation time required to generate an explanation for one activity instance (yellow in Figure 8) using a quad-core Intel i7-6700K CPU @ 4.00GHz with 32GB RAM. We observe that $\mathcal{E}_{LIME}$ and $\mathcal{E}_{LIME+}$ have the shortest computation (5.91 seconds), while $\mathcal{E}_{Anchors}$ explanations have the longest computation (739.1 seconds). We additionally see that $\mathcal{E}_{SHAP}$'s computation time (695.2 seconds) is slightly shorter than Anchors, but similarly longer than $\mathcal{E}_{LIME}$ and $\mathcal{E}_{LIME+}$. While we do not claim superiority of one XAI model over another based on computation time, we demonstrate that with increased accuracy of explanations (*EAcc%*), comes an increased cost in computation time as well.

6.2.2 Analyzing Explanation Content. We additionally analyze the types of information presented from each XAI model to understand the similarities and differences between the explanation content by each XAI model. From Figure 9, we observe that the motion sensors events (i.e., movement in a particular location of the home), is the most utilized sensor type by each XAI model for describing all activities. However, we see that $\mathcal{E}_{LIME}$ explanations utilize motion sensors less frequently than $\mathcal{E}_{Anchors}$ and $\mathcal{E}_{SHAP}$, while frequently using temperature sensors events (i.e., temperature is at given setting) compared to $\mathcal{E}_{Anchors}$ and $\mathcal{E}_{SHAP}$. This shows that the LIME XAI model tends to focus on the temperature changes as an important contribution towards an activity more than the SHAP and Anchors XAI models. Furthermore, we observe from Figure 10 that $\mathcal{E}_{Anchors}$ and $\mathcal{E}_{SHAP}$ leverage the absence of movement in a location in addition to the presence of movement, whereas $\mathcal{E}_{LIME}$ and $\mathcal{E}_{LIME+}$ only leverage the presence of movement in a location. These results imply that the SHAP and Anchors XAI models view both the 'on' and 'off' states of sensors as important factors towards an activity whereas LIME and LIME+ XAI models focus solely on the 'on' states.

7 USER STUDY: COMPARISON OF XAI MODELS

In this section, we describe the user study in which we validated each explanation type. These results contribute to the analysis of RQ1, where we seek to evaluate whether users prefer natural language explanations or simple activity labels, and to RQ2, where we explore whether explanations give users more confidence in the activity recognition model.

7.1 Study Design

In the study, participants were presented with a scenario in which they played the role of a remote caregiver attempting to monitor the well-being of an elderly patient with Dementia named Pat. For each validation scenario participants were informed of Pat's activity (e.g., "Pat is currently bathing."). They were then presented with five explanations of the activity ($\mathcal{E}_{LIME}$, $\mathcal{E}_{LIME+}$, $\mathcal{E}_{Anchors}$, $\mathcal{E}_{SHAP}$ and $\mathcal{E}_{Base}$). Note that $\mathcal{E}_{Base}$ is the baseline condition in which no explanation is given, only the activity label is repeated (e.g "The activity is 'Bathing'"). Participants were then asked to respond to a number of questions designed to evaluate the effectiveness of the explanations.

The study was designed as a within-subjects study, and each participant saw and compared all explanation types. In order to control for explanation accuracy, and ensure that we are independently analyzing accuracy and explanation type, the user study included only example explanations that were selected as 'accurate' through the process described in Section 6.1. Specifically, we evaluated 93 activities for which all XAI-models generated accurate explanations. To reduce participant fatigue, each participant was asked to evaluate at most 8 activities and their explanations; the 93 activities were randomly separated into 12 groups, in which groups 1-11 included explanations for 8 activities, and group 12 included explanations for the remaining 5 activities. Additionally, to reduce participant bias, the order in which explanations were presented was randomized in each question.
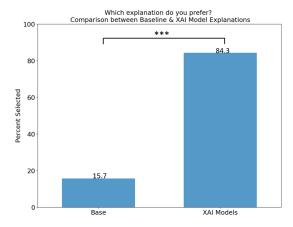
Fig. 11. Comparison of user preferences *(PId)*, showing the preferences between a baseline explanation, $\mathcal{E}_{Base}$, or an XAI-Model based explanation. Statistical significance reported as: *** p<0.001.
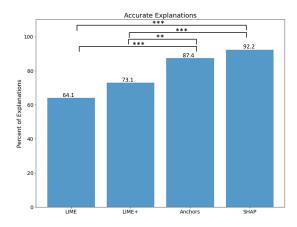

Fig. 12. Comparison of user preferences *(PId)*, showing the preferences among explanations from each XAI model. Statistical significance reported as: *** p<0.001.

