Graph-Based Methods for Multimodal Indoor Activity Recognition: A Comprehensive Survey

Graph-Based Methods for Multimodal Indoor Activity Recognition: A Comprehensive Survey

Saeedeh Javadi, Daniele Riboni, Luigi Borzì, *Member, IEEE*, and Samaneh Zolfaghari

***Abstract*—This survey article explores graph-based approaches to multimodal human activity recognition in indoor environments, emphasizing their relevance to advancing multimodal representation and reasoning. With the growing importance of integrating diverse data sources such as sensor events, contextual information, and spatial data, effective human activity recognition methods are essential for applications in smart homes, digital health, and more. We review various graph-based techniques, highlighting their strengths in encoding complex relationships and improving activity recognition performance. Furthermore, we discuss the computational efficiencies and generalization capabilities of these methods across different environments. By providing a comprehensive overview of the state-of-the-art in graph-based human activity recognition, this article aims to contribute to the development of more accurate, interpretable, and robust multimodal systems for understanding human activities in indoor settings.**

***Index Terms*—Graph-based methods, human activity recognition, indoor environments, interpretable models, multimodal learning, reasoning techniques, sensor data.**

I. Introduction

In recent years, the landscape of social computing has experienced a significant shift, marked by increasing variety of data modalities ranging from text and images to audio and videos [1], [2], [3]. This paradigm shift has not only enriched our understanding of social phenomena but has also necessitated the development of advanced techniques for representation and reasoning to effectively leverage these multimodal data sources [1].

The abundance of multimodal data presents both a challenge and an opportunity [4], [5], [6]. On one hand, the heterogeneity of these data sources demand sophisticated representation techniques capable of encoding information from disparate modalities into a cohesive structure. On the other hand, the inherent relationships and dependencies between different modalities offer fertile ground for reasoning mechanisms that can extract meaningful insights and uncover hidden patterns. Indeed, the increasing complexity of multimodal data requires advanced techniques to effectively represent and reason across diverse modalities. In this context, graph-based methods offer a cohesive framework for capturing spatial and temporal dependencies across multiple modalities, which is essential for applications like human activity recognition (HAR).

HAR can be categorized by a broad range of dimensions, including the type of data source, the environment, and the activity type [7]. Based on the data source, HAR includes sensor-based approaches, which use wearable and ambient sensors for privacy-preserving monitoring, and vision-based approaches, which rely on image or video data but face privacy and lighting challenges. In terms of environment, HAR can be classified as indoor or outdoor. Indoor HAR emphasizes confined spaces such as homes and healthcare facilities, while outdoor HAR involves activities performed in open or public settings, often relying on vision-based or GPS data. Activity types further refine HAR studies, encompassing physical activities, daily living behaviors, and health-related actions. In this regard, this survey focuses on the pressing need for graph-based multimodal representation and reasoning techniques, focusing specifically on the domain of sensor-based HAR in indoor environments. Indeed, HAR from sensor data is rapidly advancing in the past decades. It involves automatically detecting human activities from a range of sensors which has numerous applications. These applications include remote patient monitoring and medical diagnosis to shorten hospital stays [8], [9], elderly care, emergency assistance both at home and at assisted living facilities [10], [11], reminder systems [12] for people with cognitive disorders and chronic conditions [13], [14], and recognition of sports and leisure activities [15] to improve people's quality of life [16].

By leveraging *graph-based activity recognition* (GAR) approaches, we aim to bridge the gap between different modalities, facilitating the seamless integration of disparate data sources for more accurate and interpretable activity recognition and remote sensing [17], [18], [19]. While we acknowledge broader advancements in multimodal graph-based methods, such as recent studies on emotion recognition in conversations [20], cross-modal retrieval from educational slides [21], and

Received 30 July 2024; revised 12 November 2024 and 5 December 2024; accepted 13 December 2024. *(Corresponding author: Samaneh Zolfaghari.)*

Saeedeh Javadi and Luigi Borzì are with the Department of Computer and Control Engineering, Polytechnic University of Turin, 10129 Turin, Italy (e-mail: saeedeh.javadi@studenti.polito.it; luigi.borzi@polito.it).

Daniele Riboni is with the Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy (e-mail: riboni@unica.it).

Samaneh Zolfaghari is with the School of Innovation, Design and Engineering, Mälardalen University, 72123 Västerås, Sweden (e-mail: samaneh.zolfaghari@mdu.se).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCSS.2024.3523240, provided by the authors.

Digital Object Identifier 10.1109/TCSS.2024.3523240

student engagement prediction [22], these applications lie beyond the specific focus of our work. Additionally, recent models such as SpectralGPT [23] and multimodal foundation models [24], [25], [26] are designed primarily for outdoor remote sensing tasks, including land cover segmentation, environmental monitoring, and cross-regional analysis. Similarly, methods focused on skeleton-based action recognition and visual data-based activity recognition [27], [28], [29], [30], [31] emphasize single-modalities or visual cues, often lacking the multimodal sensor fusion necessary for indoor activity recognition.

Given the increasing prevalence of HAR applications, numerous surveys and reviews have been recently published that tackle techniques and methodologies for recognizing human behaviors and activities based on sensor data on different perspectives. While some papers, including [32] and [33], provide a broad overview of the different approaches, others focus on specific methodologies. Given their effectiveness when large sets of training data are available, deep learning approaches to HAR have been extensively reviewed in [34] and [35]. In contrast to these surveys, our paper focuses on a variety of graph-based approaches for HAR, including not only graph neural networks but also graphical models and knowledge-based frameworks. Other existing surveys, including [36] and [37], are specifically addressed to graph-based neural network approaches to HAR. However, unlike these paper, our survey considers a wider range of graph-based methodologies. Moreover, the former papers address image-based and video-based techniques, while our survey considers sensor-based HAR. Few survey papers have examined the HAR based on multimodal data. Deep learning approaches to multimodal HAR have been reviewed in [38]. Multimodal knowledge graph techniques to indoor scene recognition have been reviewed in [39]; however, in that paper, only vision-based methods are considered, whereas our survey focuses on sensor-based HAR recognition.

Our article is the first to comprehensively survey diverse graph-based methods for multimodal sensor fusion HAR. We critically analyze their strengths and weaknesses across several dimensions, including training data requirements, integration with emerging technologies such as large language models (LLM) to enhance interpretability and facilitate human–computer interaction in social computing spaces, real-time recognition capabilities, and comparative analysis in terms of performance and scalability. Our primary contributions involve an in-depth exploration of the theoretical foundations of graph-based representations and reasoning, establishing a solid framework for understanding their application in indoor HAR. This work is dedicated to indoor environments, excluding outdoor or non-sensor-based methods. We highlight state-of-the-art methodologies and their effectiveness in capturing the inherent relationships and dependencies between modalities for HAR. We also examine the use of existing tools and technologies for multimodal GAR, particularly in social computing domains such as smart homes and assisted living environments. Based on our analysis, we identify key research directions and open challenges, paving the way for future advancements in graph-based multimodal HAR. Our survey serves as a valuable resource for researchers and practitioners, guiding the development of more robust, innovative, and interpretable solutions for multimodal indoor HAR.

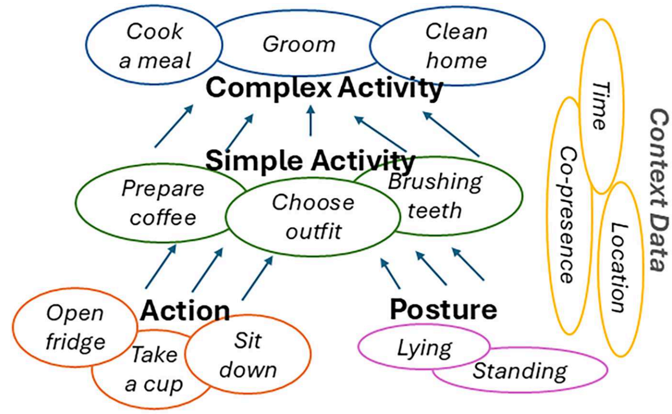

Fig. 1. Hierarchy of activities considered in this survey.

We divided the study into four distinct sections, aligning with the primary aim of employing multimodality for GAR in smart indoor environments. Section II investigates the background information related to HAR, including activity hierarchy, activity categories, sensing technologies, and evaluation methodologies and metrics. Section III outlines our search and selection strategy to gather relevant studies based on our inclusion criteria, followed by a comprehensive science mapping analysis of the selected studies. In Section IV, we conduct an in-depth analysis of various GAR approaches to explore their characteristics, advantages, and disadvantages. In Section VI, we discuss the findings from our thorough analysis of the relevant studies, highlighting key insights and open challenges. Finally, Section VII concludes the article.

II. Background on Activity Complexity and Activity Recognition Metrics

Human activities are variegate, and may range from simple actions, like *open fridge*, to complex activities such as *cooking a meal*. Moreover, they may be executed by a single individual or by multiple people, either in a concurrent or interleaved fashion. The complexity of an activity significantly affects how challenging it is to recognize. In this section, we present a categorization of human activities, and discuss the metrics commonly used in the literature to evaluate the effectiveness of activity recognition systems.

1) Activity Hierarchy: To recognize human activities based on the observation of sensor data, different researchers proposed hierarchical frameworks, such as the one proposed by Lukowicz et al. [40]. In those frameworks, activities at increasing level of complexity are recognized based on the recognition of simpler activities or actions. In Fig. 1, we illustrate a high-level hierarchy of human activities, which is general enough to cover most of the research studies considered in our survey paper.

The lowest side of the hierarchy includes *actions* and *postures*. In this context, we define an action as a unit of activity that cannot be decomposed in simpler ones. Examples of actions are: opening the fridge, sitting down, and taking a cup. In general, actions have a very brief duration, and are considered as instantaneous events. Some actions may be further decomposed in terms of sequences of atomic or manipulative gestures [40]. However, gestures are outside the scope of this survey; hence, they are not included in the hierarchy. Postures, such as lying or standing, are frequently part of activity recognition frameworks, since they are useful to recognize specific activity types, such as physical activities.

Simple activities are characterized by the execution of certain sequences of actions. For instance, prepare coffee can be characterized by the sequence of actions: open the kitchen cabinet, take coffee jar, put coffee in the machine, turn the coffee machine on. Simple actions typically have a rather short duration, limited to some seconds or few minutes.

Complex activities are characterized by the execution of a set of simple activities that concur to fulfill a certain goal. For instance, the complex activity *grooming* can be characterized by the execution of simpler activities such as showering, dressing up, and washing teeth. Normally, complex activities have a sensible duration, and may last from some minutes to few hours.

Even though they are not strictly part of an activity hierarchy, *context data* [41] such as current location, time of the day, and environmental parameters are often included in activity recognition frameworks, since they provide additional information to characterize human activities.

2) Activity Categories: Human activities can be further classified in categories; i.e., groups of activities that share common features. For the sake of this survey, we consider the following main categories. *Daily living activities* include everyday routine tasks such as watching television, eating, making coffee, and sleeping.

General activities cover basic and common activity such as sitting, standing, walking, going upstairs and downstairs.

Health-related activities are related to healthcare or abnormality detection, rehabilitation, and fall detection.

Physical activities include vigorous actions such as jumping, exercising, and cycling.

Posture analysis focuses on specific movements such as head rotations, arm movements, and leaning.

Work activities occur primarily in office or school environments, including tasks such as teaching, computer work, writing, and reading.

3) Sensing Technologies: In the field of HAR, different sensing technologies offer distinct advantages and disadvantages. Wearable sensors, such as accelerometers, gyroscopes and heart rate monitors, are widely used due to their ability to record motion and physiological signals with high accuracy. These sensors provide continuous, real-time monitoring, making them very effective for detailed HAR and personalized data collection. However, their effectiveness is often hampered by compliance issues, as users have to wear these devices constantly, which can be inconvenient or annoying. In addition, the need for regular charging and the potential for wear and tear pose practical problems.

Cameras, on the other hand, offer a noncontact method of accurately recording human movement. They are not intrusive in terms of user interaction, as they do not have to be worn. This allows for a natural recording of activities. However, cameras raise significant privacy concerns, as continuous video recording can be invasive. Moreover, their limited receptive field means that they can only monitor activities within their line of sight. Surveillance cameras can mitigate this problem to some extent by covering wider areas, but privacy issues remain a significant obstacle.

Radar-based systems are an interesting alternative, as they solve the privacy problems associated with cameras. These systems can detect movements and deduce activities without capturing detailed visual information, thus protecting users' privacy. However, like cameras, radar systems have a limited receptive range, which limits their monitoring capabilities to specific areas.

