IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 November 2023, accepted 11 January 2024, date of publication 19 January 2024, date of current version 31 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3356352

RESEARCH ARTICLE

Enhancing Kitchen Activity Recognition: A Benchmark Study of the Rostock KTA Dataset

SAMANEH ZOLFAGHARI[1], **TEODOR STOEV**[2], **AND KRISTINA YORDANOVA**[2]

[1]School of Innovation, Design, and Engineering, Division of Intelligent Future Technologies, Mälardalen University, 722 20 Västerås, Sweden
[2]Institute of Data Science, University of Greifswald, 17489 Greifswald, Germany

Corresponding author: Samaneh Zolfaghari (samaneh.zolfaghari@mdu.se)

This work was supported in part by the BehavE Project through the German Research Foundation under Grant YO 226/3-1, and in part by Mälardalen University. The work of Kristina Yordanova was supported by the Defense Advanced Research Projects Agency within the Framework of the Perceptually-Enabled Task Guidance (PTG) Autonomous Multimodal Ingestion for Goal Oriented Support (AMIGOS) Project HR001122C0009.

ABSTRACT With the global population aging, the demand for technologies facilitating independent living, especially for those with cognitive impairments, is increasing. This paper addresses this need by conducting a comprehensive evaluation of the Rostock Kitchen Task Assessment dataset, a pivotal resource in kitchen task activity recognition. Our study begins with an in-depth introduction, emphasizing the increasing prevalence of neurodegenerative disorders and the crucial role of assistive technologies. Our contributions encompass a systematic literature review, design and implementation of a working prototype of our envisioned system, refinement of the Rostock Kitchen Task Assessment dataset, creation of a semantically annotated dataset, extraction of statistical features, comparative analysis, and rigorous model performance assessment. The core of our work is the thorough evaluation and benchmarking of different activity recognition approaches using the aforementioned Rostock Kitchen Task Assessment dataset. Our experimental results demonstrate that despite encountering an imbalance problem in the dataset, the fusion of the Hidden Markov Model and Random Forest leads to superior results, achieving a weighted-averaged F_1-score of 74.10% for all available activities and 81.40% for the most common actions in the Rostock Kitchen Task Assessment dataset. Moreover, through systematic analysis, we identify strengths and suggest potential refinements, thereby advancing the field of kitchen activity recognition. This offers valuable insights for researchers and practitioners in assistive and remote care technologies.

INDEX TERMS Pervasive healthcare, neurodegenerative disorders, multistage activity, kitchen task assessment, action detection, activity recognition.

I. INTRODUCTION

Aging brings with it a growing concern for neurodegenerative conditions, notably Alzheimer's disease (AD) [1]. Projections from the World Health Organization (WHO) indicate that by 2030, an estimated 65.7 million individuals will be primarily afflicted with Alzheimer's disease (AD) [2]. Amongst these, individuals with cognitive impairments, particularly the elderly, face the dual challenges of declining the level of independence and safety issues [3].

As a response to this demographic shift, remote care applications have gained attention. People with dementia (PwD) and cognitive impairments frequently encounter challenges while performing routine Activities of Daily Living (ADLs), such as cooking, which can be complex and mundane. Assistive Technology Devices (ATDs) have assumed a key role in providing support to PwD and cognitive impairments [4]. These devices serve as beacons, guiding navigation and delivering timely medication reminders, while

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

providing expert guidance through the intricate process of culinary tasks. Moreover, these technologies act as virtual sentinels, remaining observant against potentially dangerous behaviors by reminding individuals to switch off cooking appliances or even taking control remotely. Contemporary innovations, such as applications on smartphones and computers, advocate for a user-driven, step-by-step prompting approach [5]. This method is a cornerstone in assisting multi-step tasks like cooking or other vital daily activities for the elderly, and underscores a critical aspect of assistive technologies: discerning precisely when and where intervention is warranted, all tailored to the specific circumstances and profile of the user [4], [6].

Effective care is built on a nuanced comprehension of human behavior within a given environment [7]. This understanding is often facilitated through a combination of sensors, audio inputs, and visual cues. Insights gained from these sources enable precise inferences about a person's actions, all framed within the contextual information of their surroundings [8], [9]. However, modeling, which hinges on annotated data, poses a significant challenge: it requires precise and time-intensive data collection and labeling.

