Received 8 January 2026, accepted 1 February 2026, date of publication 6 February 2026, date of current version 11 February 2026.
Digital Object Identifier 10.1109/ACCESS.2026.3661415

RESEARCH ARTICLE

Transformer-Based Explainable Framework for Human Activity Recognition in IoT-Enabled Smart Homes

VISHALINI VIJAYANANTH, (Member, IEEE),
THINAGARAN PERUMAL, (Senior Member, IEEE), AND **ERZAM MARLISAH**
Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor 43400, Malaysia
Corresponding author: Thinagaran Perumal (thinagaran@upm.edu.my)

This work was supported by the Ministry of Higher Education through the Fundamental Research Grant Scheme (FRGS), Universiti Putra Malaysia under Cost Center 5540651 and Grant FRGS/1/2023/ICT08/UPM/02/1.

ABSTRACT The growing integration of the Internet of Things (IoT) devices in smart homes has enabled applications in comfort, energy efficiency, healthcare, and security. A key enabler of these applications is Human Activity Recognition (HAR), which enables smart systems to understand residents' activities and deliver personalised services. While single-resident HAR has been well studied, multi-resident scenarios are more complex due to collaborative and parallel activities, requiring effective modelling of temporal dependencies. Transformer-based models have shown promising results in this context as their self-attention can capture long-range temporal dependencies and allows parallel computation. However, standard attention uses both past and future information, causing information leakage in real-time HAR. It also treats all time steps equally, ignoring the fact that event correlations in smart homes naturally decay over time and require locality bias. Since these systems involve real residents, explainability is also crucial to ensure trust and transparency in activity classification. This paper proposes a Powerformer-based multi-output explainable framework incorporating a Weighted Causal Multi-Head Attention (WCMHA) mechanism with power-law decay to emphasise recent events while maintaining causal order, improving accuracy. A Multi-Output Temporal Local Interpretable Model-agnostic Explanations (MO-TLIME) technique is proposed, which identifies key sensors and time segments influencing each resident's activity. Evaluated on publicly available multi-resident datasets, the proposed model outperforms existing deep learning baselines in both accuracy and efficiency. Statistical validation and ablation studies confirm its effectiveness, while explainability analyses demonstrate that the model focuses on contextually relevant sensor data and temporal segments.

INDEX TERMS Explainable AI (XAI), Internet of Things (IoT), local interpretable model-agnostic explanations (LIME), multi-resident human activity recognition, smart homes, transformer.

I. INTRODUCTION

Exponential technological growth and advancements in the Internet of Things (IoT) have led to the rise of automation, making human lives easier and revolutionising how humans interact with the environment. From simple thermostats to life-saving Magnetic Resonance Imaging (MRIs), progressions in sensing technologies have played a vital role in the development of automation systems by collecting real-time data from various environments, which is further processed and analysed to make informed decisions by the system. In recent years, due to the availability of these sensors at a very low cost, embedding them in living spaces has become a practical choice for consumers, paving the way for the tremendous strides that development in smart homes has undergone. The smart home market worldwide is expected to grow at an average rate of 9.55% per year from 2025 to 2029, resulting in a market volume of USD 250.6 billion by 2029, encouraging innovation and expansion [1]. Smart homes embedded with IoT devices enable diverse applications

The associate editor coordinating the review of this manuscript and approving it for publication was Enrico Zio.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

that improve residents' quality of life across domains such as convenience (automatic lighting and temperature control), security (smart locks and intrusion alarms), energy management (smart meters), and healthcare (fitness trackers and in-home monitoring). Among these, healthcare has gained prominence due to the ageing global population, as elderly individuals increasingly depend on in-home medical assistance. By 2030, one in six people worldwide will be aged 60 or above [2], underscoring the need for smart home–based health monitoring systems. Furthermore, the COVID-19 pandemic has transformed daily living, with remote work and health concerns leading individuals to spend more time at home, reinforcing the role of smart homes as multi-functional spaces that promote well-being, safety, and productivity for all age groups [3]. To implement these meaningful resident-centric applications, Human Activity Recognition (HAR) is essential for identifying residents' activities and deriving contextual insights.

HAR has been an active research area for over two decades [4], with early studies primarily focusing on single-resident environments [5], [6], [7]. In the past few years, research in Multi-resident Human Activity Recognition (MHAR) has grown significantly due to the inherent challenges of accurately identifying activities when multiple residents may perform the same or different activities at the same time or collaboratively, in addition to possible sequential, interleaving, and concurrent activities performed by each individual residing in smart homes [8], [9]. MHAR comprises implementing activity recognition with the ability to recognise the activities accurately and to associate corresponding activities with the respective person who performed them to provide smart services by understanding each resident's requirements. For example, in a living room, one resident may be watching television while another is reading, yet shared motion, light, and ambient sensors capture overlapping activations that complicate accurate activity–resident association.

Transformer-based algorithms have demonstrated strong performance across various domains and have emerged as state-of-the-art for MHAR, owing to their self-attention mechanism, which enables parallel computation and effectively captures long-range temporal dependencies by allowing each time step to attend to all others [10]. However, traditional transformer models use an all-to-all attention mechanism that allows each time step to consider information from all other time steps, including future ones. This can lead to the model accessing future data, violating the natural temporal order required for real-time applications. Furthermore, recent sensor events tend to be more informative for identifying ongoing activities, especially in multi-resident settings where multiple activities may occur simultaneously or in close succession. Since traditional transformers do not enforce causal ordering or sufficiently emphasise local temporal context [11], they often struggle to accurately capture the dynamic and overlapping nature of activities in shared living environments.

Moreover, these techniques involved in recognition are non-interpretable [12] opaque models, offering limited insight into how specific sensor activations influence activity predictions. Explainable AI (XAI) techniques are therefore essential to improve transparency, interpretability, and user trust [13]. Although eXplainable AI (XAI) has made significant progress in single-resident HAR, its adoption in multi-resident scenarios remains limited. In particular, time-segment-based explanations play a vital role in identifying specific parts within the input sequence that most influence the model's decision [14], [15], thereby enhancing transparency and providing more natural, intuitive interpretations in real-world smart home applications. Resident-specific explanations further help the system identify subtle distinctions between overlapping activities, enhancing trust and aiding in the analysis and correction of misclassifications. Motivated by these challenges, this study aims to address the following objectives:

- To develop a multi-output MHAR model that preserves temporal causality while effectively modelling both local and global temporal dependencies in residents' activities, thereby overcoming the limitations of conventional Transformer-based MHAR approaches.
- To introduce a resident-specific, time-segment–based explainability technique that provides explanations by identifying influential temporal segments and corresponding sensor activations for each resident's predicted activity.

Accordingly, this work proposes a Powerformer-based multi-output model for MHAR, addressing the limitations of traditional Transformer architectures in MHAR. By incorporating a Weighted Causal Multi-Head Attention (WCMHA) mechanism governed by a power-law function, the model effectively captures temporal locality, reduces the impact of irrelevant long-range dependencies, and maintains causality. To improve transparency and user trust, the model is further integrated with a Multi-Output Temporal Local Interpretable Model-agnostic Explanations (MO-TLIME) technique, enabling the identification of critical time segments and sensor inputs contributing to the classified activity of each resident.

The main contributions of this research study are as follows:

- A Powerformer-based multi-activity classification model utilising Weighted Causal Multi-Head Attention for efficient and accurate recognition of activities in shared environments.
- An extended Local Interpretable Model Agnostic Explanations (LIME) technique, MO-TLIME, incorporating a multi-output perturbation strategy for MHAR, providing resident-specific, time-segment-based explanations that provide influential temporal intervals along with their corresponding active sensors.
- A comprehensive evaluation comparing the proposed system with baseline deep learning models,

supported by statistical validation using the Wilcoxon Signed-Rank and paired t-tests. Ablation studies further examined the impact of causal and decay masks and assessed explanation reliability through input-level analyses across datasets.

This paper substantially extends our earlier conference paper [16] with an integrated explainability framework, expanded multi-dataset evaluation, additional baselines, and enhanced analysis. The remainder of this paper is organised as follows: Section II reviews related works in MHAR and explainable AI in smart home HAR. Section III presents the proposed MHAR framework. Section IV details the experimental setup, datasets used, pre-processing, and evaluation. Section V discusses the results, including performance comparisons with baseline models, statistical validation of the results, an ablation study, and explainability analysis. Section VI concludes the paper and outlines potential directions for future research.

II. RELATED WORKS

MHAR in smart homes typically relies on non-intrusive ambient sensor networks, enabling continuous monitoring while preserving residents' privacy. In recent years, deep learning models have shown significant potential in recognising the complex and overlapping activities of multiple residents from raw sensor data. This section reviews key advancements in MHAR models and explainability techniques that have been applied for real-world smart home HAR deployments.

A. STATE-OF-THE-ART RECOGNITION MODELS FOR MHAR

A wide range of techniques, including probabilistic models, especially Hidden Markov Models (HMMs) [17], [18], [19], [20], rule-based systems, traditional machine learning algorithms [21], advanced deep learning models, and hybrid approaches, have been proposed in the literature to recognise activities in multi-resident settings. Traditional machine learning models require manual feature extraction to capture spatial and temporal characteristics from the sensor data. These features serve to guide the models in managing the complexity inherent in MHAR, thereby facilitating accurate activity recognition. With the rise of deep learning, MHAR has increasingly adopted these models capable of automatically learning complex temporal and spatial patterns directly from the raw sensor data without manual feature extraction. Tran and Zhang [22] introduced a Recurrent Neural Network (RNN), especially a Gated Recurrent Units-based MHAR model, to capture temporal dependencies effectively and supported real-time online learning via stochastic gradient descent. Standard RNNs face challenges like vanishing gradients, limiting their ability to capture long-term dependencies where residents' actions can span across long temporal intervals, prompting the adoption of Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU). LSTMs, with their gating mechanisms and memory cells, mitigate vanishing gradients and allow long-term dependency learning. GRUs offer similar benefits with reduced computational overhead. Jain et al. [23] proposed deep and mini-batch LSTM models trained on time-windowed sensor data. Their stacked LSTM architecture handled long-range temporal dependencies through hierarchical temporal modelling, and dropout with bitmasking reduced overfitting. Liang et al. [24] similarly applied LSTM-based models, reinforcing their utility in MHAR tasks. To jointly model spatial and temporal dynamics, hybrid models combining Convolutional Neural Networks (CNN) with LSTMs have been explored. In the study by Arrotta et al. [25], CNN layers were used to extract spatial features from environmental and inertial sensor streams, which were then processed by LSTM layers to capture temporal dependencies. A reasoning module was further added to filter inconsistent activity predictions based on contextual rules. Similarly, Natani et al. [26] employed CNN-RNN hybrids to achieve effective spatiotemporal modelling. To improve the real-world applicability, Mehri et al. [27] proposed a Hierarchical Multi-label Classification (HMC)-LSTM model that imposed structural constraints across location, activity, and object hierarchies. This structure improved robustness but relied heavily on predefined hierarchies.