Environmental sensors, such as microphones, offer another possibility for recognizing human activity. These sensors are environmentally friendly and can be integrated into the environment. However, their effectiveness in recognizing human activity is limited compared to other types of sensors.

Sensors embedded in objects or parts of the home, such as door sensors, water flow meters and light switches are noninvasive and do not require direct user interaction, making them convenient and unobtrusive. They can effectively infer certain activities based on the user's interaction with the environment. However, their ability is limited by the user's need to interact with these objects and they may not capture activities that do not involve such interactions.

Each of these sensing technologies has a number of strengths and limitations. Understanding these trade-offs is crucial to selecting the most suitable sensing technology for a given application in HAR.

4) Evaluation Methodologies: Typically, the evaluation of activity recognition techniques relies on executing experiments with datasets collected from subjects performing various activity instances, each labeled with the correct activity class [42]. In unsupervised methods, the recognition performance of the algorithm is evaluated using the entire dataset. In methods relying on supervised machine learning techniques, the dataset is partitioned into *training set* and *test set* [43]. The former is used to train the recognition model, while the latter is used to measure the recognition performance.

Different approaches exist for choosing which data to use for training and testing the model. A simple approach, named *hold-out* validation, consists of using a fixed percentage (e.g., 70%) of the dataset for training and the remaining part for testing purposes [43]. A shortcoming of this approach is that it uses only a portion of the dataset for testing the recognition performance, reducing the significance of the experiments. This limitation is avoided by a widely used approach known as *k-fold cross validation* [44]. With this approach, the dataset is partitioned into k (e.g., 10) folds of equal size. One fold is used as test set, while the remaining folds are used for training the model. This procedure is repeated k times, with each fold serving as test set once. This ensures that every instance in the dataset is used for testing, providing a more robust evaluation of the recognition performance.

Regardless of the chosen validation approach, various methods exist for partitioning data into training and test set folds [45]. A common method is *random partitioning*, where data are shuffled and randomly assigned to folds. This approach should generally be avoided for time series data, such as wearable sensor data, as they are often preprocessed using sliding windows to extract feature vectors. Indeed, consecutive vectors can be highly similar, especially with overlapping windows, leading to similar vectors in both training and test sets and giving an artificial advantage to the algorithm. This can result in overfitting and poor model generalization. *Stratified partitioning* ensures each fold has the same class distribution as the dataset, which is useful for imbalanced datasets to maintain representative training and test sets.

Partitioning methods can be categorized into subject-dependent and subject-independent ones [46]. In *subject-dependent partitioning*, data are split so that instances from the same subject appear in both the training and test sets, allowing the activity recognition algorithm to learn subject-specific patterns and potentially improve recognition performance. This method simulates a scenario where a generic model is fine-tuned with examples from the current user, but may be intrusive due to the training effort required. It is well suited for challenging tasks where data from other individuals may not suffice for achieving high recognition rates. *Subject-independent partitioning*, on the other hand, ensures that each subject's data is entirely in either the training or test set, evaluating the model's ability to generalize across different individuals. This method is ideal for activity recognition systems that need to perform well for unseen users without needing subject-specific training.

5) Activity Recognition Metrics: In the simplest formulation, activity recognition can be modeled as a binary classification task where the goal is to predict whether an activity instance i belongs to a class C (e.g., "cooking") or not. Given an instance i, the outcomes include true positive (TP), where i is correctly assigned to C; false positive (FP), where i is wrongly assigned to C; true negative (TN), where i is correctly predicted not to belong to C; and false negative (FN), where i belongs to C but is not predicted as such.

To evaluate the recognition performance of activity recognition algorithms, different metrics are used in the literature [43], which include *accuracy*, calculated as $(\mathrm{TP}+\mathrm{TN}/\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$; *precision*, defined as $(\mathrm{TP}/\mathrm{TP}+\mathrm{FP})$; *recall* (also known as true positive rate or sensitivity), expressed as $(\mathrm{TP}/\mathrm{TP}+\mathrm{FN})$; and the F_1 *score*, which is the harmonic mean of precision and recall. *Specificity*, also known as true negative rate, is measured as $(\mathrm{TN}/\mathrm{TN}+\mathrm{FP})$. *Mean Absolute Error* (MAE) is sometimes used, calculated as $1-$accuracy.

For multiclass activity recognition, these metrics are extended by averaging scores for each class using either microaveraging (weighted by the number of instances of each activity class) or macroaveraging (unweighted). When class imbalance is an issue, metrics like geometric mean (G-mean) are preferred to balance recognition across majority and minority classes [47]. For binary classification, G-mean is computed as the square root of the product of recall and specificity, while for n-class problems, it is the nth root of the product of recalls for each class.

Other commonly used metrics in activity recognition focus on the efficiency of algorithms (e.g., computational costs and power consumption), scalability to the number of users, and system durability. Since they provide different perspectives on the effectiveness of an activity recognition system, it is beneficial to evaluate multiple metrics to gain a comprehensive understanding of the activity recognition system.

III. Materials and Methods

This section details our systematic literature review process for selecting studies that meet our inclusion criteria concentrating on the examination of GAR through the deployment of different variety of sensors technologies within an indoor settings. In addition to this, we have also outlined the criteria for selecting and excluding studies, as well as provided an comprehensive science mapping analysis of the chosen research works.

A. Search and Selection Strategy

Fig. 2 presents the flowchart of our search and selection strategy adopted to conduct the literature review and retrieving relevant papers. To gather a comprehensive collection of relevant studies, we conducted a search across reputable scientific databases such as *Web of Science*, *ScienceDirect*, *PubMed*, and *Scopus*. The procedure for our literature review and selection is organized into three key phases. This structured approach allowed for a systematic and thorough review of the literature, guaranteeing that only studies of significant relevance and impact are selected for our analysis. This detailed and systematic strategy highlights our dedication to achieving thoroughness and accuracy in mapping out the relevant studies in the field.

1) Paper Extraction: Our methodology aims to search and select relevant studies focusing on GAR using sensor technologies in indoor environments. This approach ensures that we are able to identify the most comprehensive information available in this field and make informed decisions and recommendations. Therefore, to extract relevant studies, we deployed a carefully designed search query. This query aimed at identifying studies of interest by focusing on specific keywords, titles, and abstracts.

We customized the query to align with the unique search mechanisms of each database, ensuring our search goals are consistently meet across different platforms, as outlined in Table I. The whole query is carefully designed into five distinct sections, joined by the "AND" operator. The first section covers a broad range of research on "graph-based" studies, encompassing a diverse array of terminology related to various activities and behaviors, particularly those that are relevant to daily life and routines. The second section concentrates on literature that utilizes sensor technologies and pervasive computing, including everything from IoT applications to wearable and non-invasive devices, highlighting smart environments and assistive technologies. The third section narrows the search to studies

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

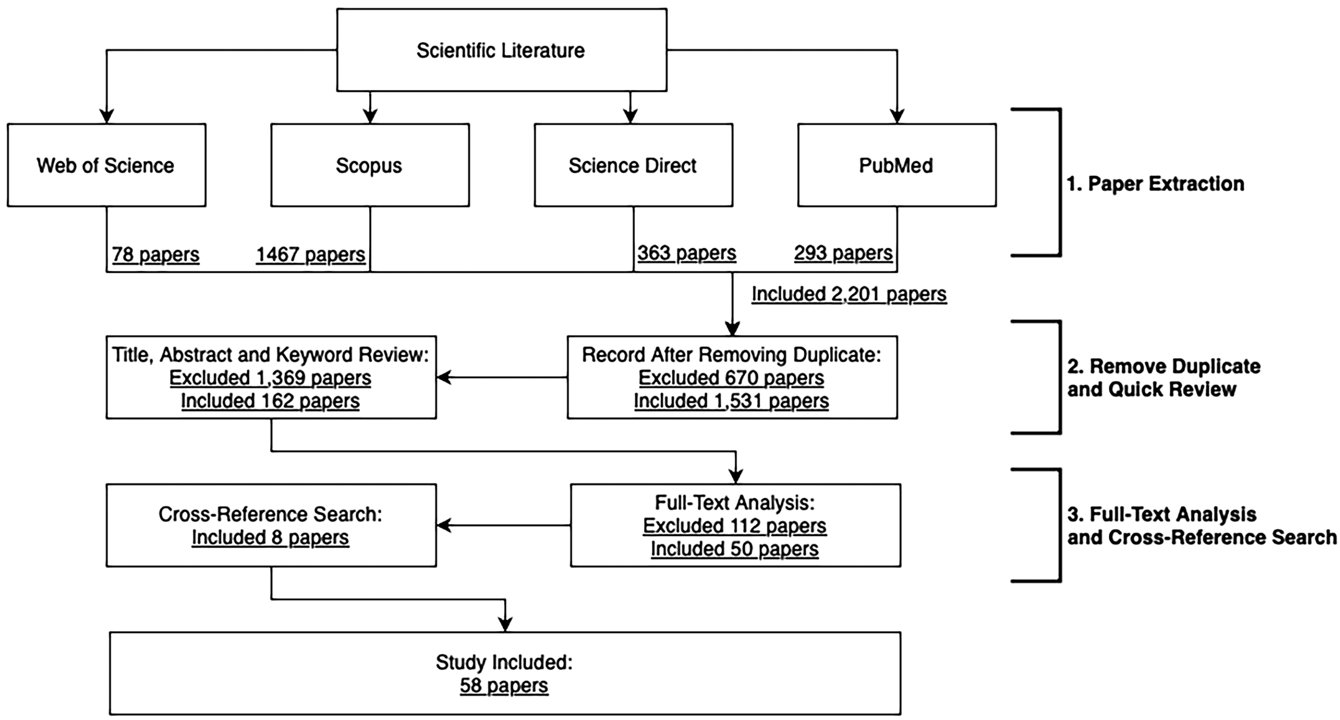

Fig. 2. Flowchart of our literature search and selection method.

TABLE I
LITERATURE SEARCH QUERY

TITLE-ABS-KEY ((“graph”) AND (“activity” OR “behavior” OR “behavior” OR “activities of daily living” OR “ADL” OR “daily life” OR “activity-aware” OR “daily routine” OR “action”) AND (“IOT” OR “indoor” OR “web of things” OR “WoT” OR “sensor” OR “GPS” OR “accelerometer” OR “gyroscope” OR “mobile device” OR “mobile data” OR “wearable” OR “positioning technology” OR “smart home” OR “ambient intelligence” OR “intelligent technology” OR “intelligent system” OR “noninvasive” OR “intelligent assistive” OR “device” OR “smartphone” OR “human computer interaction” OR “mobile health” OR “healthcare” OR “ambient assisted living” OR “assistive technology” OR “smart environment”) AND (“detection” OR “analysis” OR “classification” OR “reconstruction” OR “monitoring” OR “assessment” OR “recognition”) AND PUBYEAR $>$ 2009 AND PUBYEAR $<$ 2025 AND (LIMIT-TO (DOCUMENT TYPE, “(PEER REVIEWED JOURNAL) ARTICLE”) OR LIMIT-TO (DOCUMENT TYPE, “CONFERENCE PAPER”) OR LIMIT-TO (DOCUMENT TYPE, “REVIEW”) AND (LIMIT-TO (LANGUAGE, “English”))

that provide methods for detecting, analyzing, classifying, and monitoring these activities. In the fourth section, only publications dated between 2010 and February 2024 are considered, ensuring that the results are current. The last section retrieves only papers published in conference proceedings or scientific journals, all in English.

This approach led us to successfully identify 2201 papers, establishing a comprehensive foundation for our investigation into the utilization of graph-based approaches and technological tools in the monitoring and analysis of daily activities.

2) Duplicates Removal and Quick Review: During the second step, we first removed duplicate articles, reducing the count to 1531 papers. Next, we conducted a quick review by examining the title, abstract, and keywords of each paper to ensure alignment with our criteria. This preliminary screening was necessary due to the high volume of unrelated papers initially retrieved, many of which lacked a focus on graph-based methods, activity recognition, or were irrelevant to ambient and wearable sensors within smart indoor environments. After this refinement, 162 papers met our research criteria.

3) Full Text Analysis and Cross-Referenced Search: In the third phase, we thoroughly reviewed the full texts of remaining papers, focusing on studies with relevant topics, robust data and methodologies, and a clear contribution to indoor activity recognition. We excluded studies without experimental evaluation, those unrelated to daily activity recognition, or those focused solely on outdoor environments. This phase also allowed us to remove preliminary versions of previously considered studies, ensuring a selection of high-quality research aligned with our criteria.