7.2 Participants

We recruited 144 individuals from Amazon's Mechanical Turk, which included 71 males and 73 females, all of whom were 18 years or older (M=38.3, SD = 10.8). Each of the 12 groups of activities included 12 participants, thus each activity-explanation pair was reviewed by 12 individuals. Participants in groups 1-11, those who evaluated explanations for 8 activities, took on average 50-60 minutes to complete the study, and were compensated US$7.00. Participants in group 12, those who evaluated explanations for 5 activities, took on average 20-30 minutes to complete the study, and were compensated with US$4.00.

7.3 Metrics

Prior evaluation methods for explanations have examined qualities such as preference [46], perceived accuracy [46], user confidence [15], and adequacy in justification [15]. Thus, we evaluated the effectiveness of each explanation type using similar metrics:

Preference Identification (*PId*): measures percentage of activities for which a given explanation type was preferred over others. Evaluated as a response to the question "Which explanation do you prefer?"

Perceived System Accuracy (*SAcc*): measures participants' self-reported perceptions on the system's ability to identify an activity correctly based on each explanation. Evaluated based on response to the statement "The smart home correctly identified Pat's activity", measured on a 5-point Likert scale between Strongly Agree and Strongly Disagree.

Perceived Confidence (*Conf*): measures the participants' self-reported confidence in the system's tracking abilities. Evaluated based on response to the statement "Given the explanation, I am confident in the smart home's ability to accurately track Pat's activity", measured on a 5-point Likert scale between Strongly Agree and Strongly Disagree.

Perceived Justification Adequacy (*JAdq*): measures participants' self-reported perceptions on how adequately each explanation justifies a given activity. Evaluated based on response to the statement "The smart home provides adequate justification as to why Pat is <activity>", measured on a 5-point Likert scale between Strongly Agree and Strongly Disagree.

7.4 User Study Results

Using the above metrics, we now examine the study results as they relate to our research questions.

7.4.1 RQ1: User Preference. The core research question of our work is whether users prefer natural language explanations over simple activity labels. In the study we presented participants with five statements and asked which explanation they preferred as description of Pat's activity. Figures 11 and 12 summarize the results. As seen in Figure 11, explanation-based statements (grouped across all conditions) were preferred by participants for 84.3% of tested activities compared to the baseline $\mathcal{E}_{Base}$ statements. *Statistical analysis:* Validated by conducting an unpaired t-Test we conclude that XAI-model explanations have a significantly higher *PId* compared to $\mathcal{E}_{Base}$ (t(2158) = -43.7, p<0.001).

We further break down the results by explanation model type in Figure 12. As can be seen, participants preferred results generated by LIME+, Anchors and SHAP over those of LIME, with no strong preference among the top three methods. The distinction between LIME and LIME+ is notable because there are no differences between the two explanation types with respect to which sensory features are being described – both LIME and LIME+ explanations incorporate the same features. What differs is that LIME+ explanations incorporate duration of activities, and LIME explanations do not. Thus, we conclude that communicating sensor activation duration is found to be more intuitive and valuable by users. *Statistical analysis:* We conducted a one-way ANOVA with a post-hoc Tukey test, since our data follows a normal distribution (Shapiro-Wilk's Test, p<0.001) and includes homoscedasticity of error variances (Brown-Forsythe Levene's Test,p<0.001). The ANOVA test determined a significant difference among the XAI-based explanations (F(3,3636)=21.3, p<0.001). Specifically, we see that $\mathcal{E}_{LIME+}$ (t(3636)=-5.67, p<0.001), $\mathcal{E}_{Anchors}$ (t(3636)=5.67, p<0.001) and $\mathcal{E}_{SHAP}$ (t(3636=-7.53, p<0.001) have a significantly higher *PId* compared to $\mathcal{E}_{LIME}$.