Remote Care Technologies (RCTs) are emerging as technologies within the domain of assistive technologies for individuals suffering from cognitive impairments. These technologies initiate real-time monitoring within smart homes, providing invaluable insights for surveillance of ADLs and providing support to patients and caregivers [10]. Collaborative tools, ranging from unassuming paper calendars to sophisticated user-driven digital prompts, alleviate cognitive burdens and are especially efficacious for tasks like cooking among the elderly [11], [12].

Individuals with Traumatic Brain Injury (TBI) often exhibit abnormal behavior during multi-stage activities, which can be categorized into three main types: challenges related to experience, following steps, and cognitive/emotional factors. Among daily tasks, independent kitchen activities have been found to significantly impact overall well-being, health, social roles, self-esteem, and emotional balance [11], [13], [14]. Therefore, the demand for recognizing multi-step activities, particularly in cooking scenarios, has increased due to their pivotal role in nurturing independent living, especially for those suffering from cognitive impairments. However, the term of Kitchen Task Assessment (KTA), interwoven with the domain of remote care, presents a challenge: the need for a robust, versatile object and Activity Recognition (AR) system. Such a system must discerningly discriminate between diverse cooking items and activities while detecting symptoms and tracing the progression of neurodegenerative disorders in real-world conditions [7], [15]. The KTA, a context of diverse objects shared across multiple activities, has been significantly reinforced by several approaches and datasets, investigated by the KTA dataset for object usage detection, AR, and error measurements due to cognitive impairments by Yordanova et al. [16], [17].

TABLE 1. Description of kitchen task evaluable skills [13].

Evaluable skills	Description
Initiation	Evaluation of the subject's ability to begin the tasks
Organization	Evaluation of the subject's ability to gather the necessary tools and use them appropriately while performing tasks
Inclusion of all steps	Evaluation of the subject's ability to perform all the major steps alone without assistance
Sequencing	Evaluation of the subject's ability to perform tasks in a functional sequence
Judgment and Safety	Evaluation is based on how the person manages the use of tools that may cause injuries such as stove or hot pot.
Completion	Evaluation of the subject's ability to complete the tasks

II. PROBLEM STATEMENT

As previously noted, among ADLs and multi-step activities, kitchen tasks play a crucial role in people's cognitive abilities essential for independent living. Research has consistently demonstrated that performing these tasks autonomously significantly contributes to an individual's overall well-being, success in social roles, self-esteem, and control of emotions [11], [13], [14].

Indeed, the KTA, which occupational therapists developed, provides a functional measurement evaluating cognitive processes that affect task performance [13]. Moreover, it can be performed either in a clinic or in the person's home in a short period of time [18]. In general, the aim of the KTA is assessing the skills of initiation, organization, the inclusion of all steps, sequencing, safety and judgment, and completion while the task is being performed [13], [16], [19]. These skills are described in Table 1.

Considering the described kitchen task evaluable skills and observation of abnormal behavior, KTA can achieve the following goals [18]:

- Evaluate the cognitive processes that affect task performance and record the level of cognitive support necessary for successful task completion.
- Allow the clinician to observe and translate the person's performance into strategies the caregiver may use to manage the cognitively impaired person in other ADLs and instrumental tasks.
- Generate a score to measure changes in performance over time (either progression or improvement).

The initial step in addressing these challenges lies in the system's proficiency in AR. It is important to note that AR within the kitchen environment presents a particularly demanding task, as it encompasses the assessment of initiation, organization, and completion skills, adding an extra layer of complexity to the evaluation process. Indeed, a significant challenge arises during the process of data acquisition and labeling for creating robust AR models.

Moreover, refining annotations for detailed datasets like the KTA dataset [16], [17] presents a critical task.

Therefore, in this study, we start by investigating different existing methods in a systematic review. We then focus on a specific cooking situation, using the very detailed Rostock KTA dataset. We create a prototype of a system and improve the available semantic annotation. Finally, we compare different classification methods based on the improved annotations.

III. CONTRIBUTIONS

The contributions of the paper are as follows:

- **Systematic Literature Review**: Our work is grounded in a rigorous systematic literature review, which involved the relevant studies centered around AR in kitchen environments.
- **Enhancement of the Rostock KTA Dataset Quality**: We extend and correct the existing semantic annotations and evaluate their quality by calculating the inter-rater reliability between two annotators. This step ensures the accuracy and reliability of the annotated data.
- **Development of the Prototype System**: We design and implement a working prototype of our envisioned system.
- **Statistical Feature Extraction and Comparative Analysis**: We conducted an extensive analysis by extracting various statistical features and comparing them with different categories of features and sensors. This detailed examination included an in-depth investigation into the impact of various sensor types on the recognition process.
- **Model Performance Evaluation**: We carried out extensive experiments with the Rostock KTA dataset gathered in a smart kitchen test-bed with 12 participants to evaluate the system performance considering several state-of-the-art techniques such as Support Vector Machines (SVM), k-Nearest Neighbours (kNN), Decision Trees (DT), and Bidirectional Long short-term memory (BiLSTM), as well as fusion of Hidden Markov Models (HMM) with Random Forest (RF).
- **Publication of the Annotated Dataset**: The annotated dataset, along with its detailed methodology, will be made publicly available on GitHub[1] upon the publication of this manuscript. This resource is intended to facilitate further research and applications in the domain of AR.