After this filtering, 50 papers were retained. In a final step, we examined references in these papers to identify additional

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

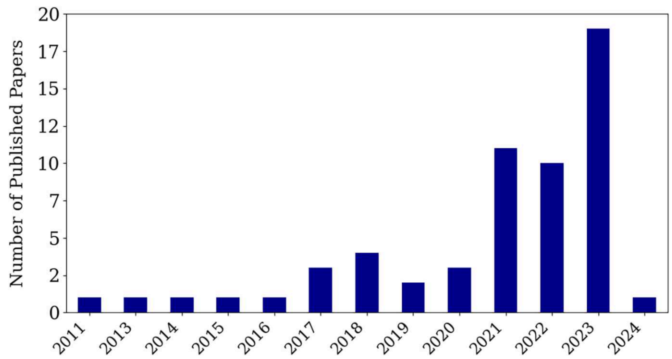

Fig. 3. Number of papers based on the different years.

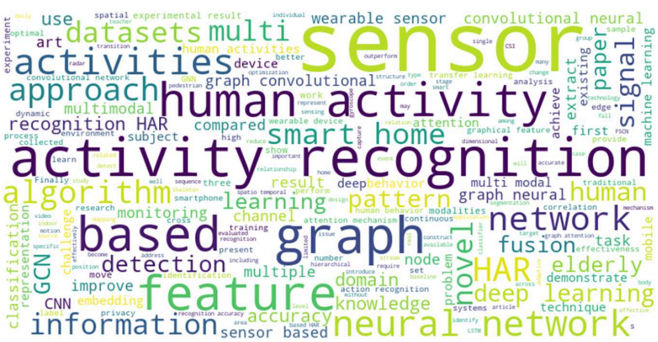

Fig. 4. Visual representation of key term frequencies.

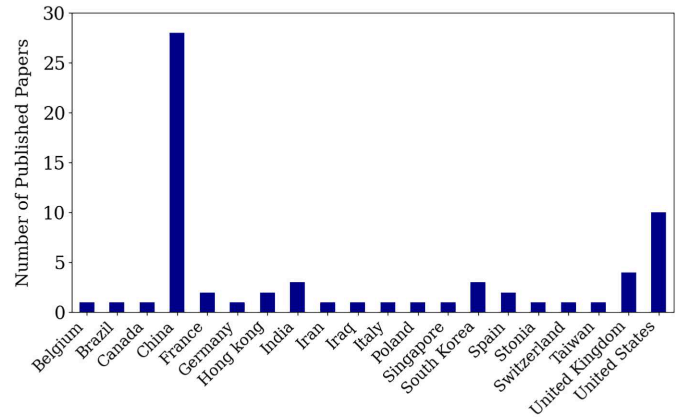

Fig. 5. Number of papers based on the different countries.

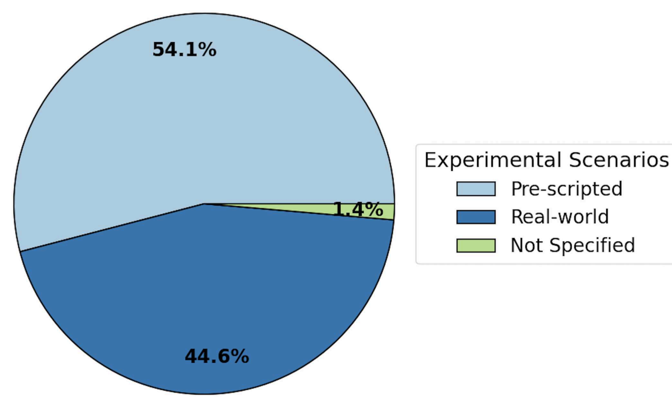

Fig. 6. Percentage of papers based on different experimental scenarios.

relevant studies, adding eight more papers to our review. Ultimately, our selection was refined to a total of 58 papers, listed in the *Supplementary Materials*.

B. Comprehensive Science Mapping Analysis

Recently, bibliometric measurements have been gaining significant attention across diverse fields [48], [49]. This method is especially adept at facilitating comprehensive science mapping analyses of published studies in fragmented and contentious realms of research.

1) Annual Scientific Production: The annual trend of scientific production is demonstrated in Fig. 3, which shows a notable rise in the number of studies focusing on graph-based analyses of activities and behaviors using pervasive sensor technologies. This rise reflects growing research interest and progress in the fields of IoT, wearable devices, and ambient assisted living, particularly within the domain of activity recognition and monitoring from 2010 to February 2024.

2) Word Cloud: Word clouds are a visual tool that highlights the primary themes related to a specific subject by condensing and emphasizing the most frequently used terms. As shown in Fig. 4, the most common terms in the selected literature are presented, revealing the prominent role of sensors and GAR and behavior analysis. This pattern reflects the central themes and interests that are currently driving research trends.

3) Country-Specific Production: Fig. 5 illustrates the geographic distribution of research contributions covered in this survey, which encompasses 20 countries where applicable case studies have been conducted. This visual representation indicates that the most significant contributions in the field of GAR in indoor environments come from China, the United States, the United Kingdom, and the South Korea. Upon closer examination, it becomes apparent that China holds the largest share of publications (42.43%), with the United States (15.16%), the United Kingdom (6.07%), the South Korea and India (4.54%) ranking subsequent.

4) Scientific Production Based on Experimental Scenarios and Dataset Usage: The field of GAR has been evolving rapidly over the years, with new experimental scenarios and dataset usage being explored constantly. In recent explorations within GAR, the diversity in experimental scenarios and dataset applications presents a distinctive narrative of current research trends. Our analysis, illustrated in Fig. 6, categorizes the scenarios predominantly used in the studies surveyed. Studies with pre-scripted scenarios make up the majority at 54.1%, followed by modeling real-world settings at 44.6%. The careful selection of experimental scenarios is crucial for the advancement of GAR technology.

5) Scientific Production Based on Experimental Environment: The categorization of experimental environments in GAR research offers significant insights into the preferred settings for data collection. Based on the methods used for data

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

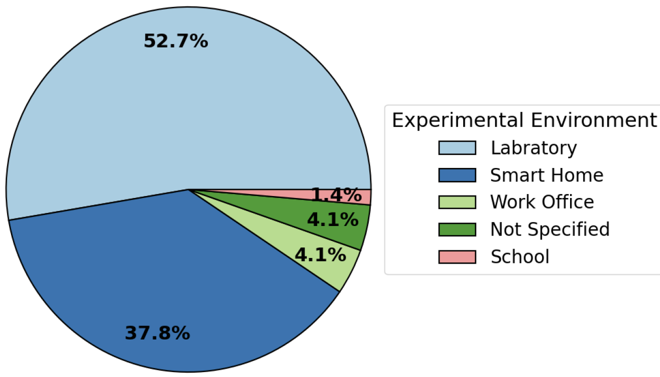

Fig. 7. Percentage of papers based on the experimental environment.

collection and the datasets employed, research environments can be broadly classified into several key categories.

The primary environments identified, as shown in Fig. 7, include Smart Homes, Laboratories, Schools, Work Offices, and some cases where the environment has not been specified. As anticipated, a substantial majority of the studies, accounting for more than half, are conducted in laboratory settings. This preference for laboratories can be attributed to the controlled conditions they offer, facilitating the precise calibration and evaluation of sensor systems and algorithms. Following laboratories, Smart Home environments represent over 37.8% of the research settings. These environments are crucial for the development and testing of practical applications, reflecting a growing interest in enhancing everyday life through smart technologies.

Schools and work offices, although less represented, provide unique contexts for studying group dynamics and individual behaviors in structured settings. Meanwhile, the category "Not Specified" highlights a gap in reporting that could affect the reproducibility and applicability of the research findings.

6) Dataset Analysis: A total of 50 datasets were used in the 58 selected studies, most of them publicly available. This testifies to the large amount of data that can be used to train and evaluate HAR approaches. On the other hand, the heterogeneity of the different datasets in terms of number of sensors, type and location may pose a challenge in maximizing these resources to develop effective and robust GAR systems. Table II provides a comprehensive summary of the most used datasets for GAR, in terms of number of subjects, number of devices, type of sensors, their placement (on the body, on objects, environmental) and number of activities. This allows the completeness and complexity of each dataset to be assessed and facilitates comparison between different databases. Only the datasets used in at least three studies are reported. It is important to note that these datasets were selected based on our inclusion criteria and the studies we analyzed. While other benchmarks, such as SensorLLM [50], DAGHAR [51], KU-HAR [52], HARSENSE [53], and RealLifeHAR [54], are valuable for HAR research, they are not included in our analysis as they do not appear in the studies we reviewed for multimodal indoor GAR—a central focus of our research.

UCI-HAR [55], UCI-HAPT [56], and WISDM [57] represent the largest public datasets in terms of the number of subjects (29–30 subjects), followed by the MMAct [58] dataset (20 subjects). The remaining have less than 15 participants involved in the experiments. HAR datasets are often collected within the context of specific applications, such as detecting cognitive impairment. Consequently, collected activity data may be biased toward certain populations, such as older adults. To mitigate potential bias, when evaluating HAR methods it is crucial to consider participant diversity, including attributes such as age, gender, and health conditions. However, while most dataset companion papers provide age and gender distributions, other important information, such as health status and education level, is sometimes omitted.

In terms of sensors, CASAS [59], Opportunity [60], and MMAct datasets provide the most comprehensive set of sensing technologies. Specifically, the CASAS and Opportunity datasets include more than 70 devices positioned on the subject's body, on specific objects (e.g., glass) and home furnitures, as well as environmental sensors. In more detail, accelerometers, gyroscopes, switches, water flow, light, temperature, and pressure sensors are employed. The MMAct dataset includes on-body, on-object and environmental sensors. Specifically, camera, smart glasses, smartphones, and wi-fi and pressure sensors record user activity data. The UTD-MHAD [61] dataset includes multimodal data extracted from an RGB camera and an inertial sensor on the subjects' wrist or upper leg. The remaining datasets include wearable devices, mostly inertial sensors embedded in smartphones or inertial measurement units (IMUs), sometimes in combination with heart rate monitors. In four datasets (UCI-HAR, USC-HAD [62], UCI-HAPT, WISDM), a single wearable device is used for data collection (either a smartphone or IMU). Some dataset acquisition testbeds are designed to support a wide range of scenarios, while other ones are tailored for specific applications. For instance, CASAS smart homes are well-suited for diverse applications within smart home environments, particularly when the recognition of complex activities is required. However, the limited presence of fine-grained sensors in CASAS makes it challenging to detect simpler actions and activities. Conversely, datasets relying solely on inertial sensors are well suited to detect gestures and physical activities, but are often insufficient for recognizing more complex activities and behaviors. Environmental constraints are another critical factor when evaluating HAR datasets. Certain sensor types require fixed placement and are limited to specific locations, such as passive infrared sensors for localization or item sensors in CASAS testbeds. Those sensors are effective for monitoring localized activities but are restricted in coverage, making them less suitable for tracking activities across larger or dynamic spaces. Other sensors, such as wearable sensors, provide greater flexibility, allowing continuous tracking of users across different environments. However, in a real-life implementation, they may introduce challenges related to user compliance, battery life, and data transmission. Hence, when choosing a dataset for HAR experimentation, it is necessary to carefully consider the real-life requirements of the target applications.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
DESCRIPTION OF THE MOST WIDELY USED DATASETS IN MULTIMODAL GAR

Dataset	# Subjects	# Devices	Sensors	Location	# Activities	# Studies
UCI-HAR	30	1	Smartphone (Acc, Gyr)	On-body	6 (postures and simple activities)	10
PAMAP2	9	3	Acc, Gyr, Mag, ECG	On-body	24 (postures, simple and complex activities, physical activities)	9
CASAS	variable	up to 71	Acc, Gyr, Switches, flow, temperature, light, pressure	On-body, on-objects, environmental	Variable	9
UTD-MHAD	9	2	IMU (Acc, Gyr), Kinect	On-body, in-front	27 (gestures, actions, activities)	7
Opportunity	12	73	Microphone, Acc, Gyr, localization, switches, pressure)	On-body, on-object, environmental	35 (actions, gestures, activities)	7
mHEALTH	10	3	IMU (Acc, Gyr, Mag), ECG	On-body	12 (postures, activities, physical exercises)	5
USC-HAD	14	1	IMU (Acc, Gyr, Mag)	On-body	12 (postures and simple activities)	4
UCI-HAPT	30	1	Smartphone (Acc, Gyr)	On-body	12 (postures, postural transitions, and simple activities)	3
WISDM	29	1	Smartphone (Acc)	On-body	6 (postures and simple activities)	3
MMAct	20	7	Camera, smart glass, smartphone (Acc, Gyr), Wi-Fi, Pressure	On-body, in-front, environmental	37 actions (complex, simple, desk)	3

Note: Acc, accelerometer; Gyr, gyroscope; ADLs, activities of daily living; IMU, inertial measurement unit; Mag, magnetometer.