More recently, transformer-based models have emerged as a strong candidate for MHAR due to their parallel computation and capacity to capture long-range dependencies via self-attention. Unlike RNNs, transformers process sequences in parallel, improving both efficiency and contextual understanding. Chen et al. [10] introduced TRANS-BiGRU, a hybrid model combining Transformer encoders with Bidirectional GRUs. While Transformers enhanced global temporal dependency modelling, GRUs focused on localised temporal features. This architecture achieved high performance in recognising concurrent and collaborative activities.

However, although this transformer-based architecture has demonstrated strong performance in MHAR by capturing global temporal dependencies, it relies on non-causal self-attention mechanisms that permit access to future information, thereby limiting its applicability in real-time smart home deployments where strict temporal causality must be preserved. Moreover, this model employs a hybrid architecture, such as integration with GRU, to compensate for local temporal modelling, which increases architectural complexity and computational overhead while placing limited emphasis on recent sensor events that are critical in multi-resident scenarios.

These limitations motivates the development of a causal unified transformer-based MHAR architecture that supports short and long-range dependencies for efficient multi-resident activity modelling. A comparison of key papers is presented in Table 1. Overall, the evolution from probabilistic to hybrid deep learning models reflects a growing emphasis on accuracy, scalability, and robustness in MHAR systems. Despite significant progress, challenges remain in real-world generalisation, particularly in transparent decision-making.

B. EXPLAINABILITY TECHNIQUES FOR ACTIVITY RECOGNITION IN SMART HOMES

With the growing use of MHAR systems in domains like healthcare, eldercare, and behaviour monitoring, explainability has become critical. Opaque decisions can hinder user trust, limit clinical adoption, and obscure system accountability. While XAI has advanced for single-resident HAR, its application in multi-resident settings remains limited. Recent efforts have applied various XAI techniques to enhance the explainability of deep learning-based smart home HAR models. In the study proposed by Hussain et al. [28], healthcare-specific explanations were obtained in an Electroencephalography (EEG)-based HAR system, where one of the most popular post-hoc XAI techniques, LIME, is used to explain model decisions based on EEG spectral features. Similarly, Javed et al. [29] leveraged LIME to derive feature importance scores for transparency. In contrast, the authors of the pre-print [30] introduced a novel model-agnostic approach where, instead of perturbation-based feature attribution as in LIME, it generates explanations by analysing how the model's predictions change in response to strategic input transformations (jittering, masking, temporal clipping) during both training and inference. This counterfactual-style explainability method highlights which temporal components of the input data were most influential, offering complementary insights to traditional LIME-based methods that focus on feature importance.

XAI techniques are generally classified into model-agnostic and model-specific approaches. Model-agnostic methods operate independently of the model's internal architecture, allowing them to be applied across various algorithms, from traditional machine learning classifiers to deep learning networks. This versatility makes them especially valuable for conducting comparative analyses across different modelling approaches. LIME comes under model-agnostic approaches. In addition to LIME, other model-agnostic XAI techniques, such as SHapley Additive exPlanations (SHAP), have also been employed to improve interpretability in smart home HAR systems. Kim et al. [31] applied SHAP summary plots to visualise and quantify the impact of features on activity predictions across two distinct activity sets and utilised them for feature selection. Unlike LIME, which uses local surrogate models and perturbs input features to generate instance-level explanations, SHAP attributes the output of a model to each feature using cooperative game theory. Furthermore, the study integrated HAR with anomaly detection using Isolation Forest algorithms. Recently, Fiori et al. [32] examined the integration of Large Language Models (LLMs) for generating natural language explanations, making system decisions more understandable to non-experts. These studies demonstrate significant progress in explainability for single-resident HAR. LIME

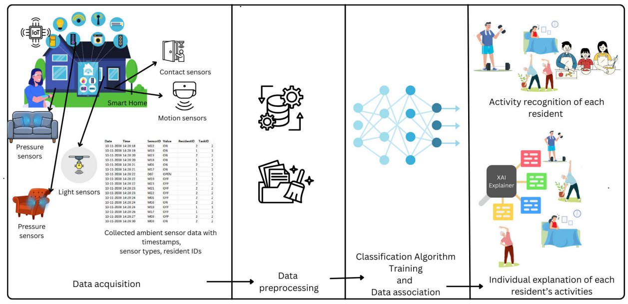

FIGURE 1. Overall System Architecture.

has gained broader adoption than SHAP mainly because of its computational efficiency, making it more suitable for deployment in resource-limited HAR systems. Visualisation-based explanation methods enhance interpretability by offering intuitive, instance-specific insights, enabling domain experts to better understand and validate model behaviour.

However, these explainability techniques applied to smart home HAR, such as LIME and SHAP, primarily provide static, feature-level explanations and are originally designed for tabular data. Although effective in single-resident HAR, these methods are less suitable for multi-resident scenarios, where overlapping activities and shared sensor streams require temporally grounded and resident-specific explanations. They generate explanations for a single output, making them inadequate for MHAR systems that jointly predict activities for multiple residents.

Therefore, these limitations highlight the need for an explainability technique capable of producing resident-specific, time-segment–based explanations tailored to multi-output MHAR models, thereby ensuring trustworthy activity recognition in multi-resident scenarios.

III. PROPOSED MHAR FRAMEWORK

The proposed framework is a powerformer-based multi-output classification model with Multi-Output Temporal Local Interpretable Model-agnostic Explanations (MO-TLIME) for MHAR in smart home environments. The system consists of four key components: data acquisition from ambient sensors, data preprocessing, a powerformer-based classification model with multi-output heads, and per-resident, time-segment-specific explanations via the proposed MO-TLIME technique. An overview of the system is illustrated in Fig.1.

A. MULTI-OUTPUT POWERFORMER-BASED CLASSIFICATION MODEL

Powerformer, introduced by Hegazy et al. [11], is a Transformer-based model specifically designed for time-series forecasting. While Transformer models have gained immense popularity for Natural Language Processing (NLP) tasks due to their powerful attention mechanism, encouraging their implementation in time-series data tasks, their direct application to time-series data poses challenges. The standard

TABLE 1. Key Works on MHAR.

Sl.No.	Study	Classification Algorithm	Dataset	Accuracy	Positive Outcomes	Limitations
1	Tran et al. (2020) [19]	Mixed Dependency Model	Centre for Advanced Studies in Adaptive Systems (CASAS) Multi-resident dataset, Activity Recognition with Ambient Sensing (ARAS) dataset	**CASAS:** 57.27% **ARAS House A:** 50.43% **ARAS House B:** 75.03%	Comprehensive modelling of dependencies – Introduced six HMM variants that capture individual, cross-resident, and group-level dependencies.	State space complexity increases with residents.
2	Jethanandani et al. (2020) [21]	Classifier chain with Decision Tree as base classifier with Majority Vote classifier	ARAS	**House A:** 88.02% **House B:** 97%	Multi-label classification approach. Ensemble learning framework to improve overall accuracy.	Didn't model time-series dependencies explicitly.
3	Jain et al. (2021) [23]	Deep and mini-batch LSTM	ARAS	**Mini-batch:** House A: 67.309% House B: 83.095% **Deep LSTM:** House A: 67.309% House B: 83.095%	Effective use of sequential networks. Multiresident handling via separate outputs.	Spatial and contextual reasoning have not been explored.
4	Natani et al. (2021) [26]	CNN + LSTM	ARAS	**House A:** R1: 81.81%, R2: 90.82% **House B:** R1: 89.21%, R2: 86.74%	Combined spatial extraction (CNN) and temporal sequencing (LSTM) performed significantly better than traditional sequential models.	Limited long-range dependency handling.
5	Chen et.al. (2022) [10]	Transformer with Bidirectional Gated Recurrent Unit (TRANS-BiGRU)	CASAS Multi-resident dataset, ARAS	**CASAS:** R1: 93.28% R2: 92.73% **ARAS House A:** R1: 89.70% R2: 91.63% **ARAS House B:** R1: 91.76% R2: 90.93%	Achieved greater accuracy capturing local and distant dependencies. Parallelism through Transformer.	Complex architecture. Non-causal – not suitable for real-time applications.

Multi-Head Attention (MHA) mechanism in the transformer architecture computes attention scores in an all-to-all manner, meaning that each time step can attend to every other time step in the sequence. Transformers' parallelism and the ability to model long-range temporal dependencies make them attractive, provided that the temporal structure is properly addressed. Although this global attention mechanism is beneficial in NLP, where word dependencies are often long-range and weakly ordered, it can be less effective in time-series data, where strong short-range dependencies dominate and are causal in nature. Nonetheless, time-series tasks also exhibit significant long-range dependencies, such as periodic or recurring activity patterns, which must be modelled in a manner that preserves causality while highlighting short-term temporal dynamics. In many physical systems, including smart home environments, events remain correlated over long time intervals, with these correlations decaying gradually over time rather than abruptly [33]. This decay often follows a power-law distribution, where the correlation strength between time steps decreases as a function of the logarithm of their temporal distance. For example, in multi-resident smart home environments, for instance, sensor events triggered by one resident cooking in the kitchen may strongly correlate with sensor activations in the kitchen occurring within minutes of each other, but less so with activities from hours earlier, such as "sleeping," highlighting the importance of local temporal context.

Capturing such long-range dependencies with appropriate weighting is critical for accurate modelling. To address this, Powerformer replaces the standard MHA with a WCMHA mechanism. The Powerformer's core concept is to enhance temporal attention efficiency by weighting the self-attention mechanism through both causal and decay masks [11]. In this work, we integrate the WCMHA module within our framework to effectively capture temporally dependent activities in smart home sensor data.