The remaining sections of the paper are structured as follows. In Section IV we present a thorough literature review of existing approaches to AR in kitchen settings. We analyzed several features of the systems and we found: their ability to recognize fine- vs. coarse-grained actions, their public availability for benchmark purposes, the sensor technology they make use of, and their applicability in the detection of erroneous behaviour.

Section V elaborates on the methodology utilized to establish the benchmark for the KTA dataset. This includes comprehensive insights into sensor technologies, data collection, and annotation procedures. Additionally, it addresses crucial aspects related to the consideration of normal behavior and the simulation of erroneous behavior. These measures are employed to evaluate errors in kitchen task proficiency and track the progression of cognitive decline symptoms. In Section VI, we present the experimental setup, results, and analyses. It provides an analysis of the effectiveness of our proposed annotation approach and feature extraction methodology. Section VIII summarizes our findings, delves into their implications, and future research directions.

IV. SYSTEMATIC LITERATURE REVIEW

For our study we conducted a comprehensive systematic literature review, mirroring the approach outlined in [20], considering relevant AR studies within a kitchen environment. Our primary emphasis was on approaches developed or evaluated using openly accessible datasets, with a reliance on sensor data from ambient, wearable, cameras, objects and even radar and audio sources. Furthermore, these selected approaches exhibit proficiency in managing activities, whether fine-grained or coarse-grained, and a subset of them demonstrates effectiveness in addressing erroneous behaviors.

This literature review illuminates crucial facets regarding the utilization of the Rostock KTA dataset in this research. Notably, the KTA dataset stands out as a rare artifact in kitchen AR, encompassing instances related to cognitive impairment and erroneous behavior. Furthermore, it serves as a notable benchmark for sensor technology integration when compared to other datasets. Drawing on insights from related literature, approaches can be skillfully applied to yield novel insights when employed with the KTA dataset. This dataset occupies a distinctive niche, boasting features like public accessibility, diverse sensor observation types, and expertise in handling fine-grained actions.

A. SEARCH AND SELECTION STRATEGY

For our literature search, we utilized esteemed publication databases including IEEE Xplore, PubMed, Scopus, and Web of Science (WoS) considering academic studies published in peer-reviewed journals and in proceedings of international conferences in English. These databases were chosen for their relevance to computer science publications and their user-friendly features, allowing for specific term searches within titles, abstracts, and keywords. While each database may have slight variations in query syntax, it doesn't alter the underlying meaning or intent of the query.

The methodology we applied for our literature study comprised three distinct phases for extracting and finalizing pertinent papers, including *Search Phase and Paper Extraction*, *Pre-Selection Phase*, and *Full Text Analysis and Selection Phase*. We explain each of these phases below.

[1] https://github.com/DataScienceLab-HGW/Benchmark_KTA_Rostock

```
TITLE-ABSTRACT-KEYWORDS ( ( "kitchen" OR "cook" OR "food
    preparation" OR "meal-taking" ) AND ( "kitchen task"
    OR "kitchen activity" OR "cooking" OR "activities of
    daily living" OR "ADL" OR "daily life" OR "Activity-
    aware" OR "Situated Actions" OR "Situation awareness"
    OR "daily routine" ) AND ( "IOT" 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 "non-invasive" OR "
    intelligent assistive" OR "assistive device" OR "
    device free" OR "smartphone" OR "human computer
    interaction" OR "hci" OR "mobile health" OR "
    healthcare" OR "ambient assisted living" OR "aal" )
    AND ( "detection" OR "analysis" OR "classification"
    OR "reconstruction" OR "monitoring" OR "assessment"
    OR "recognition" ) ) AND ( LIMIT-TO ( DOCTYPE , "cp"
    ) OR LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO (
    DOCTYPE , "re" ) ) AND ( LIMIT-TO ( LANGUAGE , "
    English" ) )
```

FIGURE 1. Our search query.