In terms of sensor position, the waist (UCI-HAPT, UCI-HAR, USC-HAD) and front pocket (WISDM, MMAct) were the preferred location for inertial sensors when using a single wearable device for HAR. When multiple on-body sensors were used, they were placed on the chest, wrist and ankle (PAMAP2 [63], mHEALTH [64]). On-objects inertial sensors were positioned on the table, chair, or doors for motion monitoring (Opportunity), while pressure sensors on table and plate (Opportunity) or medicine container (CASAS) can detect usage.

The use of sensors in GAR presents heterogeneity. In some studies, nodes represent sensors in the same device. For example, in [65], three nodes corresponding to accelerometer, gyroscope, and magnetomer embedded in a smartphone placed on the pocket, or in [66], accelerometer, gyroscope, and heart rate monitor in a chest-worn wearable device. When data were available for multiple devices on different body locations, each sensor (in the same or different device) was used as node [65]. In some cases, nodes represent aggregated feature vectors (statistical and spectral features) from different sensor readings over specific time windows [67], [68]. In other cases, nodes represents different sets or clusters of sensor data [69]. Notably, different graph approaches were explored in [70], including body graph (edges of devices connected in the human body are set to 1), sensor modality graph (edges of sensors of the same type are set to 1), and data pattern based graph (edges are constructed by measuring the similarity of data patterns); the three graph modalities provided similar results. On the other hand, most studies defined edges by similarity measures between sensors, sets of sensors data, or clusters.

In terms of activities, datasets comprising only on-body sensors have between 6 and 24 activities. This rises to 27–37 when combined with cameras, on-object and environmental sensors. A detailed description of the activities included in the different datasets is reported in the following, from simple to more complex activities/datasets. The UCI-HAR includes general activities (walking, walking upstairs, walking downstairs) and postures (sitting, standing, laying). The UCI-HAPT dataset further includes six postural transitions (stand-to-sit, sit-to-stand, sit-to-lie, lie-to-sit, stand-to-lie, lie-to-stand). In WISDM, five activities are in common with the UCI-HAR dataset, and the only difference is the presence of jogging instead of laying. The USC-HAD dataset includes diverse walking modalities (forward, left, right, upstairs, downstairs), physical activity (running, jumping), postures (sitting, standing, sleeping), and elevator (up, down). mHEALTH comprises postures (standing, sitting, lying), general activities (walking, walking upstairs), physical activities (jogging, cycling, running, jump) and stretching exercises (waist/knees bending, arms elevation). MMAct includes postures (sitting, standing), postural transitions (sit-to-stand, stand-to-sit), physical activities (jumping), interactions with objects (carrying, open/close fridge, pull, pick up), talking, enter/exit room, and using PC. Opportunity includes postures (lying, sit, stand), postural transitions (sit-to-stand, stand-to-sit), walk, drink, daily activities (prepare sandwich, clean, prepare coffee, eat), and actions (open/close the fridge/dishwasher/drawers/door, turn on/off the lights). The UTD-MHAD dataset comprises hand gestures (arm swipe to the right/left, wave, clap, throw), cross arms, draw (X, circle, triangle), bowling, boxing, basketball shoot, swing (baseball, tennis), actions (catch object, pick up and throw,

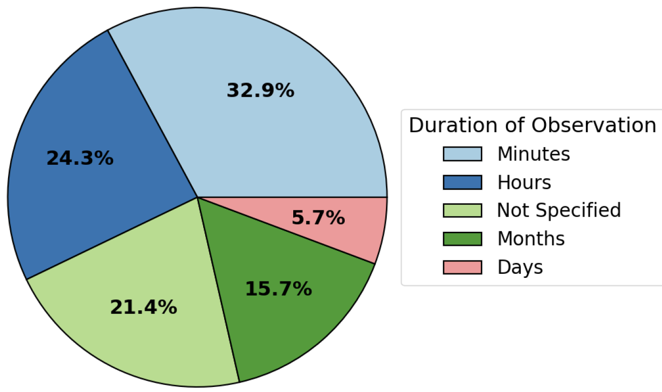

Fig. 8. Percentage of papers based on the duration of observations.

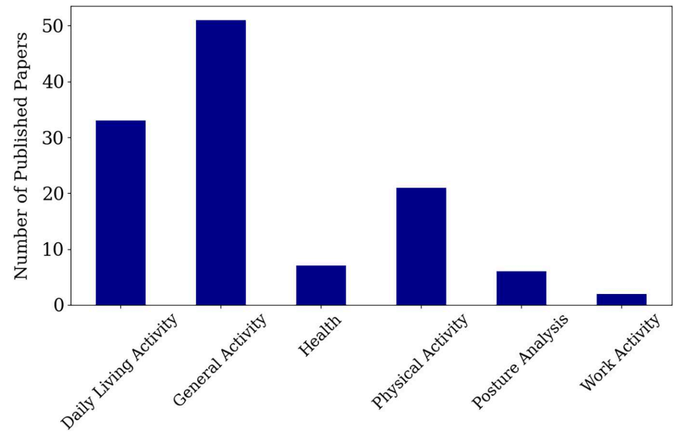

Fig. 9. Frequency of activity types across research papers.

knock on doors), general activities (walking), postural transitions (sit-to-stand, stand-to-sit), and physical exercises (lunge, squat, jogging). PAMAP2 includes postures (standing, sitting, lying), general activities (walking, walking upstairs, downstairs), physical activities (running, cycling, nordic walking, playing soccer, rope jumping), daily activities (watching TV, car driving), and work activities (computer work). Other activities include vacuum cleaning, ironing, folding laundry, house cleaning, and other (transient activities). Finally, there is a large number of datasets comprised in CASAS, with different samples, experiments, sensor settings, and activities. Daily activities can include sleeping, eating, cooking, bathing, watching TV, and other routine tasks. Instrumental activities can include preparing meals, housekeeping, medication management. Finally, complex activities comprise social interactions, concurrent activities, and transitions between different activities.

7) Scientific Production Based on Duration of Observations: The duration of observations in GAR research varies widely, influencing the depth and applicability of findings. As shown in Fig. 8, 32.9% of studies conduct experiments lasting minutes, typically within prescripted scenarios under controlled conditions. Following this, 24.3% of studies span hours, providing a broader range of activity data. Notably, 21.4% of studies do not specify their observational duration, highlighting a reporting gap that can impact the reproducibility and context of results.

Long-term observations are less frequent but vital for capturing sustained behavior patterns. Studies extending over months account for 15.7%, often linked to real-world settings such as smart homes. This trend suggests that real-world scenarios require longer observation periods, while pre-scripted scenarios favor shorter durations. Matching the observation period to the research objectives is critical to advancing GAR technology and ensuring its practical applicability.

8) Scientific Production Based on Activity Recognition Type: This section examines the distribution of activity recognition systems based on the activity categorization presented in Section II-2. The distribution is shown in Fig. 9. This segmentation of sensor types and activities highlights GAR's diverse applications and technological innovations, demonstrating its potential to improve lifestyles and healthcare.

IV. Graph-Based Activity Recognition

In this section, we will go through classified multimodal algorithms for GAR. The classified methods are summarized in Table III.

A. Graphical Models

In this group of approaches we explored the application of *dynamic graphical models (DGMs)* and *graph network* for HAR in smart indoor environments. Both models provide a robust framework for mapping the relationships between sensor data and activity labels, each with its own advantages and disadvantages. In the following sections, we will examine these models in more details.

1) Dynamic Graphical Models for HAR: DGMs offer a powerful framework for capturing the relationships between sensor data and activity labels. Traditional DGM methods like naive Bayes classifiers (NBCs) and hidden Markov models (HMMs) provide a foundation, but limitations arise in capturing temporal information or dependencies between sensor readings [71], [72]. Researchers have addressed these limitations through various techniques that enhance the capabilities of DGM-based methods.

One approach, incorporates virtual sensors representing time or activity context, as demonstrated by Ghasemi et al.'s [72] work with periodic time spans. Keyvanpour et al. [73] further expanded on this concept by introducing multimodal virtual sensors that integrate statistical features with activity context details such as the number of triggered sensors, day of the week, and sensor locations. Graph-based regularization to identify informative sensor combinations is other approach investigated by Twomey et al. [74]. Their approach learns the sensor topology of a home, represented by an adjacency matrix, and identifies useful combinations in advance, achieving better performance than exhaustive methods with minimal computational cost. This method offers a computationally efficient way to exploit the inherent relationships between sensors for activity recognition.

DGMs can also be applied to indoor positioning tasks [75] by combining activity recognition algorithms such as HMM with

TABLE III
GAR APPROACHES AND CHARACTERISTICS

GAR Approaches		Main Idea	Characteristics	Challenges	# Studies
Graphical models	*DGMs for HAR*	DGMs are widely used in HAR for leveraging relative temporal information and activity duration and capturing the relationships between sensor data and activity labels.	- Sequence modeling and leveraging temporal and spatial information; - Leverage sensor topologies to identify useful sensor combinations efficiently; - Applicable to diverse smart environments and activities.	- Computational complexity; - Data dependency and quality; - Energy efficiency; - Robustness.	6
	Graph Networks for HAR	Graph networks are using to translate sensor readings into a network of interconnected points as a common underlying approach for creating informative features from sensor data and HAR.	- Construction of dependency graphs to capture activity structure and facilitate knowledge transfer; - Facilitate sensor fusion, reduce manual labeling needs, and limited data of new users and labeled data from other users; - Hierarchical structures to distinguish similar activities effectively; - Capability of modeling transitions between activities.	- Handling complex, dynamic environments and creating efficient graph structures for diverse sensor data types; -Real-time adaptation, data integration, and scalability while managing computational costs.	14
Knowledge-based approaches		Knowledge-based approaches in HAR use domain expertise, ontologies, knowledge graphs, and rule-based systems to enhance accuracy and interpretability.	- Leverage ontologies, knowledge graphs, and rule-based systems to integrate expert knowledge into HAR; - Develop sensor logical topologies and learn sensor embeddings to improve GAR performance; - Use inference algorithms to identify activities based on predefined parameters and contextual information; - Capability of adaptive knowledge distillation frameworks to transfer knowledge between sensor-based and vision-based models.	- Handling and integrating diverse sensor data; addressing computational intensity of pre-processing, feature extraction, and embedding learning; - Ensuring scalability, real-time adaptation, and expert knowledge integration; improving model interpretability; - Managing data sparsity; and maintaining high accuracy in activity recognition while minimizing manual labeling.	6
Deep learning GAR approaches	*GCN-based approaches*	GCNs improve HAR by modeling complex relationships with diverse graph structures, enhancing feature interactions and performance, particularly in scenarios with limited data.	- Handling complex activities and data scarcity by using diverse graph structures for different sensor modalities, body anatomy, and temporal dependencies; - Extracting interconnections to improve feature interactions and model performance; - Addressing data limitations, privacy, and battery concerns; - Combining GCNs with other methods; - Employing data augmentation, few-shot learning, and multi-task learning to capture temporal and spatial dynamics.	- Acquisition of large-scale labeled datasets; - Ensuring privacy and data security; - Battery maintenance for continuous monitoring; - Need for real-time processing and adaptation to diverse environments; - Reliance on quality of sensor data.	22
	CNN-based approaches	Leveraging advanced deep learning techniques, particularly CNNs with various data processing and graph modeling strategies to effectively recognize and classify human activities from sensor data.	- High accuracy, efficiency, and enhanced robustness; - Spatial and temporal data processing; - Minimal preprocessing since they can directly process raw sensor data, minimizing the need for extensive preprocessing; - Versatility.	- Model complexity; - High computational cost and memory requirements; - Optimizing data preprocessing techniques to reduce computational overhead while maintaining accuracy; - Generalization.	2
	GANs	GANs improve activity recognition by capturing temporal and spatial correlations in sensor data, leading to more accurate and robust models.	- GANs enhance feature extraction with attention mechanisms, while models like GrapHAR use graph-based representations to capture sensor data correlations, combining temporal and spatial features. Integration with methods like Convolutional layers, TCCNs, Multi-head GAT, and AGCN improves performance, generalization, and noise handling. Solutions for dynamic feature extraction and scalable models are also explored.	- High computational resources required. - Managing data reduction and efficiency; - Ensuring model interpretability; - Handling noisy and complex environments; - Adapting models to new tasks with limited data.	8

Note: DGM: dynamic graphical model; GCN: graph convolutional neural network; CNN: convolutional neural network; GAN: graph attention network.

crowdsourced Wi-Fi fingerprints to match walking trajectories to an indoor map represented as a directed graph. This approach effectively tracks individuals without prior knowledge of their starting position and handles structural complexities of indoor environments. DGMs are also able to analyze complex activities within smart environments. Prieto et al. [76] have been explored extracting "orchestration graphs" from wearable sensor data to analyzing dynamic classroom activities and assessing teaching practices in real time. These graphs depict teaching activities and social interactions over time.