Given an input sequence of encoded sensor activations $\mathbf{X} \in \mathbb{R}^{T \times d_{\text{model}}}$, where T is the sequence length and d_{model} is the model (embedding) dimension per time step, each attention head h produces query, key and value projections via learned linear maps:

$$\mathbf{Q}_h = \mathbf{X}\mathbf{W}_h^Q, \qquad \mathbf{K}_h = \mathbf{X}\mathbf{W}_h^K, \qquad \mathbf{V}_h = \mathbf{X}\mathbf{W}_h^V, \tag{1}$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are trainable projection matrices for head h. The resulting projections have shapes $\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h \in \mathbb{R}^{T \times d_k}$. Here d_k denotes the per-head dimensionality so that a multi-head attention with H heads typically satisfies $d_{\text{model}} = H \cdot d_k$. Attention logits are computed as the scaled dot-product between queries and keys. The attention scores $\mathbf{S}_h$ in WCMHA are computed by incorporating both causal and temporal decay masks $\mathbf{M}^{(C,D)}$:

$$\mathbf{S}_h^{(C,D)} = \frac{\mathbf{Q}_h \mathbf{K}_h^{\top}}{\sqrt{d_k}} + \mathbf{M}^{(C,D)}. \tag{2}$$

For a given pair of time steps (t_i, t_j), where t_i and t_j denote the temporal indices (timestamps) of the i-th and j-th tokens respectively, the combined mask $\mathbf{M}^{(C,D)}$ is defined as:

$$\mathbf{M}_{i,j}^{(C,D)} = \begin{cases} -\infty, & j > i, \\ -\alpha \log\big(1 + t_i - t_j\big), & j \le i, \end{cases} \tag{3}$$

where $\alpha > 0$ is a tunable decay coefficient controlling the rate of temporal attenuation. The first condition ($j > i$) enforces the causal constraint, preventing a time step from attending to any future positions. For ($j \le i$), a logarithmic decay penalty proportional to the temporal distance $(t_i - t_j)$ is applied, emphasising more recent events and suppressing long-range dependencies. This temporal weighting enables the model to maintain causal order while adaptively reducing attention strength over longer time lags.

The normalized attention weights $\mathbf{C}_h$ and the head-wise output $\tilde{\mathbf{X}}_h$ are then obtained as:

$$\mathbf{C}_h = \text{Softmax}\big(\mathbf{S}_h^{(C,D)}\big), \qquad \tilde{\mathbf{X}}_h = \mathbf{C}_h \mathbf{V}_h. \tag{4}$$

The outputs from all heads are concatenated and passed through a linear projection to generate the final attention output. In our proposed model, we adopt the core WCMHA mechanism from Powerformer and extend it for multi-output classification by incorporating a global pooling layer followed by multiple classification heads corresponding to each resident, each comprising fully connected layers with a softmax activation to classify activities of different residents. This design enhances parameter efficiency, facilitates resident-level feature representation, and yields a compact yet informative summary of both short- and long-term dependencies for final classification. Furthermore, we utilise learnable positional embeddings, allowing the model to directly learn and adapt temporal position information from the training data. The workflow of the proposed model is summarised in Algorithm 1.

In Step 1, an input projection and patching operation is performed using a one-dimensional convolutional layer, Conv1D(·), with d_{model} filters. The convolutional kernel size determines the temporal patch length, while the stride parameter (*patch_stride*) controls the overlap between adjacent patches. This operation maps the raw input sequence into a sequence of patch-level embeddings $\mathbf{X}'$, enabling local temporal aggregation and improving computational efficiency. In Step 2, learnable positional encoding is added to the patch embeddings to obtain $\mathbf{X}''$, thereby adaptively capturing temporal ordering information directly from the training data.

Algorithm 1 Powerformer-Based Multi-Output Classification

Input: Segmented time series data X, number of outputs M, labels per output L
Output: Classified activity probabilities $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M\}$

1: **Input Projection and Patching**
   $X' \leftarrow$ Conv1D(X, filters $= d_{model}$, kernel_size, stride $=$ *patch_stride*)
2: **Add Positional Encoding**
   $X'' \leftarrow X' +$ PositionalEncoding(X')
3: **Pass through depth Powerformer blocks**
   for $i = 1$ **to** *depth* **do**
      $X'' \leftarrow$ PowerformerBlock(X'')
   end for
4: **Global Average Pooling**
   $Z \leftarrow$ GlobalAveragePooling1D(X'')
5: **Multi-output classification heads**
   for $m = 1$ **to** M **do**
      $h_1 \leftarrow$ ReLU(Dense(128)(Z))
      $h_1 \leftarrow$ Dropout(p)(h_1)
      $h_2 \leftarrow$ ReLU(Dense(128)(h_1))
      $\hat{y}_m \leftarrow$ Softmax(Dense(L)(h_2))
   end for
   return $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M\}$

In Step 3, the encoded sequence $\mathbf{X}''$ is passed through a stack of Powerformer blocks, where *depth* denotes the number of stacked blocks. Each Powerformer block incorporates the WCMHA mechanism, which enforces temporal causality and applies a power-law decay function to emphasise recent sensor events while suppressing less relevant long-range dependencies. In Step 4, a Global Average Pooling operation is applied along the temporal dimension to produce a fixed-length global representation $\mathbf{Z}$. This pooling strategy aggregates temporal features while reducing the number of parameters, making it well-suited for multi-output classification. In Step 5, the pooled representation $\mathbf{Z}$ is forwarded to M parallel classification heads, one corresponding to each resident. Each classification head consists of two fully connected layers with 128 hidden units and ReLU activation functions, followed by a dropout layer with dropout rate p to mitigate overfitting. The final dense layer uses a Softmax activation to output a probability distribution $\hat{\mathbf{y}}_m$ over activity classes for the m-th resident.

The overall model output is a set of predicted activity probability vectors, $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_M\}$, enabling simultaneous and resident-specific activity recognition in shared smart home environments.

B. MULTI-OUTPUT TEMPORAL LIME (MO-TLIME) FOR MULTI- RESIDENT ACTIVITY EXPLANATION

Local Interpretable Model-Agnostic Explanations (LIME) [34] is a widely adopted post-hoc interpretability technique that explains predictions of complex machine learning models by learning an interpretable surrogate model locally around a specific input instance. In our work, we adapt and extend LIME for explaining time-series classification in an MHAR setting using sensor data. To generate explanations, standard LIME perturbs the input instance multiple times and observes the model's prediction for each perturbed sample. Perturbation, in this context, refers to slightly altering the input data to simulate different variations of the original input. This can help the surrogate model capture how the true model behaves locally. For tabular data, LIME generates perturbations by randomly sampling and replacing feature values, while for text and images, it removes words or toggles superpixels. Each perturbed instance is evaluated by the opaque model, and predictions are weighted by their similarity to the original input using a kernel function. LIME then fits a simple surrogate model, such as linear regression, to approximate the complex model's local behaviour around that instance. The use of random perturbation is central to the effectiveness of LIME, as it enables the approximation of a opaque model's local decision boundary without requiring access to internal model parameters.

For time-series sensor instances, we apply segmentation-based perturbation by dividing each sensor instance into fixed-length temporal segments and applying random binary masks. A masked segment simulates the absence of sensor reading within that segment. Randomly masking temporal segments allows the explanation process to simulate the absence or suppression of specific sensor events, thereby revealing their contribution to the model's predictions. This is particularly important in MHAR scenarios, where overlapping activities and shared sensor streams cause actions from multiple residents to activate the same sensors over time, making it challenging to attribute specific temporal segments to each resident's predicted activity. Random perturbation ensures diverse local neighbourhood sampling around the original instance, resulting in a more robust and faithful surrogate model that captures non-linear dependencies in the local region.

Our extension of LIME is tailored for models with multi-output heads, where multiple separate classifications are made simultaneously, one for each resident. For each output head, we generate separate perturbation matrices and collect the corresponding predicted probabilities. By independently perturbing temporal segments for each output head, the proposed approach enables resident-specific explanations that separate the influence of shared sensor activations across multiple residents. We compute the cosine similarity between each perturbed instance and the original input to determine their proximity, which is converted into weights using an exponential kernel. A local surrogate model, typically a weighted linear regression, is trained using these inputs and prediction scores, producing coefficients that indicate the contribution of each sensor instance segment to the final classification output. These coefficients are then sorted to highlight the most influential segments in determining each resident's activity. Algorithm 2 provides the workflow of the extended LIME technique. This adaptation allows us to generate segment-level explanations for multi-activity classification in smart homes. Detailed workflow of the proposed Powerformer-based explainable framework is presented in Fig. 2.

Algorithm 2 MO-TLIME for Multi-Resident Time-Series Classification

Input: x_{original}: time-series instance, f: trained R-head model, N: number of perturbations, S: segment size, π_x: similarity kernel

Output: Explanation weights $\{w^{(r)}\}_{r=1}^{R}$ for each resident

1: Segment x_{original} into T segments of length S

2: Generate N random binary perturbation vectors $M \in \{0, 1\}^{N\times T}$

3: For each perturbation $m \in M$:

Apply m to x_{original} to obtain perturbed input z_i

Query model: $(f^{(1)}(z_i), f^{(2)}(z_i), \ldots, f^{(R)}(z_i))$

Compute similarity weight:

$$\pi_x(z_i) = \exp\left(-\frac{D(x_{\text{original}}, z_i)^2}{\sigma^2}\right)$$

For each resident $r = 1, \ldots, R$:

Each perturbed data point consists of $(z_i, f^{(r)}(z_i), \pi_x(z_i))$, added to dataset $Z^{(r)}$

4: For each resident $r = 1, \ldots, R$:

Fit a weighted linear model on $Z^{(r)}$, restricted to at most K features

5: Return $\{w^{(r)}\}_{r=1}^{R}$ as the explanation for each resident r for x_{original}

IV. EXPERIMENTAL SETUP

To assess the performance and interpretability of the proposed framework, this section details the experimental setup used in our study. The datasets and their preprocessing procedures are first described, followed by the system configuration and hyperparameter selection. Finally, the evaluation protocol is discussed, highlighting the metrics and comparative baselines employed to ensure a fair and comprehensive performance analysis.

A. DATASET

The datasets used in this study are the publicly available Activity Recognition with Ambient Sensing (ARAS) datasets [35] and the Centre for Advanced Studies in Adaptive Systems (CASAS) [36] multi-resident dataset, which captures multi-resident activities in diverse smart home environments. The ARAS datasets consist of recordings from two smart homes (House A and House B), each occupied by two residents (Resident A and Resident B) over a continuous 30-day period. Sensor data were captured every second, producing 86,400 events per day and 30 daily files for each house. Each log covers the full day (00:00:00 to 23:59:59), documenting all activities performed

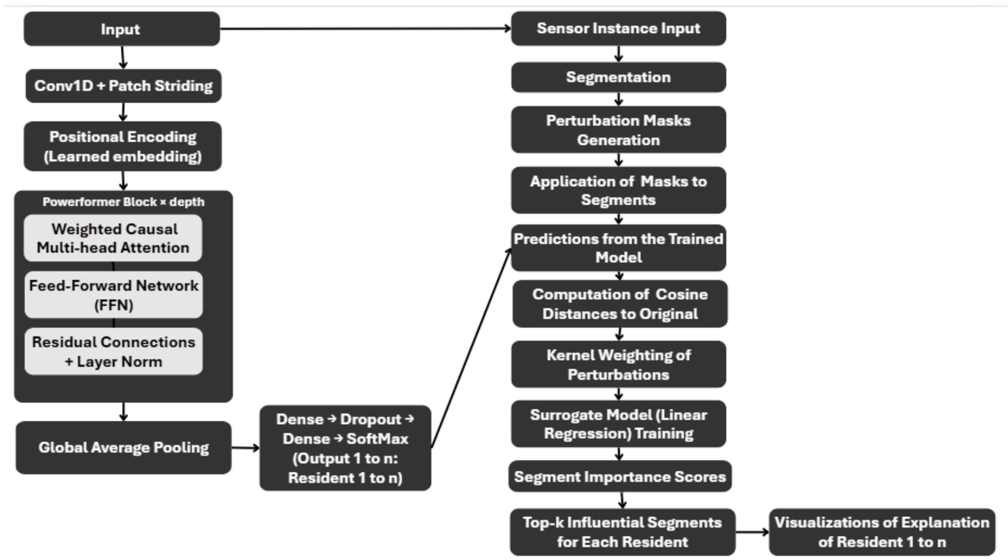

FIGURE 2. Proposed Powerformer-based Explainable Framework.

by both residents. The dataset employs ambient binary sensors, such as motion, contact, and other non-intrusive environmental sensors, to monitor household interactions. The CASAS multi-resident dataset was collected in a smart apartment testbed at Washington State University. The apartment includes motion, temperature, and analog sensors for monitoring water flow and stove use, along with contact sensors for tracking object interactions. Forty participants performed predefined activities, with two residents occupying the apartment simultaneously to simulate multi-resident scenarios. Unlike the continuous 30-day recordings in ARAS, CASAS data are event-based with shorter, scripted sessions, comprising 26 annotated files of about 30 minutes each.