1) SEARCH PHASE AND PAPER EXTRACTION

This phase encompassed a thorough delineation of terms, keywords, and logical formulas critical for the precise identification of relevant articles within our interest domain. These elements were instrumental in shaping our search query and its subsequent execution across the publication databases. Indeed, the formulation of the query (refer to Figure 1) was the result of a comprehensive brainstorming session, integrating multiple terms relevant to our topic of interest, coupled with insights derived from our extensive experience with scientific publications in this domain.

The entire query is broken down into five sections linked by the "AND" operator. The initial section encompasses various terms associated with kitchen environments. Given our focus on recognizing kitchen activities, the second section targets publications centered around cooking, kitchen tasks, daily living activities, and situated actions. In the third segment, we narrow down to papers specifically addressing sensor devices and pervasive computing technologies. The fourth section filters for papers concerning diverse monitoring techniques, while the final segment exclusively retrieves papers published in conference proceedings or scientific journals in English. We executed the formulated search query and stored the paper's link along with its metadata (including authors, publication year, journal, link, doi, etc.). This phase yielded an initial pool of 691 papers.

2) PRE-SELECTION PHASE

In this phase we removed all duplicate papers, keeping 548 papers. Then each paper's title and abstract were carefully examined to make an initial assessment of its relevance. In case of disagreement, the paper was discussed between all of this articles's authors to reach a consensus which was subsequently refined to 154 papers. The majority of papers were excluded as they either did not utilize sensor devices or pervasive computing technologies, focused on activities unrelated to kitchen settings, or concentrated on monitoring aspects such as indoor air quality, smoke alarm systems, identifying water usage patterns, detecting objects, or supervising cooking assistants in the form of robots, text, or mobile applications.

3) FULL TEXT ANALYSIS AND SELECTION PHASE

Papers which we considered potentially relevant underwent precise reading and the information needed for our study has been extracted from them. Further refinement during this phase by considering our inclusion criteria led us to select 54 research papers.

The outcomes of our literature review are summarized in Table 2.

B. FINDINGS FROM THE SYSTEMATIC LITERATURE REVIEW

In this section, we systematically analyze and evaluate our literature findings, considering several features which play a pivotal role in kitchen AR. First, we discuss the *public availability* of the datasets used in the works we found. Ensuring that a dataset is freely available ensures reproducibility and enables benchmarking. The second feature we discuss is the *sensor technology*. For us it is important to find out if single-modality or multi-modality approaches are predominantly utilized. The particular types of sensor technology is also of interest, since it might provide information about advantage or limitations. The third important aspect is the *action granularity*. We focus on delineating between coarse-grained and fine-grained actions. Here, we examine how activities are defined and recognized in the studies, offering an in-depth understanding of the level of detail considered in the recognition process.

Lastly, we turn our attention to how the studies account for anomalous or erroneous behavior. This subsection addresses whether the research performs AR for abnormal situations or behaviors, underlining the critical importance of robust recognition systems in kitchen AR.

1) PUBLIC DATASET AVAILABILITY

Conducting comprehensive experiments in naturalistic settings is vital for evaluating tools and methods effectively. However, this domain presents challenges in establishing experimental test-beds and involving a diverse participant pool, encompassing both normal and erroneous behavior. Additionally, the involvement of sensitive information often restricts data accessibility. As a result, there is a scarcity of publicly available, large-scale datasets spanning extended time periods and including a substantial number of individuals.

Out of the papers examined (as presented in Table 2), approximately 35% (19 papers) discussed the utilization of publicly available datasets for kitchen AR. Notable examples include Carnegie Mellon University's Multi-Modal Activity (CMU-MMAC) Database [31], Epic Kitchens [36], [37], Cooking activity dataset with macro and micro activities [74],

TABLE 2. **The results of our literature study. Abbreviations as folow: CS: contact sensor, THS: thermal sensor, ACC: accelerometer, LRS: laser range sensors, IRC: infrared cam, CCDC: CCD camera, IRSMD: infrared sensors for movement detection, IMU: Inertial measurement unit, TEMP: temperature, CO2: CO2 presence, MAG - magnetic switch, PIRMS: passive infrared movement sensor, LUM: luminiscence sensor, HUM: humidity, USD: ultra sound sensor to measure distance, ELCONS: electrical consumption sensor, WCONS water consumption sensor, WUS: water usage sensor, MS: movement sensor, RS: reed sensor, WPS: water pressure sensor, PS: pressure sensor, LLS: light level sensor, NLS: noise levels sensor, DLS: dust levels sensor, LIM: limit switch,RFID Radio-Frequency Identification, RFIDT: RFID-tags, DC: RGB camera with depth information, PWD: power load sensor, VS: vision sensor, C: RGB camera, SS: smoke sensor, SWACC: smart-watch accelerometer, MACCG: mobile device accelerometer and gyroscope, MACCGMAG: mobile accelerometer, gyroscope, magnetometer, PT: paper-tags, SC: stereoscopic camera, FS: float sensor.**