Future research in DGM-based HAR focuses on key advancements such as real-time adaptation to new activities and sensor configurations, energy-efficient computation, and managing model complexity for practical deployment. Automatic indoor map generation could simplify positioning, while integrating more sensor data sources would enhance activity analysis. Long-term and synthetic data collection can improve model accuracy and generalizability. Addressing these challenges will make DGM-based methods essential for robust HAR systems in smart environments.

2) Graph Networks for HAR: Different studies have shown that graph Networks offer innovative ways to interpret sensor data by representing sensors as nodes, consecutive activations are connected by undirected edges, and repeated activations are denoted by self-loops [77], [78], [79], providing a deeper understanding of human activities and sensor activation.

In smart home environments, Augustyniak et al. [10] decomposes activities into elementary poses, represented as nodes in a graph. This time series of graphs provides a compact way to represent and predict human behavior within a smart home environment. Similarly, Li et al. [80] leverage a behavior-aware flow graph to identify activities and daily routines of elderly individuals, promoting better care and support.

Moreover, ActiLabel [67], paves the way for efficient recognition in new domains by constructing dependency graphs where nodes represent activity clusters identified from sensor data, and edges depict the relationships between them. These graphs capture the inherent structure of activities, allowing knowledge to be transferred across situations without the need for extensive data collection in each new environment. Complementing this, NN-FGO [81] enhances indoor navigation through sensor fusion by integrating data from wearable sensors with UWB radio signals. This approach constructs a dynamic graph that accurately represents the user's location and movements, ensuring precise navigation in GPS-unavailable environments. By identifying loops and movement patterns, the system imposes virtual constraints, thereby maintaining accurate navigation even with limited UWB data.

Additionally, OptiMapper [66] tackles the challenge of personalized recognition with minimal data by creating a network that links a new user's limited activity data to labeled data from others. This network identifies patterns in the new user's data by leveraging similarities with existing observations, although initial calibration may be needed for best results. Aljanabi et al. [82] enhance this by combining data from multiple sources, such as smartphones and smartwatches utilizing a multilevel optimization model for comprehensive data analysis. This approach involves creating a graph structure from the sensor data and then employing sophisticated algorithms to optimize subgraphs within the network, leading to enhanced accuracy but increased computational complexity. Additionally, Hu et al. [83] utilize semisupervised learning with graph-based label propagation for labeling activities in smart homes, where significantly reduces the need for manual labeling while maintaining high accuracy.

Several studies delve into hierarchical GAR such as HierHAR [84] and the work by Ataya et al. [85], utilize hierarchical and temporal graph structures to distinguish between similar activities, model transitions as well as understanding of the flow and sequence of movements. These methods improve the recognition of complex and sequential activities but come with higher computational demands. In the realm of crowd behavior analysis, researchers like Roggen et al. [16] used pattern recognition and graph clustering of accelerometer data to gain insights into group dynamics, although scalability remains a challenge. In this approach they tried to translate individual movements as nodes, and connections between similar movements as the edges in the graph.

For real-time recognition, Vieira et al. [86] proposed speed-aware action graphs for immediate recognition of unsegmented data streams. A graph where nodes represent key poses within an action, and edges depict the transitions between them. This approach provided promising results in terms of accuracy in real-time activity recognition but requiring high processing power.

Overall, Graph Networks are becoming key tools for analyzing movement and activities in sensor data. By enabling personalized recognition and real-time applications, they help devices better interpret and respond to human activities. As research advances, more sophisticated methods will emerge, leading to a future where devices can seamlessly understand complex movement patterns.

B. Knowledge-Based Approaches

Knowledge-based approaches in HAR leverage domain-specific knowledge to enhance the accuracy, to fuse multimodal data and reduce the dependence on large-scale training samples, and interpretability of recognition systems [87]. These methods often utilize ontologies, knowledge graphs, and rule-based systems to integrate expert knowledge into the recognition process. Compared to traditional graph networks, knowledge-graph approaches improve HAR by incorporating complex semantics and relationships between domain entities, enhancing interoperability and reasoning. These graphs use subject–predicate–object triples, typically represented in the resource description framework (RDF) [88], to capture relationships between people, objects, locations, actions, and activities. To derive further information, rule-based query languages like SPARQL [89] are commonly used to infer additional knowledge from these triples. Knowledge graphs are often integrated with ontologies, expressed in OWL web ontology language [90], to formalize entity relationships and support reasoning, such as logical subsumption. This integration enhances system interpretability and effectiveness across diverse systems, which is particularly valuable in heterogeneous domains like healthcare and smart environments [91].

Recently, Wang et al. [92] proposed a novel technique comprised of two phases including sensor topology learning using data-driven adjacency and graph embedding. The method entails the preprocessing of streaming data, the extraction of features, the determination of correlation thresholds, and the fusion of adjacency matrices to obtain sensor adjacency. This is followed by the learning of embeddings from the sensor topology graph. The proposed model can generate explainable and computable sensor logical topologies.

To enhance activity tracking, Steenwinckel et al. proposed tracking activities by linking knowledge (TALK) [93] which creates a knowledge graph where nodes represent activities, objects, and contextual information, and edges represent relationships between them. This structure allows the system to infer activities by reasoning about the connections and patterns in the knowledge graph. Moreover, Liu et al. [94] introduced a semantics-aware adaptive knowledge distillation framework for action recognition. This method utilizes knowledge distillation to transfer knowledge from sensor-based models to vision-based models, improving the recognition performance across different modalities. The framework adapts to the semantic context of the data, ensuring robust and accurate activity recognition. Furthermore, Bermejo et al. [95] introduced an unsupervised real-time CPD method using neural embeddings for detecting fin-grained activity transitions, enhanced with graph-based postprocessing that incorporates expert knowledge. This hybrid approach uses three undirected knowledge graphs (activities, locations, and combined) and offers lower computational costs and improved generalization across environments. Finally, the intelligent activity inference engine developed by Hao et al. [69] leverages nonintrusive sensors in smart homes to mine complex behavioral patterns. This system uses rule-based methods and machine learning to interpret sensor data, providing insights into residents' activities and detecting anomalies that may indicate health issues.

These knowledge-based approaches demonstrate the potential of integrating expert knowledge and advanced reasoning techniques into GAR systems. By leveraging ontologies, knowledge graphs, and adaptive learning frameworks, these methods enhance the system's ability to understand and interpret complex human behaviors in smart environments.

C. Deep Learning GAR Approach

Deep learning, especially convolutional neural networks (CNNs), has revolutionized HAR, boosting accuracy and efficiency. The integration of graph-based approaches further enhances performance by modeling sensor data as graphs and using attention mechanisms to capture complex relationships. This leads to more robust and accurate HAR systems. In this regard, readers may find a comprehensive survey on deep neural networks for graphs in [96] of interest. In the following, we have categorized deep learning GAR approaches into three main categories including: *CNN-based approaches*, *graph CNN (GCN)-based approaches*, and *graph attention networks*.

1) CNN-Based Approaches: CNNs are pivotal in machine learning for HAR, primarily due to their ability to process and recognize patterns in complex data. Basic CNN-based methods involve direct processing of raw or minimally preprocessed sensor data, extracting spatial and temporal features that are crucial for accurate HAR as demonstrated in [97]. This work introduces an innovative cross-channel communication block with a feature encoder, a graph neural network (GNN)-based message passing mechanism, and a feature decoder. This design enables interchannel interaction for comprehensive information exchange, capturing discriminative features, making it ideal for real-time applications and resource-constrained environments.

Moreover, Ouyang et al. [98] proposed a multimodal approach using weighted voting among CNNs to enhance HAR accuracy. Sensor data is converted into a 2-D graph representing different axes, processed by two CNN models for handling one-dimensional data streams and the combined 2-D graph, respectively. The final activity recognition is determined through weighted voting, capturing a range of activities such as running, walking, and sleeping with high robustness.

Despite significant advancements, CNN-based HAR systems face several challenges. Model complexity and computational cost are significant issues, particularly for real-time applications. Solutions like Huang et al.'s shallow CNNs with cross-channel communication blocks [97] and Ouyang et al.'s multimodel weighted voting [98] aim to balance efficiency and accuracy. Future direction research could focus on improving data preprocessing, computational efficiency, temporal information processing, and privacy to enhance the robustness and applicability of CNN-based HAR systems across various environments.

2) GCN-Based Approaches: GCNs outperform traditional methods, especially in complex activities and data-limited scenarios. Key innovations include the construction of diverse graph structures based on sensor modalities, body anatomy, or temporal dependencies, as shown in [70]. These structures enable GCNs to extract complex interconnections, improving feature interactions and model performance.

There are several studies addresses privacy concerns and battery maintenance such as what is done in [68], [99], [100], [101], and [102]. Wang et al. [99] proposed an innovative method using body RFID skeletons and GCNs. Their model enhances the RFID skeleton activity graph convolution network (FSCN) by using parallel convolution modules for improved efficiency and accuracy in classifying activities such as walking, sitting, and standing, achieved high performance and robustness in various smart environment with different layouts. In [68], Sarkar et al. leveraged GCNs and federated learning (FL) for semisupervised HAR and to address above mentioned challenges. By constructing similarity graphs from sensor data, the model captures the interrelatedness and closeness of activities. FL will enhance accuracy, training speed, and preserving data privacy, making it suitable for real-world applications while ensuring scalability. In [103], [104], [100] and [105] to address privacy, the authors scanned the environment via radar-based sensors and skeleton data from Kinect sensors using a multimodal two-stream GNN approach as well as a GNN model utilizing pretrained 3-D human-joint coordinates which estimates these 3-D joint coordinates from the sparse radar data.

In [100], the collected point cloud data is processed to generate trajectory graphs of human movements. The spatial–temporal GCN (ST-GCN) model then classifies activities and detects abnormal ones. While in [104], a dynamic edges algorithm identifies key connections for each activity, which are then used by a GNN architecture. Combined with an long short-term memory (LSTM) network, the GNN captures temporal sequences, accounting for the order of data points within an activity. In other study by Zhao et al. [106], they utilized RFID technology and ST-GCNs to addressing portability and privacy challenges. Their method achieved high recognition accuracy through a tag array, comprehensive preprocessing, and action segmentation.

Furthermore, to address data limitation problem several studies proposed data augmentation techniques [103], few-shot learning [107], multitask learning [102], and chronological graph construction method [108] for HAR. In [103] data augmentation techniques such as zero-padding, Gaussian noise, and agglomerative hierarchical clustering. It used ST-GCNs to extract spatio–temporal features and an attention mechanism

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

to align multimodal features effectively for HAR. While in [102] they used graph-based LSTM and multitask learning for HAR. Multiple graphs capture diverse data aspects, and a Graph LSTM learns complex temporal dynamics. In [107] they transformed channel state information (CSI) data into a graph and optimize subgraph selection. This approach enables learning from few samples but relies on CSI data quality. Additionally, in [109] the authors used deep transfer learning of residual GCN (ResGCN) to first trained on a large dataset from one source domain. Then the learned knowledge from this pretrained model is then transferred to a new ResGCN for the target domain.

Moreover, to capturing temporal and spatial dynamics several researches in this domain demonstrated promising results such as [17], [18], [110], [111]. Their contributions mostly rely on their graph constructions approach. In [17] the graph, constructed from aligned sensor data, enhances feature extraction and classification, while in [110] the multichannel CNN captures temporal features, and the GCN models the human body as a graph to extract spatial features. This combined approach improves dynamic feature representation and activity recognition accuracy. Furthermore, in [18], an adaptive supervised multimodal fusion method (AMFM) for HAR in mobility-impaired patients is introduced. Combining spatio–temporal graph convolution networks with LSTM fully convolutional network (LSTM-FCN) and adaptive loss functions, AMFM fuses visual and inertial data, capturing spatio–temporal dynamics and excelling in detecting critical fine-grained activities such as falling.

Moreover, the PC-GNN study [112] used sEMG signals and a Pearson correlation-based graph for lower limb HAR. GNNs analyze the graph from time-domain features, enhanced by wavelet denoising and ensemble empirical mode decomposition, capturing both individual and relational data to improve classification accuracy. As research progresses, graph-based methods for HAR will lead to more efficient algorithms for real-time processing and adaptation to dynamic environments, expanding applications in smart homes, healthcare, fitness tracking, and more.