The physical layouts and sensor placements for each house are illustrated in Fig. 3 for ARAS datasets and Fig. 4 for CASAS dataset, providing spatial context for interpreting sensor activations and resident movements. Further details are summarised in Table 2. The list of recognised activities in the ARAS datasets is provided in Table 3 and the CASAS dataset in Table 4.

B. DATASET PRE-PROCESSING

The input configuration for the proposed explainable multi-output Powerformer framework varies by dataset. In the ARAS dataset, continuous time-series data from ambient binary sensors (0 = OFF, 1 = ON) installed in real homes are used. In contrast, the CASAS dataset provides event-based logs with Sensor ID, Sensor Value, and Timestamp. To prepare both datasets for the proposed model, a 60-second fixed sliding window was applied, with each segment labelled according to activities performed within that interval.

The use of a fixed 60-second sliding window follows the protocol of the benchmark study [10] used for comparison, ensuring fair and consistent evaluation. Assigning a single activity label per resident to each fixed window implicitly

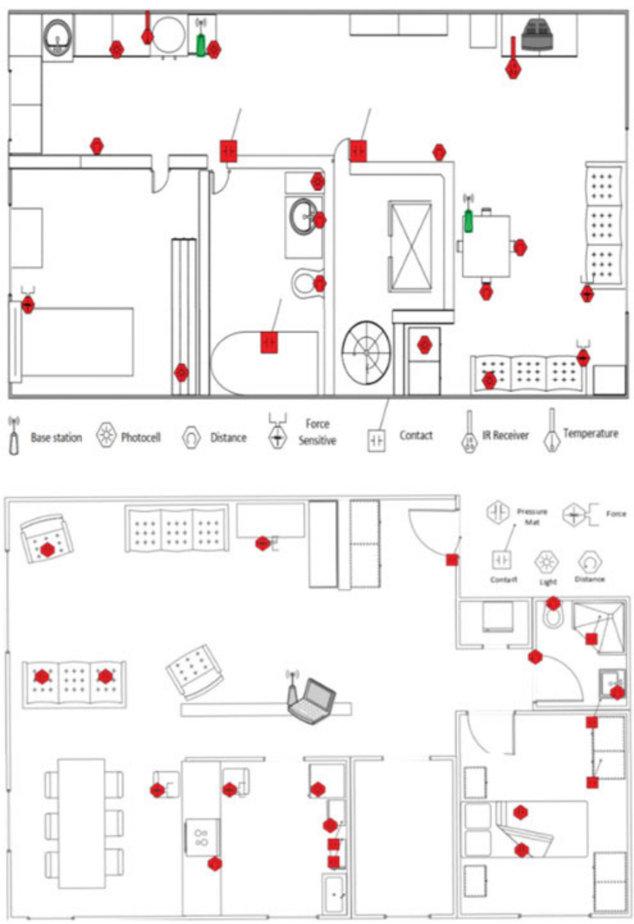

FIGURE 3. Layouts for House A (at the top) and House B (at the bottom) with sensors' placements [35].

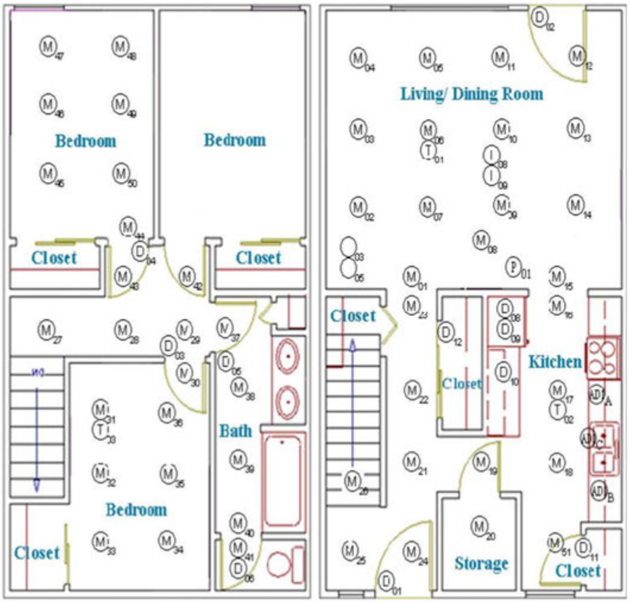

FIGURE 4. Layout of WSU smart apartment where M01-M51 are motion sensors, I01-I08 are item sensors, D01-D12 are cabinet sensors, AD1-A and AD1-B are water sensors, AD1-C is a burner sensor, P01 is a phone sensor, and T01-T03 are temperature sensors [36].

TABLE 2. Dataset Properties of ARAS and CASAS datasets [35], [36].

Property	ARAS House A	ARAS House B	CASAS
No. of Ambient Sensors	20 sensors of 7 different types	20 sensors of 6 different types	78 sensors of 7 different types
Sensor Types	Force-Sensitive Resistor (FSR), pressure mat, contact, proximity, sonar distance, photocell, IR	FSR, pressure mat, contact, proximity, sonar distance, photocell	Motion, item, cabinet, water, burner, phone, and temperature sensors
House Information	One bedroom, one living room, one kitchen, one bathroom	Two bedrooms, one living room, one kitchen, one bathroom	Three bedrooms, one living room, one storage room, one kitchen, and one bathroom
Residents	Two males, both aged 25	Married couple, average age 34	Two volunteers at a time (from a pool of 40)
Duration of Data Collection	30 days	30 days	26 sessions (approximately 30 minutes each)
No. of Activities	27	27	15

assumes that the activity remains dominant within that interval. This assumption is reasonable for a short window duration such as 60 seconds, as most daily living activities typically persist longer than one minute, making brief transitions unlikely to dominate the window. While intra-window

TABLE 3. Activity Labels of ARAS datasets [35].

ID	Activity Label	ID	Activity Label
1	Other	15	Toileting
2	Going Out	16	Napping
3	Preparing Breakfast	17	Using Internet
4	Having Breakfast	18	Reading Book
5	Preparing Lunch	19	Laundry
6	Having Lunch	20	Shaving
7	Preparing Dinner	21	Brushing Teeth
8	Having Dinner	22	Talking on the Phone
9	Washing Dishes	23	Listening to Music
10	Having Snack	24	Cleaning
11	Sleeping	25	Having Conversation
12	Watching TV	26	Having Guest
13	Studying	27	Changing Clothes
14	Having Shower		

TABLE 4. Activity labels of CASAS dataset [36].

Activity ID	Activity Label	Resident ID
1	Filling the medication dispenser	A
2	Hanging up clothes	B
3	Moving furniture	B
4	Reading magazine	A
5	Watering plants	A
6	Sweeping floor	B
7	Playing checker	A and B
8	Preparing dinner	A
9	Setting the table	B
10	Reading magazine	A
11	Paying bills	A and B
12	Gathering and packing picnic food	A
13	Retrieve dishes from a kitchen cabinet	A and B
14	Packing picnic supplies	B
15	Packing picnic food	A

transitions may introduce some label ambiguity, particularly in multi-resident settings, dominant-activity labeling is widely adopted in benchmark MHAR studies [10] to ensure stable supervision. Shorter windows may fragment activities, whereas longer windows may obscure transitions; thus, the 60-second window represents a practical compromise that balances temporal resolution and label consistency while acknowledging this inherent limitation.

The ARAS dataset, already in binary format, required minimal preprocessing. Whereas the CASAS dataset is event-driven, consisting of asynchronous sensor activations, resulting in a variable number of sensor events per time window depending on resident interactions. Therefore, it underwent a structured preprocessing pipeline: extracting temporal features (hour, minute, day of week) encoded via cyclical sine-cosine transformations, tokenising sensor IDs and values into numerical indices, and padding sequences to fixed lengths for embedding. To address potential labelling errors, Cleanlab's DataLab [37], [38], [39] was used to estimate label correctness and remove low-confidence samples, improving dataset reliability. Finally, the recognition task was formulated as a multi-output classification problem, since multiple residents can act simultaneously. One-hot encoding was applied separately to each resident's activity labels, aligning with the multiple output heads of the model for accurate, independent activity prediction.

C. SYSTEM CONFIGURATION AND HYPERPARAMETER SELECTION

All experiments were conducted on Google Colab Pro with an NVIDIA A100 GPU. The proposed multi-output Powerformer-based classification model was configured with carefully chosen hyperparameters based on empirical evaluation to ensure optimal performance in the multi-resident activity recognition task. Training hyperparameters were adopted from the benchmark study [10], while model-specific parameters were empirically tuned using validation performance rather than an exhaustive grid search, with the final values reported in Table 5 for reproducibility. The loss function was categorical cross-entropy, applied separately to each output head to support multi-output classification for Resident A and Resident B. The model was trained for 200 epochs with a batch size of 40, following the configuration by Chen et al. [10]. An early stopping mechanism with a patience of 10 was employed, monitoring the combined validation accuracy of both output heads to prevent overfitting. For comparison, baseline models including the one-dimensional CNN with LSTM (CNN1D-LSTM) and the two-dimensional CNN with LSTM (CNN2D-LSTM) from Natani et al. [26], and TRANS-BiGRU from Chen et al. [10], were implemented using the same hyperparameters reported in their respective studies to ensure fair and consistent evaluation. These baseline models were chosen to capture a diverse range of state-of-the-art deep learning approaches for modelling sequential data.