Ref.	Publicly Available Dataset	Sensors Technology						Actions Granularity		Erroneous Behaviour
		Ambient sensors	Body sensors	Object sensors	Radar sensors	Cameras inkl. depth cameras	Audio sensors	Fine-grained	Coarse-grained	
[21]				PT		SC		✓	✓	✓
[22]					✓			✓	✓	
[23]		CS, THS	ACC	CS					✓	
[24]		LRS, MAG, LIM		RFIDT		IRC, CCDC		✓	✓	
[25]		IRSMD, CO2, TEMP		MAG					✓	
[26]		THS		CS				✓	✓	✓
[27]		MS		CS					✓	
[28]				ACC				✓	✓	
[29]				ACC				✓		
[30]		PIRMS, LUM, HUM, TEMP, ACC							✓	
[31]	✓		IMU					✓	✓	
[32]						DC			✓	
[33]		USD, TEMP, HUM							✓	✓
[14]	✓	ELCONS, WCONS, MS, TEMP				DC		✓	✓	
[34]		ELCONS, PD, MS, PWD	ACC			VS			✓	
[35]		RS, PIR							✓	
[36]	✓					C	✓	✓		
[37]	✓					C	✓	✓		
[38]	✓					C		✓		
[39]		TEMP, CO2, HUM, MS, SS				C			✓	✓
[40]						DC		✓		
[41]			IMU					✓		
[42]						DC		✓		
[43]			MYOA			DC		✓		
[44]	✓	PIRM, PS, RS, FS, CS						✓	✓	
[45]	✓							✓		
[46]				OM		C		✓		
[47]			MACCGMAG, MGPS			C			✓	
[48]						DC			✓	
[49]	✓		SWACC					✓	✓	
[50]	✓					C				
[51]	✓	TEMP, HUM, LLS, NLS, DLS, MS, WUS, EUS						✓	✓	
[52]	✓					DC		✓		
[53]	✓	MAG, RFID	MACCG			C			✓	
[54]	✓					C			✓	
[55]	✓						✓		✓	
[56]						C		✓	✓	
[57]	✓					DC		✓		
[58]						DC		✓		
[59]		IRSMD							✓	
[60]						DC		✓	✓	
[61]						C		✓		
[62]	✓					DC		✓		
[63]		MAG	RFID gloves	RFID tags		C		✓		
[64]		EM, RS, PS, TEMP								
[65]		PS		RFIDT, LLS,TEMP, ACC				✓	✓	
[66]	✓			ACC				✓		
[67]				PT		SC		✓		✓
[68]		RS, PIR, PS		RFID, ACC		C			✓	✓
[69]	✓	RS, RFID readers		RFID				✓		
[70]	✓								✓	
[71]			(wrist RFID reader)					✓	✓	
[72]							✓		✓	
[73]		WPS							✓	

and SPHERE [51], among others. Surprisingly, the majority of researchers relied on their own data and observations, without making them accessible to others. This lack of data sharing hampers the reproducibility of experiments, rendering benchmarking unfeasible in such cases.

It's worth noting that in some publicly available datasets, annotations consist solely of a set of labels without an underlying structure. This leaves room for potentially annotating causally impossible sequences of events. The absence of structured annotations can pose significant challenges, particularly when validating more complex symbolic structures, as annotation labels fail to represent relations between activities [75]. To address this, the authors of [76] endeavored to provide semantic annotation for CMU-

MMAC [77], thus establishing ground truth annotation. The resulting annotation is publicly accessible for further research.[2] However, it's important to note that these structural annotations are confined to the examples within the dataset.

In light of these considerations, we posit that the KTA dataset [16], [17] stands as a crucial resource for future benchmarking endeavors.

2) SENSOR TECHNOLOGY

In our exploration of sensor technology, a critical aspect was discerning whether prior studies employed single-modality or multi-modality approaches. In single-modality approaches, a specific type of sensor technology is employed.