3) Graph Attention Networks: Graph attention networks (GAN) enhance the models' ability to identify and interpret crucial features by capturing temporal and spatial correlations in sensor data, leading to more accurate and robust activity recognition such as what is done in [113], [114], and [115]. For instance, in [113] Convolutional layers refine can these features and lead to activity recognition and prediction through a fully connected layer. This approach effectively handles irregular sensor activation times and spatial layouts. Moreover, in [114], the GrapHAR model improves efficiency by reducing reliance on large, complex models. It captures both temporal and subcarrier correlations from CSI subcarriers, preserving causality. GrapHAR represents CSI subcarrier correlations as a graph, with nodes for subcarriers and edges for correlations, learned adaptively by GANs. It combines a temporal causal convolution network (TCCN) with graph attention layers (GAL) and uses Dendrite Net [116] for final activity recognition, capturing logical relationships between inputs.

In order to overcome the limitations of traditional wearable devices, such as weak sensor durability and difficulty in capturing dynamic features, a novel system for HAR using a pedal musculoskeletal response based on a differential spatio–temporal LSTM and a pedal wearable device is presented in [117]. It incorporates both temporal differential information and spatial interactive features, captured through multihead graph attention networks, to enhance classification accuracy and generalization performance. Indeed, to enhance model generalization and scalability several studies have been shown promising results such as what is done in [65] and [118]. In [65] Zhu et al. leveraged multiple 1-D convolutional denoising autoencoders (CDAEs) to extract robust features from raw sensor data, and an attention-based GCN (AGCN) to construct new heterogeneous multisensor modalities. These modalities leverage sensor relationships, with the attentive fusion subnet improving accuracy by dynamically recalibrating features based on their importance. Moreover, in [118], Zheng et al. introduced a meta-learning framework with graph prototypical models and a priority attention mechanism. By constructing sample and distribution graphs, the model captures feature and distribution relationships, excelling at adapting to new tasks with limited data. Additionally, to improve data processing in the noisy and complex environments typical of smart homes and IoT systems, specially human-robot collaboration Multi-GAT [119] is introduced.

Multi-GAT uses a hierarchical multimodal framework with graphical attention mechanisms to extract modality-specific features via unimodal encoders and a multimodal mixture-of-experts (Multi-MoE) model. It captures cross-modal relationships through cross-modal graphical attention (Cross-GAT), improving feature representation and recognition accuracy, while handling noisy sensor data. GANs excel in capturing HAR patterns but demand significant computational resources and data. Future research should focus on efficiency, data reduction strategies, and model interpretability for practical applications.

V. Comparative Analysis

In this section, we will provide a detailed comparative analysis of GAR approaches, focusing on their performance and scalability.

A. Comparative Performance Analysis

Performing a comparative analysis of GAR approaches is challenging due to differences in datasets, sensor types, preprocessing techniques, feature extraction, graph construction, classifier algorithms, validation strategies, and evaluation metrics. This analysis organizes GAR methods on a per-dataset basis to create a structured comparison, focusing on the five most widely used datasets (see Table II). On the UCI-HAR dataset, diverse methods provided similar performance. The best results were obtained using a CNN with a C3 block for channel and feature interaction (accuracy 0.969) [97], a combination of decision tree and support vector machine (accuracy 0.964) [84], a combination of CNN with attention mechanism and multi-granular belief fusion (accuracy 0.961) [120], a combination of one-dimensional convolutional denoising autoencoder (1-D-CDAE) and attention-based GCN (accuracy 0.959) [65]. The performance of approaches based on continuous HMM (accuracy 0.932) or multiple simple CNNs (accuracy 0.901) was lower. Similar performance was obtained on the PAMAP2 dataset, where the combination of multihead graph attention and dense layers provided an accuracy of 0.967 [115]. GNN based on multilayer perceptron (accuracy 0.952) [111] and the combination of 1-D-CDAE (accuracy 0.947) [65] provided similar results. Performance was lower on the CASAS dataset, where the embedding of graphical features fed to an LSTM network [92] and random forest [79] resulted in an F-score of 0.831 and accuracy in the range 0.76–0.93, respectively. Conditional random fields provided significantly lower performance (F-score 0.675). On the UTD-MHAD dataset, a multimodality GNN combining temporal CNN and LSTM provided the best results (accuracy 0.987) [18], followed by a GCN (accuracy 0.902) [17]. The combination of multihead graph attention and MLP (accuracy 0.911), and CNN with attention combined with multigranular belief fusion (accuracy 0.908) achieved top performance on the Opportunity dataset. Lower performance was obtained using a GNN based on CNN for feature extraction and LSTM for classification (accuracy and F-score equal to 0.883). Interestingly, the model outperformed state-of-the-art, non-graph-based deep learning architectures, including combinations of CNN and LSTM, with and without attention, with the addition of gated recurrent unit (GRU) and further self-attention (2%–3% improvement in F-score).

B. Comparative Scalability Analysis

Scalability is critical when applying GAR approaches to larger datasets or complex environments, where variations in sensor quantity, computational resources, and environment complexity impact performance. This discussion examines scalability based on factors including dataset count for generalizability, sensor quantity (affecting graph size and computation), computational complexity, cross-dataset evaluations for robustness, and environment complexity. A few studies evaluated the proposed approach in terms of computational complexity. Huang et al. [97] implemented a CNN with a C3 block for channels/feature interaction, achieving an inference time of 95 ms. The model had a total of 870 K parameters and required 6.8 M floating-point operations per second (FLOPs). The algorithm was tested on four datasets comprising up to 17 activities and 30 subjects. Only wearable sensors (inertial sensors, ECG monitor) were addressed. Chen et al. [70] proposed a GNN based on CNN for feature extraction and LSTM for classification, with a total of 1 M parameters and requiring 35 M FLOPs. The algorithm was evaluated on three datasets that included up to 9 wearable motion sensors, 17 participants, and 33 activities. Yang et al. [17] developed a multicolumn activity graph method combined with a CNN model for processing activity graphs. The inference time was found to be 0.54 ms. The study was conducted on three datasets comprising up to 21 activities and 30 subjects. However, data from only two sensors (accelerometer and gyroscope) were processed. Dong et al. [120] proposed a CNN with attention mechanism for class prediction and MGBF for graph creation, resulting in an inference time of 0.13 ms. The approach was evaluated on two datasets including up to 12 sensors, 18 activities, and 30 subjects. All of these studies analyzed a few datasets, comprising data collected from wearable sensors (mostly inertial sensors) during prescripted activities executed in controlled environments. In addition, they did not perform any cross-dataset tests.

Zheng et al. [118] proposed a meta-learning-based graph prototypical model with priority attention mechanism. The method was evaluated on 14 datasets, diverse in terms of subjects (up to 66), activities (up to 27), sensors (up to 73, including cameras, wearable sensors, ambient and object sensors), and environments (office, laboratory, apartment). In addition, cross-dataset evaluations were carried out on multiple datasets to test the generalization capability, resulting in consistent performance across heterogeneous sensor settings and environments. Finally, cross-subject, cross-view, cross-scene, and cross-session evaluations were addressed. However, any information on the computational cost and real-time processing were given. Interestingly, Chen et al. [70] compared the performance and computational complexity of several graph-based and state-of-the-art deep learning methods. The comparative analysis showed that graph-based models such as STSGCN (spatio–temporal graph convolution network) and GraphConvLSTM (temporal convolution+graph convolution+LSTM) had a significantly lower computational complexity (5–35 M FLOPs and 0.21–1.1 M parameters) than state-of-the-art, non-graph-based deep learning approaches (47.3–116 M FLOPs, 1.74–9.7 M parameters). The latter included architectures based on the combination of CNN and LSTM, with and without attention, with the addition of GRU and further self-attention. Finally, Hao et al. [69] used formal concept analysis (FCA) to develop a graph inference engine based on Hasse diagram. The approach was evaluated on two datasets with apartment scenarios and complex activities. Data from various sensors (infrared, light, temperature, waterflow, power consumption, switches, pressure), actuators, passive RFID tags, tablets, and wearable devices were analyzed. The algorithm complexity was evaluated on the two datasets, in terms of number of classes, features, nodes, training time and inference time. Despite the significant complexity in some scenarios (e.g., $>$5000 nodes, the test time was found to be $<$1.7 s. The large variety of sensors and activities, coupled with unsupervised experiments in home-like environments, promise generalization of the findings.

VI. Discussion on Open Challenges

In this work, we explored various methods and technologies for multimodal GAR in indoor environments. In this section, we discuss the existing challenges and future research directions in the field.

1) Multiuser, Concurrent, and Interleaved Activities: Existing HAR systems often rely on over-simplified assumptions that do not accurately reflect the complexity of real-world situations. As a result, when these techniques are applied outside of the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

laboratory, they may provide lower recognition rates compared to those achieved in controlled experiments. Many of these systems assume that sensor data comes from a single individual, making them effective in controlled environments but inadequate in shared spaces such as smart homes and workplaces, where multiple people engage in activities simultaneously. In such *multiuser* scenarios, the systems receive mixed streams of sensor events that do not align with predefined single-user models, complicating activity recognition. To address this challenge, a preliminary step called "data association" is necessary, which identifies which individual triggered each sensor event [121] using probabilistic models or machine learning algorithms [122], [123]. Additionally, many HAR systems incorrectly assume users perform only one activity at a time, overlooking the common occurrence of *concurrent* activities (e.g., cooking while listening to music) or *interleaved* activities (e.g., pausing cooking to take a phone call). Recognizing these complex patterns requires specialized techniques based on pattern mining, probabilistic logics, or deep learning methods [124], [125], [126].

2) Transfer Learning: HAR in indoor environments has significantly advanced through the integration of multiple sensor modalities. GAR approaches have emerged as powerful tools for modeling complex relationships both within and across these modalities. However, acquiring large-scale labeled datasets for each modality and activity remains a significant challenge. Transfer learning offers a promising solution by leveraging knowledge from one domain to enhance performance in another.

Recent work in this area, particularly using GAR approaches, shows promise. These approaches can construct dependency graphs or hierarchical structures to capture the inherent structure of activities and facilitate knowledge transfer via graphs, especially when using multimodal sensor technologies. This knowledge transfer and sensor fusion reduce the need for extensive manual labeling, making it feasible to use limited data from new users or new environments and leverage labeled data from existing ones. Several works, such as [67], [83], [84], demonstrate the effectiveness of these methods in reducing manual labeling and improving GAR performance in varied settings.

3) Real-Time Analysis and Computational Complexity: Computational complexity is a key issue when performing GAR approaches on stand-alone portable modules with limited resources, which impose constraints on memory efficiency and battery consumption. Processing steps (e.g., data transformation, feature computation, and classification) largely affect the computational load. Therefore, these aspects must be strictly controlled by selecting the right tradeoff between computational complexity and performance. In this sense, the combination of simple temporal features with classical machine learning algorithms or the use of raw sensor data and lightweight end-to-end deep learning models may be a suitable compromise.

Real-time analysis in GAR is crucial, especially for health monitoring applications. In this context, the main challenge is to ensure low latency in data processing and decision-making. This involves minimizing data transmission and processing time. As for the latter, the accurate selection of proper segmentation approaches (i.e., choice of the window length and overlap) and the minimization/optimization of pre-processing steps (e.g., filtering, normalization, data transformation, and feature extraction) are essential to provide minimal prediction delay. Unfortunately, most of the evaluated studies did not directly report a measure of inference time, nor did they demonstrate possible real-time implementations. Moreover, detailed signal segmentation settings were often overlooked. Only a few studies [17], [69], [120] reported the model inference time, without considering other key processing aspects contributing to the prediction delay (e.g., sensor communication, signal windowing).

4) Challenges for Real Life Applications: Multimodal indoor HAR face several significant challenges when applied to real-world scenarios and applications. First, the high cost of integrating various devices, such as wearables and environmental sensors, limits their widespread use. Additionally, the battery life of these smart sensors poses problems, as they often require frequent recharging or replacement, disrupting continuous monitoring. Subject comfort and compliance represent significant concerns, as intrusive wearable devices can lead to noncompliance or changes in behavior, negatively impacting the accuracy of HAR systems. Furthermore, privacy and security issues arise from the sensitive nature of data collected, necessitating strong protections against unauthorized access. Finally, the generalization and adaptability of graph-based approaches are challenging. These systems must effectively handle various environments and populations while managing diverse sensor data and contextual factors. Achieving reliable models that perform well across different contexts without extensive retraining remains a complex task. Overall, addressing these challenges is crucial for the successful deployment of multimodal HAR technologies in real-world settings.