For generating time-segment explanations, each 60-second input sequence was divided into 15, 30, and 60 segments, representing roughly 4, 2, and 1 seconds of sensor data, respectively. This segmentation approach was designed to achieve a balance between temporal detail and computational efficiency, allowing the explanations to reflect both fine- and coarse-grained activity dynamics. The segmented inputs were then analysed using the proposed MO-TLIME explainer to determine the most influential segments driving the model's activity predictions. Among these, the top five influential segments were visualised to illustrate the portions of the input sequence that contributed most to the model's decisions. This threshold was chosen for interpretability, though it can be adjusted, as this technique ranks all segments by their relative importance.

TABLE 5. Hyperparameters of the Proposed Model.

Hyperparameter	Description	Value
d_{model}	Embedding & attention dimension	384
d_{hidden}	Feed-forward layer size	256
num_heads	Number of attention heads	6
decay_mode	Attention decay type	power-law
alpha	Decay intensity control	0.8
dropout_rate	Dropout applied in blocks	0.3
patch_stride	Stride for Convolutional 1D patching	4
depth	Number of Powerformer blocks	5
optimizer	Training optimizer	AdamW

D. EVALUATION

To assess model performance, both time-series and standard k-fold cross-validation strategies [39] were employed, selected according to the characteristics of each dataset to prevent future-to-past data leakage. Specifically, for the ARAS dataset, which consists of 30 consecutive days of continuous sensor readings, time-series cross-validation was applied to maintain temporal order and ensure that future data points were never used to predict past events. Unlike conventional random data splitting, the adopted time-series approach incrementally expands the training set while sequentially sliding the validation window forward in time. In contrast, the CASAS dataset comprises independent session-based event logs rather than a continuous stream, inherently reducing the risk of temporal leakage. Therefore, standard k-fold cross-validation was deemed appropriate for the CASAS dataset. Both datasets were evaluated using a 10-fold setup [10]. Evaluation metrics, including accuracy, precision, recall, and F1-score, were computed per resident. Additionally, computation time was recorded [40] to compare the efficiency of the traditional Transformer with the proposed Powerformer-based MHAR model. To assess the statistical significance of performance differences between models, the Shapiro–Wilk test [41] was first applied to determine whether the distribution of performance metrics across the folds was normal; a p-value greater than 0.05 indicated normality. If normality was confirmed, a paired t-test [42] was conducted to evaluate whether the mean difference in metrics between two models was statistically significant, assuming matched pairs across folds. When the normality assumption was violated, the Wilcoxon signed-rank test [43] was used as a non-parametric alternative suitable for non-normal or ordinal data distributions. An ablation study was also performed across both datasets to examine the effect of excluding key components, such as the decay and causal masks, on model performance.

To ensure transparency, LIME-based explanations were generated for each resident's classification output. The proposed MO-TLIME technique was evaluated on the ARAS dataset, whose continuous data stream enables fine-grained segment-level interpretability. Although applicable to other datasets, ARAS was chosen over CASAS due to its suitability for temporal analysis. For the proposed MO-TLIME technique, the dataset was split into training (70%), validation (10%), and testing (20%) sets. The trained Powerformer-based multi-output model generated predictions for the test data, which were then analysed using the MO-TLIME Explainer to produce time-segment-based explanations showing the most influential input regions for each resident's activity. To assess the significance of the generated explanations, an input-level ablation study was conducted under three scenarios, where the input data were segmented into 15, 30, and 60 segments. Important segments were alternately retained or masked to measure changes in resident-level and average accuracies, thereby evaluating how well the explanations captured the model's decision behaviour and the influence of temporal granularity.

V. RESULTS AND DISCUSSION

The performance comparison of the proposed Powerformer with baseline models is summarised in Table 6 for the ARAS House A dataset, Table 7 for the ARAS House B dataset, and Table 8 for the CASAS dataset. All metrics are reported in percentages and are weighted averages.

For the ARAS House A dataset, Powerformer achieved the highest Accuracy (84.93%) and Recall (84.93%) for Resident A, while CNN2D-LSTM obtained superior Precision (85.44%) and F1 Score (83.97%). For Resident B, Powerformer again led in Accuracy (86.37%) and Recall (86.37%), whereas CNN1D-LSTM yielded the top F1 Score (85.18%). These results confirm Powerformer's strong detection capability, with CNN-based models excelling in precision-oriented metrics. For the ARAS House B dataset, Powerformer achieved the highest Accuracy (92.76%) and Recall (92.76%) for Resident A, while CNN1D-LSTM recorded the best Precision (92.15%) and F1 Score (91.96%). For Resident B, Powerformer led in Accuracy (91.19%), Recall (91.19%), and F1 Score (90.82%), while TRANS-BiGRU achieved the highest Precision (92.47%). Powerformer maintained superior efficiency, requiring only 165.81 s (House A) and 200.24 s (House B), significantly lower than TRANS-BiGRU's 612.79 s and 473.72 s, respectively, owing to its causal attention and power-law decay mechanisms. These findings highlight Powerformer's effectiveness in balancing accuracy and computational cost.

For the CASAS dataset, Powerformer consistently outperformed all baselines. For Resident A, it achieved the highest Accuracy (87.27%), Precision (88.70%), Recall (87.27%), and F1 Score (86.78%), demonstrating strong generalisation across environments. TRANS-BiGRU followed with Accuracy (84.76%) and F1 Score (84.47%), while CNN2D-LSTM attained moderate performance (Accuracy 83.19%, F1 Score 82.15%). CNN1D-LSTM performed the lowest. For Resident B, Powerformer again excelled (Accuracy 89.00%, Precision 91.51%, Recall 89.00%, F1 Score 89.03%), surpassing TRANS-BiGRU (Accuracy 84.17%, F1 83.94%) and CNN2D-LSTM (Accuracy 83.39%, F1 83.39%). Computation time efficiency remained high, with Powerformer completing in 68.98 s, outperforming TRANS-BiGRU (70.74 s).

Tables 9, 10, and 11 present the F1-micro and F1-macro performance results for the ARAS House A, ARAS House B, and CASAS datasets, respectively, to evaluate both overall recognition and class-balanced performance. As expected, F1-micro closely follows accuracy trends, whereas F1-macro highlights minority activity recognition. For the ARAS House A and House B datasets, F1-macro values are noticeably lower than F1-micro across all models, reflecting the class imbalance and the presence of rare activities inherent to multi-resident smart home environments. For the ARAS House A dataset, F1-macro values are relatively low, with the proposed Powerformer achieving F1-macro scores of 45.25% for Resident A and 33.38% for Resident B, remaining competitive with CNN2D-LSTM, which attains 46.99% and 34.74%, respectively. For the ARAS House B dataset, the proposed Powerformer records F1-macro values of 39.47% (Resident A) and 39.02% (Resident B), outperforming TRANS-BiGRU and CNN-based baselines for both residents. In contrast, substantially higher F1-macro scores are obtained on the CASAS dataset owing to its more balanced activity distribution, where Powerformer achieves the best performance for both residents with 83.32% (Resident A) and 84.90% (Resident B), surpassing TRANS-BiGRU and CNN-based models. These results confirm that while F1-macro is constrained by class imbalance in the ARAS datasets, the proposed Powerformer consistently demonstrates stronger class-balanced performance across datasets while maintaining high overall recognition accuracy.

Although the raw performance gains on the ARAS datasets appear modest, statistical significance testing verified that Powerformer's improvements are consistent and meaningful. For ARAS House A, Powerformer significantly outperformed CNN2D-LSTM in Accuracy and Recall for Resident B ($p = 0.0211$) and CNN1D-LSTM in the same metrics for Resident A ($p = 0.0356$). For ARAS House B, it achieved significant improvements over TRANS-BiGRU for Resident B in Accuracy and Recall ($p = 0.0020$) and F1 Score ($p = 0.0371$); over CNN2D-LSTM in Accuracy, Recall ($p = 0.0020$), and F1 Score ($p = 0.0195$); and over CNN1D-LSTM in Accuracy and Recall ($p = 0.0098$). On the CASAS dataset, Powerformer demonstrated statistically significant gains across most metrics. For Resident A, it outperformed TRANS-BiGRU in Accuracy and Recall ($p = 0.0274$), CNN1D-LSTM across all metrics ($p < 0.001$), and CNN2D-LSTM in Accuracy ($p = 0.0229$), Precision ($p = 0.0370$), Recall ($p = 0.0229$), and F1 Score ($p = 0.0195$). For Resident B, it significantly outperformed TRANS-BiGRU across all metrics ($p < 0.05$) and achieved consistent improvements over CNN1D-LSTM ($p <= 0.0034$) and CNN2D-LSTM ($p < 0.05$).

These results confirm that Powerformer's gains are both substantial and statistically significant over CNN- and Transformer-based baselines. Overall, the findings highlight Powerformer's consistent advantage in achieving top Accuracy and Recall across multiple residents and environments, with statistically validated reliability. Its strength lies in effectively modelling temporal dependencies and capturing both short-term cues and long-range dynamics through patch-based tokenisation and causal attention with power-law decay. While CNN-based baselines occasionally achieved higher Precision and F1 scores, reflecting their ability to minimise false positives, Powerformer provides a more robust solution for recall-driven detection of complex and overlapping activities. Dataset characteristics further explain these trends. The event-based CASAS data favoured

TABLE 6. Per-Resident Performance Comparison of the Proposed with the Benchmark Models for the ARAS House A Dataset.

	Resident A				Resident B			
Model	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Powerformer (Proposed)	84.93	83.48	84.93	83.08	86.37	84.98	86.37	84.54
TRANS-BiGRU	84.54	84.81	84.54	83.51	86.17	85.50	86.17	84.76
CNN2D-LSTM	84.39	85.44	84.39	83.97	85.57	86.41	85.57	84.97
CNN1D-LSTM	84.05	84.32	84.05	83.08	86.27	86.12	86.27	85.18

TABLE 7. Per-Resident Performance Comparison of the Proposed with the Benchmark Models for the ARAS House B Dataset.

	Resident A				Resident B			
Model	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Powerformer (Proposed)	92.76	91.86	92.76	91.75	91.19	92.19	91.19	90.82
TRANS-BiGRU	92.14	91.58	92.14	91.52	89.45	92.47	89.45	89.05
CNN2D-LSTM	91.97	92.04	91.97	91.41	88.81	89.11	88.81	89.39
CNN1D-LSTM	92.52	92.15	92.52	91.96	89.39	92.13	89.39	89.89

TABLE 8. Per-Resident Performance Comparison of the Proposed with the Benchmark Models for the CASAS Dataset.

	Resident A				Resident B			
Model	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Powerformer (Proposed)	87.27	88.70	87.27	86.78	89.00	91.51	89.00	89.03
TRANS-BiGRU	84.76	86.44	84.76	84.47	84.17	87.32	84.17	83.94
CNN2D-LSTM	83.19	84.77	83.19	82.15	83.39	86.97	83.39	83.39
CNN1D-LSTM	75.48	76.26	75.48	73.89	76.86	79.79	76.86	75.91

TABLE 9. ARAS House A dataset — F1-Micro and F1-Macro.