Summary: Participants strongly prefer natural language explanations over simple activity labels (84.3% vs 15.7%), indicating that users find significant value in gaining access to information that elucidates the decision making process of the classifier. Out of the four XAI methods we tested, $\mathcal{E}_{LIME}$ is the least preferred explanation type. Additionally, we see no significant difference in user preference among explanations generated by LIME+, Anchors and SHAP. Beyond these results, we analyzed our data for correlations between type of activity (e.g., cooking) and explanation type, and we found no correlations between these factors. As such, we find that explanations generated by these three models perform similarly and are similarly preferred by users. For developers, this indicates that other factors should be considered when selecting among the three models. For example, as reported in Section 6.2, SHAP produces accurate explanations at a higher rate than other methods.

7.4.2 RQ2: User Confidence in Activity Model. In RQ2 we seek to determine whether explanations give users more confidence in the activity recognition system. Specifically, we analyze how each explanation type influences participant perception of the smart home's accuracy and capability via two metrics: *(SAcc)* and *(Conf)*. All statistical analysis for these results utilized a one-way, repeated measures of ANOVA (Shapiro-Wilk's Test, p<0.001 & Brown-Forsythe Levene's Test, p<0.001) with a post-hoc Tukey test.

Figure 13(a) presents the results for the *SAcc* metric, in which we ask participants to read the given explanation and respond whether they agree or disagree with the statement "The smart home correctly identified Pat's activity". Note that only accurate explanations were given to participants. However, the wording of some explanations, particularly in the case of $\mathcal{E}_{LIME}$ and $\mathcal{E}_{LIME+}$, confused participants and we see a significant increase in the number of Disagree and Strongly Disagree responses. This result indicates that, having read the explanation, participants began to doubt that the underlying classification algorithm was working properly, even though it was. An example explanation from LIME that supported this trend was: *"The activity is cooking because at 17:45 the thermostat near the kitchen read high temperatures, 5 minutes later the thermostat near the bathroom read moderate temperatures and then a minute later the thermostat near the bathroom read moderate temperatures"*. SHAP showed
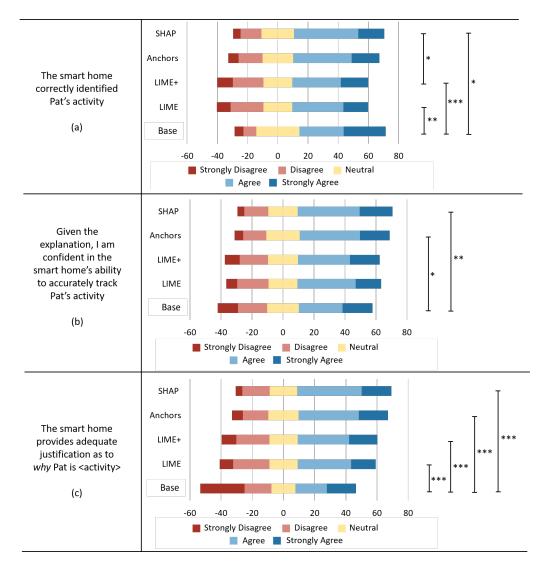
Fig. 13. Summary of user perceptions for each explanation type, specifically looking at users' perceived system accuracy *(SAcc)* (a), users' perceived confidence *(Conf)* (b), and users' perceived justification adequacy *(JAdq)* (c). Statistical significance reported as: * p<0.05, ** p<0.01, *** p<0.001.

statistically significant improvement over other conditions in improving user confidence in the correctness of the AR model, whereas Anchors did not. *Statistical analysis:* The ANOVA test determined a significant difference among the explanation types (F(4,5390)=6.24, p<0.001). Specifically, we observe that $\mathcal{E}_{Base}$ (t(5390)=3.79, p=0.0014) and $\mathcal{E}_{SHAP}$ (t(5390)=-3.08, p=0.017) have a significantly higher *SAcc* rating compared to $\mathcal{E}_{LIME}$. Additionally, we see that $\mathcal{E}_{Base}$ (t(5390)=3.91, p<0.001) and $\mathcal{E}_{SHAP}$ (t(5390)=-3.20, p=0.012) have a significantly higher *SAcc* rating compared to $\mathcal{E}_{LIME+}$.