5) Lack of Large Annotated Dataset Acquired in Naturalistic Environments: Accurate evaluation of activity recognition systems requires extensive testing in naturalistic conditions. However, setting up experimental test-beds and conducting experiments with a large number of participants is challenging. Indeed, labeling large datasets of activity data is a labor-intensive and expensive task. In particular, labelling requires direct observation of human activities, raising significant privacy concerns. Additionally, ethical and legal requirements complicate the release of public activity datasets, especially in the pervasive healthcare domain. These challenges make it difficult to experimentally compare different solutions using standardized test-bed datasets.

6) Bias Toward Healthy Young Participants: Demographic information on participants was only available in less than half datasets. Most of the samples represented young healthy volunteers, with a minimum and maximum age of 20 (range: 18–23) and 40 (range: 27–60) years, respectively. In few cases [112], [118], data from subjects with Parkinson's disease or subjects with knee injuries were analyzed. This testifies that elderly subjects and people with chronic disorders (e.g., cardiovascular and neurodegenerative diseases) and/or mobility impairment (e.g., due to a stroke) are poorly represented in the large number of works and datasets on HAR. However, elderly and persons with motor disabilities may show movement patterns (slowness, asymmetry, shaking) very different from young subjects. In this context, it is not entirely clear to what extent the large number of GAR approaches can generalize to different populations, in which activity monitoring can provide extremely important information.

7) Integration With Emerging Technologies: The integration of graph-based approaches with emerging technologies such as deep learning and LLMs holds significant promise for enhancing HAR systems. As we discussed thoroughly in Section IV-C, combining graph-based approaches with deep learning especially GNNs, captures complex spatial and temporal data relationships. GNNs are ideal for integrating sensor data in HAR, while deep learning aids scalability for larger datasets and complex environments. Additionally, enhancing interpretability with LLMs, such as GPT-4 [127], can provide richer contextual understanding and natural language processing capabilities. When combined with graph-based methods, these technologies can significantly improve the interpretability of HAR systems. Graph-based methods naturally lend themselves to visual representations, and when paired with LLMs, they can offer detailed explanations of detected activities. This combination makes the system more transparent and user-friendly, paving the way for innovative applications in various domains and offering a solid foundation for real-world applications. To the best of our knowledge, no studies yet explore the integration of GAR and LLMs specifically for multimodal HAR in indoor environments.

VII. Conclusion

Human behaviors and activities are extremely complex, and their recognition may require knowledge of variegate data including movements, environmental conditions, physiological parameters, and several context information [128]. Those data can be acquired through different modalities by different sensors. However, data acquired from heterogeneous sensors may have different formats, spatio–temporal resolution, and accuracy. Hence, the activity recognition system must cope with data integration issues. Recently, graph-based approaches emerged as appealing tools for integrating heterogeneous data [129], [130]. Consequently, different research works investigated graph-based approaches to process sensor data for recognizing human activities. In this systematic survey, we explored 58 carefully selected peer-reviewed papers that propose the use of graph-based methods for activity recognition. The publication trend of these papers clearly shows an increasing interest in GAR. We analyzed state-of-the-art methodologies, evaluating their effectiveness in capturing relationships and dependencies between data acquired through different modalities. Through our analysis, we identified key research directions and open challenges, which may guide future developments in graph-based multimodal HAR.

References

[1] J. Gao et al., "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, 2020.

[2] A. Ghorbanali and M. K. Sohrabi, "A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis," *Artif. Intell. Rev.*, vol. 56, no. Suppl. 1, pp. 1479–1512, 2023.

[3] Z. Sun et al., "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 8992–8999.

[4] D. Lahat et al., "Challenges in multimodal data fusion," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*. Piscataway, NJ, USA: IEEE Press, 2014, pp. 101–105.

[5] D. Lahat et al., "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[6] S. Qiu et al., "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Inf. Fusion*, vol. 80, pp. 241–265, 2022.

[7] D. R. Beddiar et al., "Vision-based human activity recognition: A survey," *Multimedia Tools Appl.*, vol. 79, pp. 30509–30555, 2020.

[8] H. Alemdar et al., "Daily life behaviour monitoring for health assessment using machine learning: Bridging the gap between domains," *Pers. Ubiquitous Comput.*, vol. 19, no. 2, pp. 303–315, 2015.

[9] L. Babangida et al., "Internet of things (IoT) based activity recognition strategies in smart homes: A review," *IEEE Sensors J.*, no. 9, pp. 8327–8336, May 2022.

[10] P. Augustyniak and G. Slusarczyk, "Graph-based representation of behavior in detection and prediction of daily living activities," *Comput. Biol. Med.*, vol. 95, pp. 261–270, 2018.

[11] G. L. Alexander et al., "Passive sensor technology interface to assess elder activity in independent living," *Nursing Res.*, vol. 60, no. 5, pp. 318–325, 2011.

[12] E. Castillejo et al., "Modeling users, context and devices for ambient assisted living environments," *Sensors*, vol. 14, no. 3, pp. 5354–5391, 2014.

[13] D. Arifoglu et al., "Detecting indicators of cognitive impairment via graph convolutional networks," *Eng. Appl. Artif. Intell.*, vol. 89, 2020, Art. no. 103401.

[14] A. M. Seelye et al., "Naturalistic assessment of everyday activities and prompting technologies in mild cognitive impairment," *J. Int. Neuropsychol. Soc.*, vol. 19, no. 4, pp. 442–452, 2013.

[15] T. Steels et al., "Badminton activity recognition using accelerometer data," *Sensors*, vol. 20, no. 17, p. 4685, 2020.

[16] D. Roggen et al., "Recognition of crowd behavior from mobile sensors with pattern analysis and graph clustering methods," 2011, *arXiv:1109.1664*.

[17] P. Yang et al., "Activity graph based convolutional neural network for human activity recognition using acceleration and gyroscope data," *IEEE Trans. Ind. Inform.*, vol. 18, no. 10, pp. 6619–6630, 2022.

[18] F. Lin et al., "Adaptive multi-modal fusion framework for activity monitoring of people with mobility disability," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, 2022.

[19] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observation Geoinformation*, vol. 112, 2022, Art. no. 102926.

[20] J. Shi et al., "Multimodal graph learning with framelet-based stochastic configuration networks for emotion recognition in conversation," *Inf. Sci.*, vol. 686, 2025, Art. no. 121393.

[21] M. Li, S. Zhou et al., "Educross: Dual adversarial bipartite hypergraph learning for cross-modal retrieval in multimodal educational slides," *Inf. Fusion*, vol. 109, 2024, Art. no. 102428.

[22] M. Li et al., "Multimodal graph learning based on 3d haar semi-tight framelet for student engagement prediction," *Inf. Fusion*, vol. 105, 2024, Art. no. 102224.

[23] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[24] D. Hong et al., "Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation," *Innov. Geosci.*, vol. 2, no. 1, 2024, Art. no. 100055.

[25] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[26] D. Hong et al., "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE*

Trans. Geosci. Remote Sens., vol. 61, pp. 1–12, 2023, Art. no. 5527812.

[27] J. Zhang et al., "Zoom transformer for skeleton-based group activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8646–8657, 2022.

[28] Z. Tu et al., "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 1819–1831, 2023.

[29] X. Lang et al., "Knowledge augmented relation inference for group activity recognition," 2023, *arXiv:abs/2302.14350*.

[30] Y. Liu et al., "Skeleton-based human action recognition via large-kernel attention graph convolutional network," *IEEE Trans. Visualization Comput. Graph.*, vol. 29, no. 5, pp. 2575–2585, 2023.

[31] M. S. Raj et al., "Leveraging spatio-temporal features using graph neural networks for human activity recognition," *Pattern Recognit.*, vol. 150, 2024, Art. no. 110301.

[32] L. M. Dang et al., "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, 2020, Art. no. 107561.

[33] P. K. Singh et al., "Progress of human action recognition research in the last ten years: A comprehensive survey," *Arch. Comput. Methods Eng.*, pp. 1–41, 2021.

[34] K. Chen et al., "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–40, 2021.

[35] F. Gu et al., "A survey on deep learning for human activity recognition," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 8, pp. 1–34, 2021.

[36] T. Ahmad et al., "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 128–145, 2021.

[37] L. Feng et al., "A comparative review of graph convolutional networks for human skeleton-based action recognition," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 1–31, 2022.

[38] Z. Sun et al., "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, 2022.

[39] J. Li et al., "Overview of indoor scene recognition and representation methods based on multimodal knowledge graphs," *Appl. Intell.*, vol. 54, no. 1, pp. 899–923, 2024.

[40] P. Lukowicz et al., "Recording a complex, multi modal activity data set for context recognition," in *Proc. 23th Int. Conf. Archit. Comput. Syst.*, VDE, 2010, pp. 1–6.

[41] C. Bettini et al., "A survey of context modelling and reasoning techniques," *Pervasive Mobile Comput.*, vol. 6, no. 2, pp. 161–180, 2010.

[42] S. Ranasinghe et al., "A review on applications of activity recognition systems with regard to performance and evaluation," *Int. J. Distrib. Sensor Networks*, vol. 12, no. 8, 2016.

[43] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4, no. 4.

[44] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[45] M. S. Mahmud et al., "A survey of data partitioning and sampling methods to support big data analysis," *Big Data Mining Anal.*, vol. 3, no. 2, pp. 85–101, 2020.

[46] S. Scheurer et al., "Subject-dependent and-independent human activity recognition with person-specific and-independent models," in *Proc. 6th Int. Workshop Sensor-Based Activity Recognit. Interaction*, 2019, pp. 1–7.

[47] H. Guo et al., "Logistic discrimination based on g-mean and f-measure for imbalanced problem," *J. Intell. Fuzzy Syst.*, vol. 31, no. 3, pp. 1155–1166, 2016.

[48] A. Albahri et al., "Based on the multi-assessment model: Towards a new context of combining the artificial neural network and structural equation modelling: A review," *Chaos, Solitons Fractals*, vol. 153, 2021, Art. no. 111445.

[49] S. Zolfaghari et al., "Sensor-based locomotion data mining for supporting the diagnosis of neurodegenerative disorders: A survey," *ACM Comput. Surveys*, vol. 56, no. 1, pp. 1–36, 2023.

[50] Z. Li et al., "Sensorllm: Aligning large language models with motion sensors for human activity recognition," 2024, *arXiv:2410.10624*.

[51] O. Napoli et al., "A benchmark for domain adaptation and generalization in smartphone-based human activity recognition," *Scientific Data*, vol. 11, no. 1, 2024, Art. no. 1192.

[52] N. Sikder and A.-A. Nahid, "Ku-har: An open dataset for heterogeneous human activity recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 46–54, 2021.

[53] N. A. Choudhury et al., "Physique-based human activity recognition using ensemble learning and smartphone sensors," *IEEE Sensors J.*, vol. 21, no. 15, pp. 16852–16860, 2021.

[54] D. Garcia-Gonzalez et al., "A public domain dataset for real-life human activity recognition using smartphone sensors," *Sensors*, vol. 20, no. 8, p. 2200, 2020.

[55] D. Anguita et al., "A public domain dataset for human activity recognition using smartphones," in *Proc. 21st Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.*, 2013.

[56] J.-L. Reyes-Ortiz et al., "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, Jan. 2016.

[57] J. R. Kwapisz et al., "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newslett.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

[58] Q. Kong et al., "Mmact: A large-scale dataset for cross modal human action understanding," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2019, pp. 8658–8667.

[59] D. J. Cook et al., "Casas: A smart home in a box," *Computer*, vol. 46, no. 7, pp. 62–69, 2012.

[60] D. Roggen et al., "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Networked Sens. Syst. (INSS)*, Jun. 2010, pp. 233–240.

[61] C. Chen et al., "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015, pp. 168–172.

[62] M. Zhang and A. A. Sawchuk, "Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 1036–1043.

[63] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput.* Piscataway, NJ, USA: IEEE Press, 2012.

[64] O. Banos et al., "mhealthdroid: A novel framework for agile development of mobile health applications," in *Proc. Int. Workshop Ambient Assisted Living*. Cham, Switzerland: Springer, 2014, pp. 91–98.

[65] Y. Zhu et al., "DiamondNet: A neural-network-based heterogeneous sensor attentive fusion for human activity recognition," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 11, pp. 15321–15331, Nov. 2024.

[66] R. Fallahzadeh et al., "Personalized activity recognition using partially available target data," *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 374–383, 2023.

[67] P. Alinia et al., "Model-agnostic structural transfer learning for cross-domain autonomous activity recognition," *Sensors*, vol. 23, no. 6337, 2023.

[68] A. Sarkar et al., "Grafehty: Graph neural network using federated learning for human activity recognition," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2021, pp. 1124–1129.