Resident A		
Model	**F1-micro (%)**	**F1-macro (%)**
Powerformer (Proposed)	84.93	45.25
TRANS-BiGRU	84.54	46.30
CNN2D-LSTM	84.39	46.99
CNN1D-LSTM	84.05	44.89
Resident B		
Model	**F1-micro (%)**	**F1-macro (%)**
Powerformer (Proposed)	86.37	33.38
TRANS-BiGRU	86.17	35.17
CNN2D-LSTM	85.57	34.74
CNN1D-LSTM	86.27	32.75

TABLE 10. ARAS House B dataset — F1-Micro and F1-Macro.

Resident A		
Model	**F1-micro (%)**	**F1-macro (%)**
Powerformer (Proposed)	92.76	39.47
TRANS-BiGRU	92.14	38.53
CNN2D-LSTM	91.97	37.75
CNN1D-LSTM	92.52	36.66
Resident B		
Model	**F1-micro (%)**	**F1-macro (%)**
Powerformer (Proposed)	91.19	39.02
TRANS-BiGRU	89.45	36.45
CNN2D-LSTM	88.81	33.37
CNN1D-LSTM	89.39	34.50

Powerformer's global temporal modelling, yielding superior results across all metrics. In contrast, the continuous nature of the ARAS dataset allowed CNN-based hybrids to retain slight precision advantages. Overall, Powerformer maintained a balanced, context-aware performance across both event-driven and continuous environments, validating its robustness and generalizability for real-world MHAR applications.

TABLE 11. CASAS dataset — F1-Micro and F1-Macro.

Resident A		
Model	**F1-micro (%)**	**F1-macro (%)**
Powerformer (Proposed)	87.27	83.32
TRANS-BiGRU	84.76	82.02
CNN2D-LSTM	83.19	80.67
CNN1D-LSTM	75.48	68.31
Resident B		
Model	**F1-micro (%)**	**F1-macro (%)**
Powerformer (Proposed)	89.00	84.90
TRANS-BiGRU	84.17	81.86
CNN2D-LSTM	83.39	80.53
CNN1D-LSTM	76.86	68.01

A. RESULTS OF ABLATION STUDY

To analyse the contribution of individual architectural components, ablation experiments were performed. The ablation study on the ARAS Houses A and B datasets shows that removing the decay or causal masks affects residents differently, as presented in Tables 12 and 13, respectively. For House A, Resident A exhibits slight gains when either mask is removed, with accuracy improving from 84.93% (Full) to 85.15% (No-Decay mask) and 85.04% (No-Causal mask). The No-Decay mask also slightly improves precision (84.71%) and F1 score (83.60%). However, Resident B performs best under the Full configuration, achieving the highest accuracy and recall (86.37%), while ablations provide only marginal improvements in F1 score. A similar trend is observed in House B. For Resident A, the No-Decay mask yields the best accuracy (92.90%) and F1 score (91.94%), while for Resident B, the Full configuration remains optimal with the highest accuracy (91.19%) and F1 score (90.82%). Overall, causal masking helps preserve temporal order but increases computational cost, whereas the decay mask

TABLE 12. Performance Comparison of Ablation Variants for ARAS House A dataset.

Resident A					
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Computation Time (s)**
Proposed Powerformer (Full)	84.93	83.48	84.93	83.08	165.81
Proposed Powerformer (No-Decay mask)	85.15	84.71	85.15	83.60	201.73
Proposed Powerformer (No-Causal mask)	85.04	83.50	85.04	83.07	199.94
Resident B					
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Computation Time (s)**
Proposed Powerformer (Full)	86.37	84.98	86.37	84.54	165.81
Proposed Powerformer (No-Decay mask)	86.08	84.77	86.08	84.36	201.73
Proposed Powerformer (No-Causal mask)	86.30	84.95	86.30	84.57	199.94

TABLE 13. Performance Comparison of Ablation Variants for ARAS House B dataset.

Resident A					
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Computation Time (s)**
Proposed Powerformer (Full)	92.76	91.86	92.76	91.75	200.24
Proposed Powerformer (No-Decay mask)	92.90	91.65	92.90	91.94	211.16
Proposed Powerformer (No-Causal mask)	92.86	91.70	92.86	91.82	187.36
Resident B					
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Computation Time (s)**
Proposed Powerformer (Full)	91.19	92.19	91.19	90.82	200.24
Proposed Powerformer (No-Decay mask)	90.42	92.73	90.42	90.79	211.16
Proposed Powerformer (No-Causal mask)	90.33	92.48	90.33	90.42	187.36

TABLE 14. Performance Comparison of Ablation Variants for CASAS dataset.

Resident A					
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Computation Time (s)**
Proposed Powerformer (Full)	87.27	88.70	87.27	86.78	68.98
Proposed Powerformer (No-Decay mask)	86.86	87.40	86.86	85.84	67.78
Proposed Powerformer (No-Causal mask)	86.87	88.35	86.87	86.35	68.08
Resident B					
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Computation Time (s)**
Proposed Powerformer (Full)	88.99	91.51	88.99	89.03	68.98
Proposed Powerformer (No-Decay mask)	87.82	90.18	87.82	87.77	67.79
Proposed Powerformer (No-Causal mask)	87.45	90.17	87.45	87.21	68.08

improves precision with lower overhead but reduced stability. The Full configuration thus provides the most balanced trade-off between accuracy and efficiency. In contrast, results on the CASAS dataset, presented in the Table 14, show that the Full configuration consistently outperforms both ablated variants across all metrics for both residents. This difference arises from dataset characteristics; ARAS contains dense, continuous binary streams where temporal constraints are partly redundant, while CASAS is sparse and event-based, relying heavily on both masks for temporal consistency. Overall, the Full configuration demonstrates the most stable and effective performance across diverse datasets, confirming Powerformer's robustness and adaptability.

B. SCALABILITY AND PRACTICAL CONSIDERATIONS

The computational complexity of the proposed model is primarily governed by the number of input patches, segment size, attention heads, and Transformer layers, similar to other transformer-based attention models. The segment size controls a trade-off between temporal resolution and computational cost, while increasing the number of Transformer layers improves representational capacity at the expense of higher training and inference complexity. The power-law decay coefficient controls the emphasis on local versus long-range temporal dependencies, where larger values prioritise recent events and smaller values preserve longer temporal context. These parameters were empirically tuned to balance accuracy and efficiency. Runtime and memory requirements scale with the temporal patches per window rather than raw sequence duration, as the proposed framework processes data using fixed-length windows. This design enables efficient handling of continuous, long-duration sensor streams through window-by-window inference, without unbounded growth in per-window computational cost. The causal attention mechanism further restricts attention scope, reducing unnecessary computations and supporting online inference. While long-term monitoring generates a larger number of windows over time, thereby increasing cumulative computational demand, the per-window runtime and memory usage remain bounded. As a result, the proposed design remains scalable for typical smart home activity recognition scenarios.

The generalisation observed across the datasets is primarily empirical and supported by the adaptability of the proposed model to both continuous and event-driven sensor data through dataset-specific preprocessing. The proposed

TABLE 15. Input-level ablation results for the House A dataset using the proposed MO-TLIME technique.

Segments	Evaluation Setting	Resident A Accuracy (%)	Resident B Accuracy (%)	Average Accuracy (%)
15	Only Important Segments (Top 5)	81.96	86.77	84.36
	Important Segments Masked (Top 5)	22.92	59.55	41.23
	Only Important Segments (Top 10)	85.64	89.79	86.82
	Important Segments Masked (Top 10)	19.24	58.32	38.78
30	Only Important Segments (Top 10)	64.84	87.35	76.09
	Important Segments Masked (Top 10)	40.04	58.97	49.51
	Only Important Segments (Top 20)	86.77	88.03	87.40
	Important Segments Masked (Top 20)	18.11	58.29	38.20
60	Only Important Segments (Top 20)	79.30	86.43	82.86
	Important Segments Masked (Top 20)	25.58	59.99	42.74
	Only Important Segments (Top 40)	85.98	87.96	86.97
	Important Segments Masked (Top 40)	18.59	58.36	38.63
	With All Segments	87.45	88.41	87.93

TABLE 16. Input-level ablation results for the House B dataset using the proposed MO-TLIME technique.

Segments	Evaluation Setting	Resident A Accuracy (%)	Resident B Accuracy (%)	Average Accuracy (%)
15	Only Important Segments (Top 5)	90.05	91.55	90.80
	Important Segments Masked (Top 5)	45.51	56.04	50.78
	Only Important Segments (Top 10)	93.52	92.04	92.78
	Important Segments Masked (Top 10)	42.04	55.56	48.80
30	Only Important Segments (Top 10)	89.75	91.44	90.59
	Important Segments Masked (Top 10)	45.81	56.16	50.98
	Only Important Segments (Top 20)	93.45	91.87	92.66
	Important Segments Masked (Top 20)	42.11	55.72	48.91
60	Only Important Segments (Top 20)	82.08	90.95	86.52
	Important Segments Masked (Top 20)	53.47	56.64	55.06
	Only Important Segments (Top 40)	87.71	89.51	89.61
	Important Segments Masked (Top 40)	47.85	55.49	51.67
	With All Segments	93.94	92.18	93.06

Powerformer architecture itself does not rely on assumptions about sensor layout, sensor identity, or activity sequences, operating solely on learned temporal patterns. As with most MHAR systems, the method assumes the availability of sensor streams and annotated activity segments; deployment in new environments, therefore, primarily requires preprocessing alignment and retraining rather than architectural modification. While the current evaluation focuses on two-resident scenarios due to dataset availability, the proposed framework is inherently extensible to settings with more residents. The multi-output design allows additional output heads based on the number of residents without modifying the core Powerformer architecture. Such an extension would require retraining with appropriately annotated data but not architectural redesign, supporting scalability to higher resident counts. The causal design of the proposed framework supports real-time inference by avoiding future information leakage, while the preprocessing pipeline mitigates several practical deployment challenges, including sensor noise, missing events, and asynchronous activations through temporal windowing and event aggregation. Latency in online deployment is primarily governed by the selected window length and model complexity and is explicitly managed in the proposed framework through a fixed window and causal attention, balancing responsiveness with recognition stability. Long-term deployment may still be affected by concept drift due to changes in resident behaviour, which would require periodic retraining.

C. RESULTS OF MO-TLIME EXPLANATIONS

To demonstrate the effectiveness of the proposed MO-TLIME Explainer, two representative scenarios were analysed: (i) an instance where multiple residents perform the same activity simultaneously, and (ii) an instance where residents engage in different activities at the same time.