Figure 13(b) presents results for the *Conf* metric, in which we ask participants to read the given explanation and respond whether they agree or disagree with the statement "I am confident in the smart home's ability to accurately track Pat's activity". Remember, again, that (unknown to them) all activities presented to participants are accurately classified. We observe that for the baseline condition, responses are widely spread across the full response spectrum, indicating that without any information on which to base their decision, participants vary drastically in their confidence in the model. Given explanations, however, we see an increase in overall model confidence. Specifically, users were statistically significantly more confident in the performance of the activity recognition model given SHAP and Anchors explanations compared to the baseline. *Statistical analysis:* The ANOVA test determined a significant difference among the explanation types ($F(4,5390) = 3.77$, $p<0.01$). Specifically, we observe that $\mathcal{E}_{Anchors}$ ($t(5390)=-2.95$, $p=0.027$) and $\mathcal{E}_{SHAP}$ ($t(5390)=-3.45$, $p=0.005$) have significantly higher *Conf* values compared to $\mathcal{E}_{Base}$.

Figure 13(c) presents results for the *JAdq* metric, in which we ask participants to read the given explanation and respond whether they agree or disagree with the statement "The smart home provides adequate justification as to *why* Pat is <activity>". We observe that participants very strongly disagree that the baseline condition provides sufficient justification for the activity label. This result is significant against all four explanation types, with SHAP and Anchors having the highest Agree and Strongly Agree ratings. *Statistical analysis:* The ANOVA test determined a significant difference among the explanations types ($F(5,390)=14.27$, $p<0.001$). Specifically, we observe that $\mathcal{E}_{SHAP}$ ($t(5,390)=-6.71$, $p<0.001$), $\mathcal{E}_{LIME}$ ($t(5,390)=-4.25$, $p<0.001$), $\mathcal{E}_{LIME+}$ ($t(5,390)=-4.72$, $p<0.001$) and $\mathcal{E}_{Anchors}$ ($t(5,390)=-6.33$, $p<0.001$) have significantly higher *JAdq* ratings compared to $\mathcal{E}_{Base}$.

Summary: Our results indicate that users find the activity label alone as inadequate justification for the activity label Figure 13(c), further justifying the need for explanations. We also find that explanations can have a strong positive effect in giving users confidence that the underlying activity recognition system is performing correctly (SHAP and Anchors in Figure 13(b)). However, we also observe that poorly worded explanations, even when associated with a correct activity label, can make users believe that the system is not working correctly (LIME and LIME+ in Figure 13(a)).

8 DISCUSSION

The analysis in Sections 6.2 and 7.4 provided us with answers to our first three research questions. We found that:

- users perceive simple activity labels as inadequate justifications for classification output (Section 7.4.2);
- users overwhelmingly prefer natural language explanations over simple activity labels (Section 7.4.1); and
- users can both lose and gain confidence in the accuracy of the underlying activity recognition model based on the given explanation (Section 7.4.2).

Our final research question, RQ4, is: **which explanation method is ultimately most effective?** To determine the answer to this question we must consider multiple factors, including model accuracy, computational efficiency, and user preference. Across our evaluations, we found that:

- SHAP and Anchors generated the highest percentage of accurate explanations (92.2% and 87.4% compared to 64.1% of LIME), at the cost of higher computational resources (Section 6.2);
- LIME was the least preferred explanation method in a direct comparison (Section 7.4.1);
- LIME and LIME+ explanations caused more users to doubt the correctness of the activity recognition model (Section 7.4.2);
- SHAP and Anchors increased user confidence in the accuracy of the AR model and smart home system (Section 7.4.2); and
- SHAP was more effective in improving user confidence in the correctness of the AR model than Anchors (Section 7.4.2). This may be because SHAP more frequently explained activities in terms of events that

occurred rather than the absence of events (83.2% for SHAP vs 65.5% for Anchors), which users may find more intuitive (Section 6.2).

Taking all of these factors into consideration, we find that the SHAP model results in the most effective explanations. We also note that users did not show a significant or consistent preference between the various explanations when asked to indicate preference (Figure 12), suggesting that possibly a broad spectrum of acceptable explanations may exist. More generally, these findings lead to new research questions in a number of areas:

The benefits of explanations: Our work clearly demonstrates that providing explanations to end users of activity recognition systems is of great value within the domain of smart home automation. As in other areas of XAI, AR explanations serve to elucidate the decision making process of a complex computational model and enable users to gain understanding of which factors most significantly influenced model outcome. Increased user understanding in turn leads to increased confidence in the system. This finding raises many questions for future work. For example, how can we improve computational methods for XAI, particularly as they apply to temporal decision making? Furthermore, does incorporating XAI into deployed systems smart home lead to improved user satisfaction, trust and long-term use?