[69] J. Hao et al., "Complex behavioral pattern mining in non-intrusive sensor-based smart homes using an intelligent activity inference engine," *J. Reliable Intell. Environ.*, vol. 3, pp. 99–116, 2017.

[70] L. Chen et al., "A multi-graph convolutional network based wearable human activity recognition method using multi-sensors," *Appl. Intell.*, vol. 53, p. 28169–28185, 2023.

[71] T. L. van Kasteren et al., "Human activity recognition from wireless sensor network data: Benchmark and software," in *Activity Recognit. Pervasive Intell. Environ.*, Springer, 2011, pp. 165–186.

[72] V. Ghasemi and A. A. Pouyan, "Activity recognition in smart homes using absolute temporal information in dynamic graphical models," in *Proc. 10th Asian Control Conf. (ASCC)*, Kota Kinabalu, Malaysia, 2015, pp. 1–6, doi: 10.1109/ASCC.2015.7244787.

[73] M. R. Keyvanpour and S. Zolfaghari, "Augmented feature-state sensors in human activity recognition," in *Proc. 2017 9th Int. Conf. Inf. Knowl. Technol. (IKT)*. Piscataway, NJ, USA: IEEE Press, 2017, pp. 71–75.

[74] N. Twomey et al., "Unsupervised learning of sensor topologies for improving activity recognition in smart environments," Neurocomputing, vol. 234, 2016, doi: 10.1016/j.neucom.2016.12.049.

[75] W. Yu et al., "An accurate indoor map matching algorithm based on activity detection and crowdsourced wi-fi," *Sci. China Technol. Sci.*, vol. 62, no. 9, pp. 1492–1501, 2019.

[76] L. P. Prieto et al., "Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data," *J. Comput. Assisted Learn.*, vol. 34, no. 1, pp. 1–10, 2018.

[77] S. S. Akter and L. B. Holder, "Activity recognition using graphical features," in *Proc. 13th Int. Conf. Mach. Learn. Appl.* Piscataway, NJ, USA: IEEE Press, 2014, pp. 165–170.
[78] S. S. Akter et al., "Activity recognition using graphical features from smart phone sensor," in *Int. Conf. Internet Things*. Springer, 2018, pp. 45–55.
[79] S. Akter and L. Holder, "Improving iot predictions through the identification of graphical features," *Sensors*, vol. 19, no. 15, p. 3250, 2019.
[80] C. Li et al., "Automatic extraction of behavioral patterns for elderly mobility and daily routine analysis," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 9, no. 5, 2018.
[81] M. Wang et al., "Neural network aided factor graph optimization for collaborative pedestrian navigation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 303–313, 2024.
[82] S. Al-Janabi and A. H. Salman, "Sensitive integration of multilevel optimization model in human activity recognition for smartphone and smartwatch applications," *Big Data Mining Anal.*, vol. 4, no. 2, pp. 124–138, 2021.
[83] L. Hu et al., "Activity recognition via correlation coefficients based graph with nodes updated by multi-aggregator approach," *Biomed. Signal Process. Control*, vol. 79, 2023, Art. no. 104255.
[84] A. Wang et al., "Hierhar: Sensor-based data-driven hierarchical human activity recognition," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3353–3364, 2021.
[85] A. Ataya et al., "Improving activity recognition using temporal coherence," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*. Piscataway, NJ, USA: IEEE Press, 2013, pp. 4215–4218.
[86] T. Vieira et al., "Online human moves recognition through discriminative key poses and speed-aware action graphs," *Mach. Vision Appl.*, vol. 28, pp. 185–200, 2017.
[87] Y. Huang et al., "Knowledge-driven egocentric multimodal activity recognition," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 16, no. 4, 2020.
[88] J. Z. Pan, "Resource description framework," in *Handbook Ontologies*. New York: Springer, 2009, pp. 71–90.
[89] J. Pérez et al., "Semantics and complexity of sparql," *ACM Trans. Database Syst. (TODS)*, vol. 34, no. 3, pp. 1–45, 2009.
[90] G. Antoniou and F. v. Harmelen, "Web ontology language: Owl," *Handbook on Ontologies*, pp. 91–110, 2009.
[91] D. Riboni and C. Bettini, "Owl 2 modeling and reasoning with complex human activities," *Pervasive Mobile Comput.*, vol. 7, no. 3, pp. 379–395, 2011.
[92] X. Wang et al., "Stge: Sensor topology and graph embedding learning with heterogeneous smart environment," in *Proc. IEEE 27th Int. Conf. Parallel Distrib. Syst. (ICPADS)*. Piscataway, NJ, USA: IEEE Press, 2021, pp. 699–706.
[93] B. Steenwinckel et al., "Talk: Tracking activities by linking knowledge," *Eng. Appl. Artif. Intell.*, vol. 122, 2023, Art. no. 106076.
[94] Y. Liu et al., "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 1–12, 2021.
[95] U. Bermejo et al., "Embedding-based real-time change point detection with application to activity segmentation in smart home time series data," *J. Smart Technol.*, 2021.
[96] M. Li et al., "Guest editorial: deep neural networks for graphs: theory, models, algorithms, and applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 4, pp. 4367–4372, 2024.
[97] W. Huang et al., "Shallow convolutional neural networks for human activity recognition using wearable sensors," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021.
[98] K. Ouyang and Z. Pan, "Multi-model weighted voting method based on convolutional neural network for human activity recognition," *Multimedia Tools Appl.*, 2023.
[99] Z. Wang et al., "Body rfid skeleton-based human activity recognition using graph convolution neural network," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 7301–7317, Jun. 2024.
[100] Q.-Y. Yao et al., "Human activity recognition using 2-d lidar and deep learning technology," *IEEE Sensors Lett.*, vol. 7, no. 10, 2023, Art. no. 5503204.
[101] R. Huan et al., "Two-domain joint attention mechanism based on sensor data for group activity recognition," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023, Art no. 2507315.
[102] J. Cao et al., "Sensor-based human activity recognition using graph lstm and multi-task classification model," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3s, 2022.
[103] G. Lee and J. Kim, "Mtgea: A multimodal two-stream gnn framework for efficient point cloud and skeleton data alignment," *Sensors*, vol. 23, no. 5, p. 2787, 2023.
[104] P. Gong et al., "Mmpoint-GNN: Graph neural network with dynamic edges for human activity recognition through a millimeter-wave radar," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*. Piscataway, NJ, USA: IEEE Press, 2021.
[105] G. Lee and J. Kim, "Improving human activity recognition for sparse radar point clouds: A graph neural network model with pre-trained 3d human-joint coordinates," *Appl. Sciences*, vol. 12, no. 2168, 2022.
[106] F. X. Chuanxin Zhao, Long Wang, and S. Chen, "Rfid-based human action recognition through spatiotemporal graph convolutional neural network," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 19898–19907, 2023.
[107] Y. Zhang et al., "Csi-based human activity recognition with graph few-shot learning," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4139–4150, 2022.
[108] A. Mohamed et al., "Har-GCNN: Deep graph CNNS for human activity recognition from highly unlabeled mobile sensor data," in *Proc. 2022 IEEE Int. Conf. Pervasive Comput. Commun. Workshops Other Affiliated Events*. Piscataway, NJ, USA: IEEE Press, 2022, pp. 335–340.
[109] Y. Yan et al., "Deep transfer learning with graph neural network for sensor-based human activity recognition," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 10, pp. 4389–4402, 2022.
[110] J. Wu and Q. Liu, "A novel spatio-temporal network of multi-channel cnn and gcn for human activity recognition based on ban," *Neural Process. Lett.*, vol. 55, p. 11489–11507, 2023.
[111] R. Mondal et al., "A new framework for smartphone sensor-based human activity recognition using graph neural network," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11461–11467, 2021.
[112] A. Vijayvargiya et al., "Pc-GNN: Pearson correlation-based graph neural network for recognition of human lower limb activity using semg signal," *IEEE Trans. Human-Mach. Syst.*, vol. 53, no. 6, pp. 945–953, 2023.
[113] J. Ye et al., "A graph-attention-based method for single-resident daily activity recognition in smart homes," *Sensors*, vol. 23, no. 1626, 2023.
[114] W. Meng et al., "Graphar: A lightweight human activity recognition model by exploring the sub-carrier correlations," *IEEE Trans. Wireless Commun.*, 2023.
[115] Y. Wang et al., "MhaGNN: A novel framework for wearable sensor-based human activity recognition combining multi-head attention and graph neural networks," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023, Art no. 2514314.
[116] G. Liu and J. Wang, "Dendrite net: A white-box module for classification, regression, and system identification," *IEEE Trans. Cybern.*, pp. 1–14, 2021.
[117] H. Wu et al., "A novel pedal musculoskeletal response based on differential spatio-temporal LSTM for human activity recognition," *Knowledge-Based Syst.*, vol. 261, no. C, 2022, doi: 10.1016/j.knosys.2022.110187.
[118] W. Zheng et al., "Meta-learning meets the internet of things: Graph prototypical models for sensor-based human activity recognition," *Inf. Fusion*, vol. 80, pp. 1–22, 2022.
[119] M. M. Islam and T. Iqbal, "Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 1729–1736, 2021.
[120] Y. Dong et al., "Graph-structure-based multigranular belief fusion for human activity recognition," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 10, pp. 13589–13603, Oct. 2024.
[121] A. Benmansour et al., "Multioccupant activity recognition in pervasive smart home environments," *ACM Comput. Surv.*, vol. 48, no. 3, pp. 34: 1–34:36, 2016.
[122] R. Chen and Y. Tong, "A two-stage method for solving multi-resident activity recognition in smart environments," *Entropy*, vol. 16, no. 4, pp. 2184–2203, 2014.
[123] D. Riboni and F. Murru, "Unsupervised recognition of multi-resident activities in smart-homes," *IEEE Access*, vol. 8, pp. 201985–201994, 2020.
[124] T. Gu et al., "epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.* Piscataway, NJ, USA: IEEE Press, 2009, pp. 1–9.

[125] R. Helaoui et al., "Recognizing interleaved and concurrent activities using qualitative and quantitative temporal relationships," *Pervasive Mobile Comput.*, vol. 7, no. 6, pp. 660–670, 2011.
[126] K. Thapa et al., "A deep machine learning method for concurrent and interleaved human activity recognition," *Sensors*, vol. 20, no. 20, p. 5770, 2020.
[127] J. Achiam et al., "Gpt-4 technical report," 2023, *arXiv:2303.08774*.
[128] J. Ye et al., "Situation identification techniques in pervasive computing: A review," *Pervasive Mobile Comput.*, vol. 8, no. 1, pp. 36–66, 2012.
[129] E. Damiani et al., "A graph-based meta-model for heterogeneous data management," *Knowl. Inf. Syst.*, vol. 61, pp. 107–136, 2019.
[130] F. Jie et al., "Hao unity: A graph-based system for unifying heterogeneous data," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 4725–4729.

Saeedeh Javadi is working toward the M.Sc. degree in data science and engineering with the Polytechnic University of Turin, Turin, Italy.

Her research interests include knowledge graphs, graph neural networks, human activity recognition, and the application of machine learning techniques in various domains, including healthcare, and smart home environments.

Daniele Riboni received the Ph.D. degree in computer science from the University of Milan, in 2007. He was a Postdoctoral Fellow and an Assistant Professor with the University of Milan. Currently, he is an Associate Professor in computer science with the University of Cagliari, Cagliari, Italy, since 2015. His research interests include pervasive healthcare, activity recognition, and knowledge management in pervasive and mobile computing. He served as a TPC Chair and a TPC Vice Chair for different conferences and workshops in the field, including IEEE PerCom and the International Conference on Intelligent Environments (IE). His contributions appear in major conferences and journals.

Luigi Borzì (Member, IEEE) received the B.Sc. and M.Sc. degrees in biomedical engineering and the Ph.D. degree in computer and control engineering from Polytechnic University of Turin, Turin, Italy, in 2015, 2018, and 2023, respectively.

He is an Assistant Professor with the Department of Control and Computer Engineering, Politecnico di Torino and works on medical applications of artificial intelligence and wearable technologies. His research interests include machine learning, wearable technology, digital health, data analytics, wireless body sensor networks, and medical IoT.

Samaneh Zolfaghari received the Ph.D. degree in computer science from the University of Cagliari, Cagliari, Italy, in 2023.

Currently, she is a Postdoctoral Researcher with Mälardalens University, Västerås, Sweden and works on activity recognition and fall risk assessment by utilizing various sensor technologies. Her research interests include machine learning, ambient-assisted living, mobile and pervasive systems, and human factors in pervasive computing applications.