In the first scenario, a data instance corresponding to the activity Sleeping was selected. The generated explanations, shown in Fig. 5, highlight the top five most influential temporal segments and their corresponding active sensors. For both residents, the pressure mat sensors on the bed consistently emerged as the most discriminative features. However, the specific influential time segments varied between residents, reflecting individual variations in activity execution. This demonstrates the explainer's ability to capture personalised temporal dynamics, thereby facilitating interpretable and resident-specific insights for adaptive smart home environments.

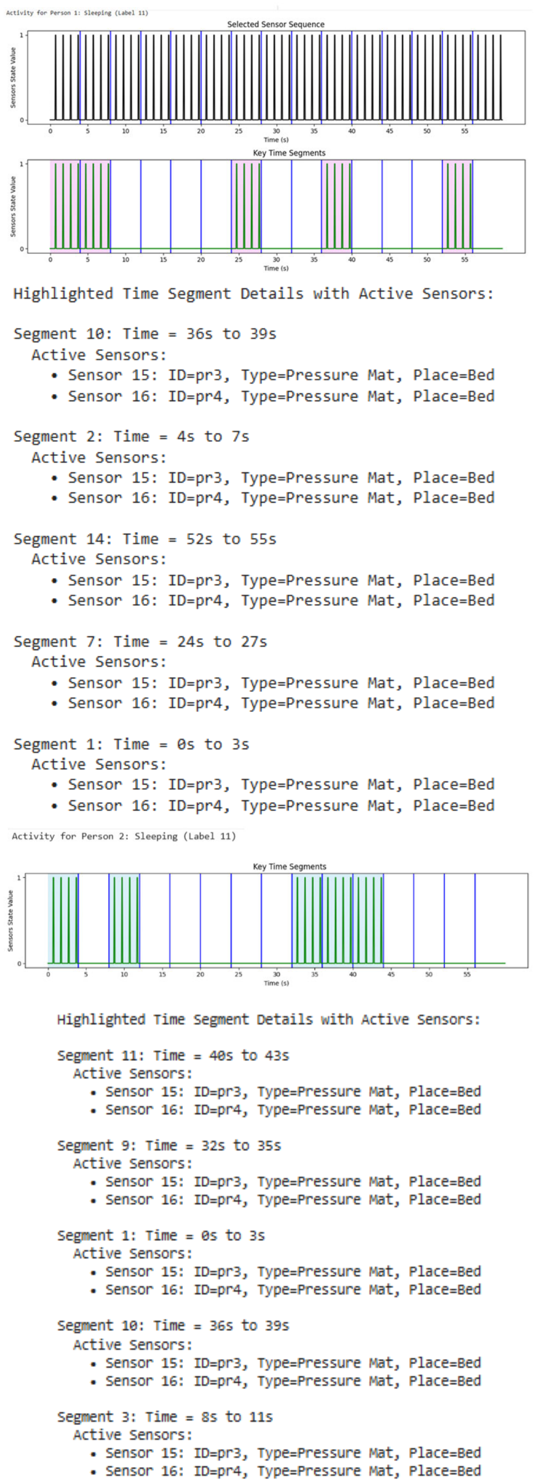

FIGURE 5. Time-segment-based explanations of activities for Resident A and Resident B exhibiting the same activity, generated using the MO-TLIME technique. The charts on the top highlight the Top 5 key time segments (highlighted in pink for Resident A and blue for Resident B), while the descriptions at the bottom provide active sensor information for the Top 5 ranked segments.

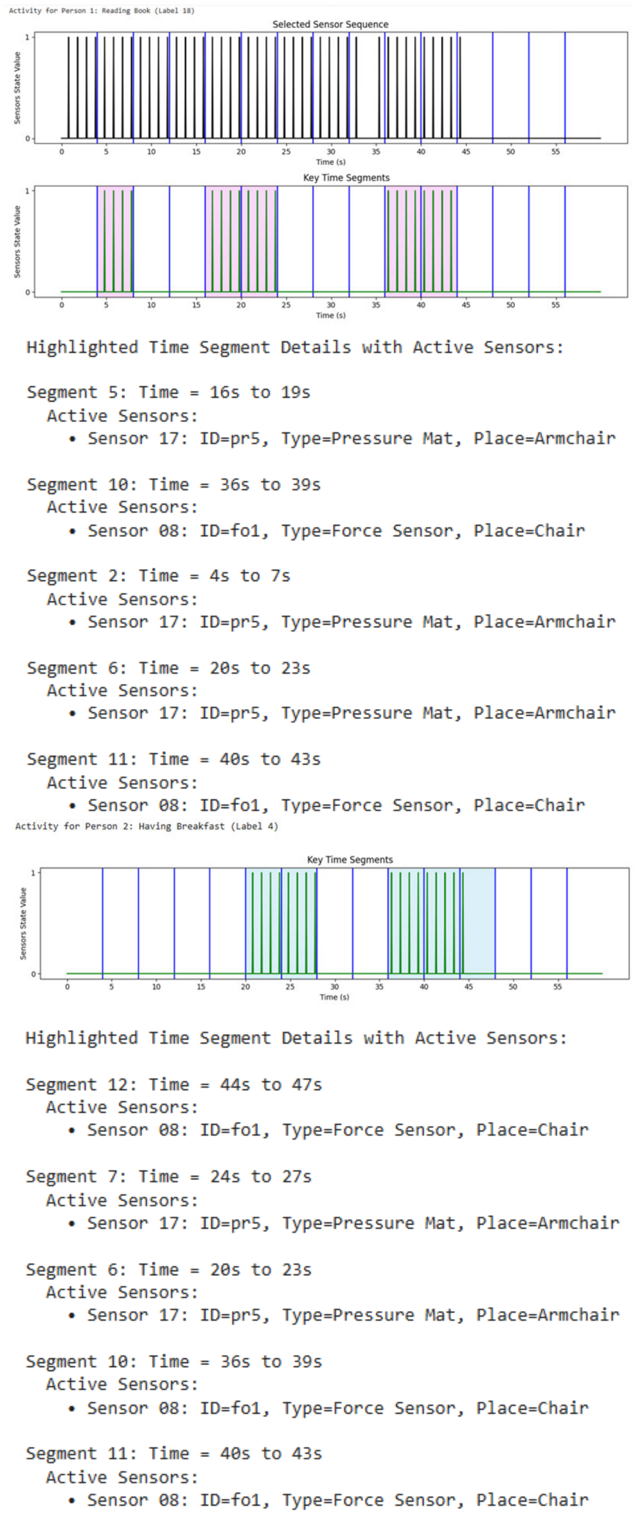

FIGURE 6. Time-segment-based explanations of activities for Resident A and Resident B exhibiting different activities, generated using the MO-TLIME technique. The charts on the top highlight the Top 5 key time segments (highlighted in pink for Resident A and blue for Resident B), while the descriptions at the bottom provide active sensor information for the Top 5 ranked segments.

In the second scenario, residents performed distinct activities at the same time: Reading Book (Resident A) and Having Breakfast (Resident B). As illustrated in Fig. 6, the pressure mat on the armchair was the dominant contributor for Resident A, while the force sensor on the chair was most influential for Resident B. The critical time segments also differed, with Resident A and Resident B. These variations confirm the explainer's capability to produce temporal-aware and resident-specific interpretations that support personalisation in multi-resident settings.

To further validate the importance of the identified temporal segments, an input-level ablation study was performed using the ARAS House A and House B datasets. Each 60-second input sequence was divided into 15, 30, and 60 segments, corresponding to approximately 4-, 2-, and 1-second intervals, respectively. Segment size introduces a trade-off between interpretability and efficiency, with finer segments enabling more precise temporal explanations at increased computational cost.Three experimental conditions were evaluated: (i) retaining only the most influential segments (top 5–40 depending on segmentation), (ii) masking the most influential segments, and (iii) using the full, unmodified input. Accuracy was used as the evaluation metric under a fixed 70%–10%–20% train–validation–test split. The results presented in Tables 15 and 16 demonstrate that retaining only the most influential segments maintains performance close to the baseline.

For instance, when segmented into 15 parts, using the Top-10 segments yields accuracies of 86.82% for House A and 92.78% for House B, compared to 87.93% and 93.06%, respectively, when all segments are used. Conversely, masking these important segments results in a substantial accuracy drop (House A: 38.78%; House B: 48.80%), confirming their critical contribution to activity recognition. Overall, the proposed MO-TLIME Explainer not only enhances interpretability but also identifies the influential temporal and sensor patterns driving model predictions. By revealing resident-specific and activity-aware insights, it strengthens transparency and supports the development of personalised, adaptive smart home systems that balance explainability and predictive accuracy.

VI. CONCLUSION AND FUTURE WORKS

The proposed Powerformer-based explainable framework effectively overcomes key limitations of conventional Transformer models by balancing accuracy, efficiency, and interpretability. It generalises well across different smart home environments and provides resident-specific explanations, making it suitable for MHAR. The framework consistently achieved superior accuracy and recall, demonstrating robustness in recognising complex overlapping activities. Although CNN models occasionally offered slightly higher precision, the Powerformer delivered a better overall trade-off between predictive performance and computational cost. The proposed MO-TLIME technique enhanced interpretability by identifying critical temporal segments and sensor activations influencing each resident's activity prediction. Temporal variations observed across residents confirmed the framework's resident-centric explainability. The key contributions include adapting the Powerformer architecture for multi-output classification and introducing a time-segment-based perturbation strategy for MHAR explanations. The ablation study further validated the importance of causal and decay masks in maintaining recognition performance. Future work will focus on edge optimisation through model compression, enhanced explainability using SHAP and attention visualisation for deeper insight and trust, and integration with smart home healthcare systems to enable adaptive automation, anomaly detection, and personalised interventions in real time.

REFERENCES

[1] Statista. *Smart Home–Worldwide*. Accessed: Nov. 10, 2025. [Online]. Available: https://www.statista.com/outlook/cmo/smart-home/worldwide

[2] World Health Organization (WHO). *Ageing and Health*. Accessed: Nov. 13, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[3] S. Ariccio and A. Theodorou, "Stay at home behavior during COVID-19: The role of person-home relationships," *J. Environ. Psychol.*, vol. 97, Aug. 2024, Art. no. 102334, doi: 10.1016/j.jenvp.2024.102334.

[4] F. Foerster, M. Smeja, and J. Fahrenberg, "Detection of posture and motion by accelerometry: A validation study in ambulatory monitoring," *Comput. Hum. Behav.*, vol. 15, no. 5, pp. 571–583, Sep. 1999, doi: 10.1016/s0747-5632(99)00037-0.

[5] A. Benmansour, A. Bouchachia, and M. Feham, "Human activity recognition in pervasive single resident smart homes: State of art," in *Proc. 12th Int. Symp. Program. Syst. (ISPS)*, Algiers, Algeria, Apr. 2015, pp. 1–9, doi: 10.1109/isps.2015.7244997.