The harmful effects of explanations: We observed that explanations can in some cases have a harmful effect on user perception of the system. Specifically, we saw that confusing explanations led some participants to doubt the accuracy of the system's activity recognition output even when it was correct. In such cases we see that an unsuitable explanation can be worse than no explanation at all. Today's systems are not designed to reason about the quality of their explanation. An important direction for future work is to consider how XAI models can communicate confidence in how well they are able to explain a particular input.

The content of an explanation: In this work we studied different computational techniques for identifying the underlying state features that contribute to an AR classification. Given such features, we used a template approach for generating explanations that sought to present those features in an interpretable way. Thus, we studied what information *could* be encoded in an explanation, but did not exhaustively study what information *should* be included. In other words, we did not try to find the "perfect" explanation (if one exists) for activity recognition in smart homes. Thus, an open question remains of what information would be ideal to include, such that future computational XAI research can be guided toward techniques that extract such information.

Differences in user preferences: Our results indicate that users have a broad range of opinions relating to any given explanation, suggesting that personal preference may play a significant factor. Existing XAI methods typically provide one type of explanation for all users. An interesting question for future work is to consider how explanations could be customized, with respect to length, type, or level of detail, for individual users.

Longitudinal effects of XAI systems: Due to the relative novelty of XAI systems, and their absence in prior activity recognition systems, we have no findings regarding the long-term effects that providing explanations will have on users. While it is encouraging to think that explainability will lead to greater trust in a system, it is also possible that users will find explanations boring, distracting or annoying.

Implications for Activity Recognition: An important research question that remains unanswered is how activity recognition systems can best facilitate explainability – are there changes that can be made to the way features are represented or models are learned that would improve the transparency of AR models? Or, alternatively, can XAI lead to improvements in activity recognition itself? Is it possible to leverage the understanding that users gain through XAI to enable users to train, teach, or instruct AR systems (through active learning or demonstrations) in order to further customize smart home systems and improve performance?

These, and many other, research questions remain open at the intersection of activity recognition and explainable AI research. This paper serves to establish the foundation that we hope will lead to many further advances in this research area.

9   CONCLUSION

While smart home systems have the potential to provide services that can help improve the quality of life of their occupants, they are not perfect systems. Inconsistencies in their underlying activity recognition systems can exist and may cause an end user to wonder "Why did the smart home do that?"

In this work, we leverage the field of Explainable AI to contribute XAI methodologies towards explainable activity recognition such that we can provide meaningful explanations to end users about a smart home's activity recognition. To do so, we first preserve the multivariate nature of smart home sensor data (both timestep and sensor event information) and leverage state-of-the-art XAI models–LIME, SHAP, and Anchors–to extract the most contributing feature information towards an activity recognition label. In this process, we introduced an improvement to LIME, which we call LIME+, that additionally provides the durational importance of a sensor event. We utilize the outputs of each XAI model to generate template-based natural language explanations. We evaluated each XAI explanation type in two manners. First, we analyzed the explanation accuracy from all XAI models as well as their computational efficiency. We then evaluated the XAI explanation types with end users to understand how effective they are in helping end users understand a smart home's activity recognition.

From our results, we find that SHAP based explanations produce the most accurate explanations (92% of explanations) compared to the other XAI models. We also learned that SHAP based explanations are most effective in increasing users' confidence in the correctness of an AR model and that users significantly prefer XAI-based natural language explanation over a simple activity label.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Gregory D Abowd, Aaron F Bobick, Irfan A Essa, Elizabeth D Mynatt, and Wendy A Rogers. 2002. The aware home: A living laboratory for technologies for successful aging. In *Proceedings of the AAAI-02 Workshop "Automation as Caregiver*. 1–7.

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[4] Fadi Al Machot, Ahmad Haj Mosa, Mouhannad Ali, and Kyandoghere Kyamakya. 2017. Activity recognition in sensor data streams for active and assisted living environments. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2017), 2933–2945.

[5] Iftikhar Alam, Shah Khusro, and Muhammad Naeem. 2017. A review of smart TV: Past, present, and future. In *2017 International Conference on Open Source Systems & Technologies (ICOSST)*. IEEE, 35–41.

[6] Hande Alemdar, Halil Ertan, Ozlem Durmaz Incel, and Cem Ersoy. 2013. ARAS human activity datasets in multiple homes with multiple residents. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 232–235.

[7] UABUA Bakar, Hemant Ghayvat, SF Hasanm, and Subhas Chandra Mukhopadhyay. 2016. Activity and anomaly detection in smart home: A survey. *Next Generation Sensors and Systems* (2016), 191–220.

[8] Serge Thomas Mickala Bourobou and Younghwan Yoo. 2015. User activity recognition in smart homes using pattern clustering applied to temporal ANN algorithm. *Sensors* 15, 5 (2015), 11953–11971.

[9] Diane J Cook, Aaron S Crandall, Brian L Thomas, and Narayanan C Krishnan. 2012. CASAS: A smart home in a box. *Computer* 46, 7 (2012), 62–69.

[10] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[11] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for Robot Failures: Generating Explanations That Improve User Assistance in Fault Recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. 351–360.

[12] Daniel Clement Dennett. 1989. *The intentional stance*. MIT press.

[13] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[14] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[15] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2018. Learning to Generate Natural Language Rationales for Game Playing Agents. In *Joint Proceedings of the AIIDE 2018 Workshops*, Vol. 2282. 1.

[16] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[17] Francisco Elizalde, Enrique Sucar, Julieta Noguez, and Alberto Reyes. 2009. Generating explanations based on Markov decision processes. In *Mexican International Conference on Artificial Intelligence*. Springer, 51–62.

[18] Labiba Gillani Fahad and Syed Fahad Tahir. 2021. Activity recognition and anomaly detection in smart homes. *Neurocomputing* 423 (2021), 362–372.

[19] Juliana J Ferreira and Mateus S Monteiro. 2020. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *International Conference on Human-Computer Interaction*. Springer, 56–73.

[20] Rebecca Ford, Marco Pritoni, Angela Sanguinetti, and Beth Karlin. 2017. Categories and functionality of smart home technology for energy management. *Building and environment* 123 (2017), 543–554.

[21] Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. 2019. Explainable AI in industry. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3203–3204.

[22] KS Gayathri, KS Easwarakumar, and Susan Elias. 2017. Probabilistic ontology based activity recognition in smart homes using Markov Logic Network. *Knowledge-Based Systems* 121 (2017), 173–184.

[23] Hemant Ghayvat, S Mukhopadhyay, B Shenjie, Arpita Chouhan, and W Chen. 2018. Smart home based ambient assisted living: Recognition of anomaly in the activity of daily living for an elderly living alone. In *2018 IEEE international instrumentation and measurement technology conference (I2MTC)*. IEEE, 1–5.

[24] Munkhjargal Gochoo, Tan-Hsu Tan, Shing-Hong Liu, Fu-Rong Jean, Fady S Alnajjar, and Shih-Chia Huang. 2018. Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN. *IEEE journal of biomedical and health informatics* 23, 2 (2018), 693–702.

[25] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[26] Stephen S Intille, Kent Larson, J Beaudin, E Munguia Tapia, Pallavi Kaushik, Jason Nawyn, and Thomas J McLeish. 2005. The PlaceLab: A live-in laboratory for pervasive computing research (video). *Proceedings of PERVASIVE 2005 Video Program* (2005).

[27] Emilie Kaufmann and Shivaram Kalyanakrishnan. 2013. Information complexity in bandit subset selection. In *Conference on Learning Theory*. PMLR, 228–251.

[28] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[29] Praveen Kumar and Umesh Chandra Pati. 2016. IoT based monitoring and control of appliances for smart home. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 1145–1150.

[30] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[31] Daniele Liciotti, Michele Bernardini, Luca Romeo, and Emanuele Frontoni. 2020. A sequential deep learning application for recognising human activities in smart homes. *Neurocomputing* 396 (2020), 501–513.

[32] Tania Lombrozo. 2012. Explanation and abductive inference. (2012).

[33] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[34] Stephen Makonin, Lyn Bartram, and Fred Popowich. 2012. A smarter smart home: Case studies of ambient intelligence. *IEEE Pervasive Computing* 12, 1 (2012), 58–66.

[35] Homay Danaei Mehr, Huseyin Polat, and Aydin Cetin. 2016. Resident activity recognition in smart homes by using artificial neural networks. In *2016 4th international istanbul smart grid congress and fair (ICSG)*. IEEE, 1–5.