[6] M. Gochoo, T.-H. Tan, S.-H. Liu, F.-R. Jean, F. S. Alnajjar, and S.-C. Huang, "Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 693–702, Mar. 2019, doi: 10.1109/JBHI.2018.2833618.

[7] J. Ye, H. Jiang, and J. Zhong, "A graph-attention-based method for single-resident daily activity recognition in smart homes," *Sensors*, vol. 23, no. 3, p. 1626, Feb. 2023, doi: 10.3390/s23031626.

[8] A. Benmansour, A. Bouchachia, and M. Feham, "Multioccupant activity recognition in pervasive smart home environments," *ACM Comput. Surveys*, vol. 48, no. 3, pp. 1–36, Dec. 2015, doi: 10.1145/2835372.

[9] Q. Li, W. Huangfu, F. Farha, T. Zhu, S. Yang, L. Chen, and H. Ning, "Multi-resident type recognition based on ambient sensors activity," *Future Gener. Comput. Syst.*, vol. 112, pp. 108–115, Nov. 2020, doi: 10.1016/j.future.2020.04.039.

[10] D. Chen, S. Yongchareon, E. M.-K. Lai, J. Yu, Q. Z. Sheng, and Y. Li, "Transformer with bidirectional GRU for nonintrusive, sensor-based activity recognition in a multiresident environment," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 23716–23727, Dec. 2022, doi: 10.1109/JIOT.2022.3190307.

[11] K. Hegazy, M. W. Mahoney, and N. B. Erichson, "Powerformer: A transformer with weighted causal attention for time-series forecasting," 2025, *arXiv:2502.06151*.

[12] D. Das, Y. Nishimura, R. P. Vivek, N. Takeda, S. T. Fish, T. Plötz, and S. Chernova, "Explainable activity recognition for smart home systems," *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 2, pp. 1–39, Jun. 2023, doi: 10.1145/3561533.

[13] D. Gaspar, P. Silva, and C. Silva, "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," *IEEE Access*, vol. 12, pp. 30164–30175, 2024, doi: 10.1109/ACCESS.2024.3368377.

[14] A. Theissler, F. Spinnato, U. Schlegel, and R. Guidotti, "Explainable AI for time series classification: A review, taxonomy and research directions," *IEEE Access*, vol. 10, pp. 100700–100724, 2022, doi: 10.1109/ACCESS.2022.3207765.

[15] T. Sivill and P. Flach, "LIMESegment: Meaningful, realistic time-series explanations," in *Proc. 25th Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 151, 2022, pp. 3418–3433.

[16] V. Vijayananth, T. Perumal, and E. Marlisah, "Efficient transformer-based classifier for multi-resident human activity recognition in smart homes," in *Proc. IEEE 15th Int. Conf. Syst. Eng. Technol. (ICSET)*, Kuala Lumpur, Malaysia, Oct. 2025, pp. 219–224, doi: 10.1109/icset65917.2025.11283709.

[17] Y.-T. Chiang, K.-C. Hsu, C.-H. Lu, L.-C. Fu, and J. Yung-Jen Hsu, ''Interaction models for multiple-resident activity recognition in a smart home,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Taipei, Taiwan, Oct. 2010, pp. 3753–3758, doi: 10.1109/IROS.2010.5650340.

[18] H. Alemdar and C. Ersoy, ''Multi-resident activity tracking and recognition in smart environments,'' *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 4, pp. 513–529, Aug. 2017, doi: 10.1007/s12652-016-0440-x.

[19] S. N. Tran, T.-S. Ngo, Q. Zhang, and M. Karunanithi, ''Mixed-dependency models for multi-resident activity recognition in smart homes,'' *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 23445–23460, Aug. 2020, doi: 10.1007/s11042-020-09093-0.

[20] R. Gómez-Ramos, J. Duque-Domingo, E. Zalama, and J. Gómez-García-Bermejo, ''An unsupervised method to recognise human activity at home using non-intrusive sensors,'' *Electronics*, vol. 12, no. 23, p. 4772, Nov. 2023, doi: 10.3390/electronics12234772.

[21] M. Jethanandani, A. Sharma, T. Perumal, and J.-R. Chang, ''Multi-label classification based ensemble learning for human activity recognition in smart home,'' *Internet Things*, vol. 12, Dec. 2020, Art. no. 100324, doi: 10.1016/j.iot.2020.100324.

[22] S. N. Tran and Q. Zhang, ''Towards multi-resident activity monitoring with smarter safer home platform,'' in *Smart Assisted Living*, F. Chen, R. García-Betances, L. Chen, M. Cabrera-Umpiérrez, and C. Nugent, Eds., Cham, Switzerland: Springer, 2019, pp. 249–267.

[23] V. Jain, D. Jhawar, S. Saini, T. Perumal, and A. Sharma, ''Human activity recognition in ambient sensing using sequential networks,'' in *Proc. Int. Conf. Paradigms Comput.*, Singapore, 2021, pp. 493–501, doi: 10.1007/978-981-15-7533-4_38.

[24] J.-M. Liang, P.-L. Chung, Y.-J. Ye, and S. Mishra, ''Applying machine learning technologies based on historical activity features for multi-resident activity recognition,'' *Sensors*, vol. 21, no. 7, p. 2520, Apr. 2021, doi: 10.3390/s21072520.

[25] L. Arrotta, C. Bettini, G. Civitarese, and R. Presotto, ''Context-aware data association for multi-inhabitant sensor-based activity recognition,'' in *Proc. 21st IEEE Int. Conf. Mobile Data Manage. (MDM)*, Versailles, France, Jun. 2020, pp. 125–130, doi: 10.1109/MDM48529.2020.00034.

[26] A. Natani, A. Sharma, and T. Perumal, ''Sequential neural networks for multi-resident activity recognition in ambient sensing smart homes,'' *Appl. Intell.*, vol. 51, no. 8, pp. 6014–6028, 2021, doi: 10.1007/s10489-020-02134-z.

[27] H. Mehri, T. Lemlouma, and N. Montavont, ''A hierarchical multi-label classification of multi-resident activities,'' in *Proc. Intell. Data Eng. Autom. Learn. (IDEAL)*. Cham, Switzerland: Springer, vol. 13113, 2021, pp. 76–86, doi: 10.1007/978-3-030-91608-4_8.

[28] I. Hussain, R. Jany, R. Boyer, A. Azad, S. A. Alyami, S. J. Park, M. M. Hasan, and M. A. Hossain, ''An explainable EEG-based human activity recognition model using machine-learning approach and LIME,'' *Sensors*, vol. 23, no. 17, p. 7452, Aug. 2023, doi: 10.3390/s23177452.

[29] A. R. Javed, H. U. Khan, M. K. B. Alomari, M. U. Sarwar, M. Asim, A. S. Almadhor, and M. Z. Khan, ''Toward explainable AI-empowered cognitive health assessment,'' *Frontiers Public Health*, vol. 11, Mar. 2023, Art. no. 1024195, doi: 10.3389/fpubh.2023.1024195.

[30] Y. Huang, Y. Zhou, H. Zhao, T. Riedel, and M. Beigl, ''Explainable deep learning framework for human activity recognition,'' 2024, *arXiv:2408.11552*.

[31] T. H. Kim, S. Ojo, M. Krichen, M. A. Alamro, A. Mihoub, and G. A. Sampedro, ''Automated explainable and interpretable framework for anomaly detection and human activity recognition in smart homes,'' *Neural Comput. Appl.*, vol. 37, no. 16, pp. 9295–9308, Jun. 2025, doi: 10.1007/s00521-025-10991-3.

[32] M. Fiori, G. Civitarese, P. Choudhary, and C. Bettini, ''Leveraging large language models for explainable activity recognition in smart homes: A critical evaluation,'' 2025, *arXiv:2503.16622*.

[33] A. Clauset, C. R. Shalizi, and M. E. J. Newman, ''Power-law distributions in empirical data,'' *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, Nov. 2009.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, '''Why should i trust you?': Explaining the predictions of any classifier,'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[35] H. Alemdar, H. Ertan, O. D. Incel, and C. Ersoy, ''ARAS human activity datasets in multiple homes with multiple residents,'' in *Proc. 7th Int. Conf. Pervasive Comput. Technol. Healthcare Workshops*, Venice, Italy, May 2013, pp. 232–235.

[36] G. Singla, D. J. Cook, and M. Schmitter-Edgecombe, ''Recognizing independent and joint activities among multiple residents in smart environments,'' *J. Ambient Intell. Humanized Comput.*, vol. 1, no. 1, pp. 57–63, Mar. 2010, doi: 10.1007/s12652-009-0007-1.

[37] C. Northcutt, L. Jiang, and I. Chuang, ''Confident learning: Estimating uncertainty in dataset labels,'' *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, Apr. 2021, doi: 10.1613/jair.1.12125.

[38] C. G. Northcutt, A. Athalye, and J. Mueller, ''Pervasive label errors in test sets destabilize machine learning benchmarks,'' in *Proc. 35th Conf. Neural Inf. Process. Syst. (NeurIPS)*, San Diego, CA, USA, 2021, pp. 1–13.

[39] F. Pedregosa, ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[40] E. Ramanujam and T. Perumal, ''MLMO-HSM: Multi-label multi-output hybrid sequential model for multi-resident smart home activity recognition,'' *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 3, pp. 2313–2325, Dec. 2022, doi: 10.1007/s12652-022-04487-4.

[41] S. S. SHAPIRO and M. B. WILK, ''An analysis of variance test for normality (complete samples),'' *Biometrika*, vol. 52, nos. 3–4, pp. 591–611, Dec. 1965, doi: 10.1093/biomet/52.3-4.591.

[42] Student, ''The probable error of a mean,'' *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908.

[43] F. Wilcoxon, ''Individual comparisons by ranking methods,'' *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, Dec. 1945.

VISHALINI VIJAYANANTH (Member, IEEE) is currently pursuing the Master of Science degree in computer and embedded systems engineering with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. Her research interests include artificial intelligence, the Internet of Things, and embedded systems.

THINAGARAN PERUMAL (Senior Member, IEEE) received the Ph.D. degree in smart technology and robotics from Universiti Putra Malaysia. He is currently an Associate Professor and the Head of the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. His research interests include interoperability in smart homes, the Internet of Things (IoT), wearable computing, cyber-physical systems, proactive IoT architectures, cognitive/semantic IoT frameworks for smart homes, and wearable devices for rehabilitation. He chairs the IEEE Consumer Technology Society Malaysia Chapter.

ERZAM MARLISAH received the Ph.D. degree from Universiti Putra Malaysia. He is currently a Senior Lecturer with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. His research interests include artificial intelligence and machine learning, specifically reinforcement learning and evolutionary computing.