[36] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[37] Meg E Morris, Brooke Adair, Kimberly Miller, Elizabeth Ozanne, Ralph Hansen, Alan J Pearce, Nick Santamaria, Luan Viega, Maureen Long, and Catherine M Said. 2013. Smart-home technologies to assist older people to live well at home. *Journal of aging science* 1, 1 (2013), 1–9.

[38] Ehsan Nazerfard, Barnan Das, Lawrence B Holder, and Diane J Cook. 2010. Conditional random fields for activity recognition in smart environments. In *Proceedings of the 1st ACM International Health Informatics Symposium*. 282–286.

[39] Tobias Nef, Prabitha Urwyler, Marcel Büchler, Ioannis Tarnanas, Reto Stucki, Dario Cazzoli, René Müri, and Urs Mosimann. 2015. Evaluation of three state-of-the-art classifiers for recognition of activities of daily living from smart home ambient data. *Sensors* 15, 5 (2015), 11725–11740.

[40] Qin Ni, Ana Belén García Hernando, and Iván Pau de la Cruz. 2016. A context-aware system infrastructure for monitoring activities of daily living in smart home. *Journal of Sensors* 2016 (2016).

[41] Mahsan Nourani, Chiradeep Roy, Tahrima Rahman, Eric D Ragan, Nicholas Ruozzi, and Vibhav Gogate. 2020. Don't Explain without Verifying Veracity: An Evaluation of Explainable AI with Video Activity Recognition. *arXiv preprint arXiv:2005.02335* (2020).

[42] Debajyoti Pal, Tuul Triyason, and Suree Funikul. 2017. Smart homes and quality of life for the elderly: a systematic review. In *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, 413–419.

[43] Donald J Patterson, Dieter Fox, Henry Kautz, and Matthai Philipose. 2005. Fine-grained activity recognition by aggregating abstract object usage. In *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*. IEEE, 44–51.

[44] Kun Qian, Marina Danilevsky, Yannis Katsis, Ban Kawas, Erick Oduor, Lucian Popa, and Yunyao Li. 2021. XNLP: A Living Survey for XAI Research in Natural Language Processing. In *26th International Conference on Intelligent User Interfaces*. 78–80.

[45] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. 2005. Activity recognition from accelerometer data. In *Aaai*, Vol. 5. Pittsburgh, PA, 1541–1546.

[46] Stephen J Read and Amy Marcus-Newhall. 1993. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology* 65, 3 (1993), 429.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[49] Vincent Ricquebourg, David Menga, David Durand, Bruno Marhic, Laurent Delahoche, and Christophe Loge. 2006. The smart home concept: our immediate future. In *2006 1st IEEE international conference on e-learning in industrial electronics*. IEEE, 23–28.

[50] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

[51] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[52] Lloyd S Shapley. 2016. *17. A value for n-person games*. Princeton University Press.

[53] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T Wells. 2020. Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*. 23–34.

[54] Emmanuel Munguia Tapia, Stephen S Intille, and Kent Larson. 2004. Activity recognition in the home using simple and ubiquitous sensors. In *International conference on pervasive computing*. Springer, 158–175.

[55] Paul Thagard. 1989. Explanatory coherence. *Behavioral and brain sciences* 12, 3 (1989), 435–502.

[56] Keshav Thapa, Abdullah Al, Zubaer Md, Barsha Lamichhane, and Sung-Hyun Yang. 2020. A Deep Machine Learning Method for Concurrent and Interleaved Human Activity Recognition. *Sensors* 20, 20 (2020), 5770.

[57] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[58] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[59] Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. 2008. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*. 1–9.

[60] Liang Wang, Tao Gu, Xianping Tao, Hanhua Chen, and Jian Lu. 2011. Recognizing multi-user activities using wearable sensors in a smart home. *Pervasive and Mobile Computing* 7, 3 (2011), 287–298.

[61] Vasanth Williams, Jude Immaculate, et al. 2019. Survey on Internet of Things based smart home. In *2019 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 460–464.

[62] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2017. Beyond sparsity: Tree regularization of deep models for interpretability. *arXiv preprint arXiv:1711.06178* (2017).

[63] Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *Interactions* 26, 4 (2019), 42–46.

[64] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. 2019. Interpreting cnns via decision trees. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 6261–6270.

[65] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.

[66] Qingyuan Zhao and Trevor Hastie. 2021. Causal interpretations of black-box models. *Journal of Business & Economic Statistics* 39, 1 (2021), 272–281.