Experimental results have convincingly demonstrated the reliable recognition of kitchen activities through embedded sensors in kitchen items. Notably, distinguishing between seemingly similar activities, such as "making tea" and "making green tea," was facilitated by object sensors, which discerned the specific objects in use [75]. However, the scalability of object-based approaches poses a substantial challenge, given the multitude of objects that require discrimination and the complexity of annotating data for each object under realistic conditions [15]. Additionally, recognizing less distinct activities, like slicing, dicing, and scraping, which are inherently fewer in number, proves to be a formidable task [78].

Consequently, multi-modality approaches, which leverage multiple sensor technologies, have been employed to infer the performed activities. Integrating sensors such as cameras and wearable Inertial Measurement Units (IMUs) alongside object sensors enhances prediction and assessment robustness. Moreover, it enriches the quality of annotation, providing a dependable ground truth for human AR. This approach also incorporates contextual information into the annotation, enabling the automatic learning of object models from video or other sensor data and the application of common-sense knowledge to activities. This ensures scalability and facilitates reliable quantification of the recognition performance [15], [16].

In certain studies ([15], [16], [79], [80]) researchers not only embedded sensors in kitchen utensils and employed wearable IMUs but also utilized cameras positioned on the wrist, head, or in a fixed location to capture the surrounding space. Although the integration of cameras significantly enhances the accuracy of ADLs and goal recognition, it raises privacy concerns, as it may inadvertently capture aspects of a user's private life [81].

In summary, our review of existing studies revealed that the predominant approaches are based on the amalgamation of multiple sensor technologies. Notably, around 31% of the papers we surveyed exclusively utilized various types of cameras (RGB and/or depth cameras) for AR, highlighting the common practice of addressing kitchen AR scenarios by exclusively relying on visual features as a singular modality. It's noteworthy that only one work [23] presented an ADLs recognition approach incorporating all three types of sensors: ambient, body, and object. Within the comprehensive review of 54 papers, [55] and [72], exclusively utilized audio sensors.

3) ACTIONS GRANULARITY

In the realm of AR, a fundamental distinction is often drawn between coarse-grained and fine-grained actions. Coarse-grained actions encompass broad ADLs, such as walking, washing dishes, and cooking food. Conversely, fine-grained actions constitute atomic actions that are integral parts of larger activities, often involving the manipulation of objects. For instance, the coarse-grained action of "washing dishes" can be decomposed into finer actions like "taking a dish" and "turning on the tap".

As delineated in Table 2 of our literature review, researchers frequently explore both types of action granularity concurrently, with only 18 papers exclusively focusing on coarse-grained actions. This reflects the recognition that activities are often deducible by identifying sequences of atomic actions and the objects involved in their execution. For example, in a study by Yordanova et al. [51], the coarse-grained action of "preparing healthy food" was discerned through the observation of fine-grained actions like "drink", "clean", and "put". Another instance is the approach by Chatterjee et al. [31], where coarse-grained actions (referred to as macro-activities) were disaggregated into finer-grained actions (micro-activities). For instance, the action "take baking pan" was represented as the sequential series of finer-grained actions: "open bottom cupboard, take the pan, and finally close the cupboard".

Ultimately, the chosen level of granularity is a pivotal consideration, contingent upon the ultimate research goal. Various combinations of sensors, as depicted in Table 2, afford multiple avenues for realizing both types of AR.

4) ERRONEOUS BEHAVIOR

Our literature review underscored a limited exploration of erroneous behavior within the context of kitchen AR. To make KTA applicable for AR and remote care scenarios, it's imperative to account for anomalous behavior.

Only in six instances did researchers endeavor to pinpoint abnormal kitchen behavior. For example, Žari'c et al. [33] linked erroneous behavior to the use of a hotplate during cooking, focusing solely on activities during the cooking stage. While they modeled situations potentially endangering individuals during cooking, such as a plate being turned on without objects on it, they did not cover cases such as the completion of a cooking task, initiation, or organization (see Table 1). In the CMU-MMAC Database [77], instances of anomalous behavior during cooking encompassed sudden fires, smoke, house robbery, distractions while cooking, and ingredient omissions while preparing a salad. Another work by Garcia et al. [26] introduced a petri-net based model capable of detecting errors in recipe executions involving elderly individuals. However, the model in the

[2] http://purl.uni-rostock.de/rosdok/id00000163

paper specifically caters to precisely modeled sequences, with limited information on real-world application and dataset specifics provided.

Magherini et al. [67] presented a temporal logic approach for detecting deviations from the normal execution of kitchen tasks, primarily focusing on correct coffee preparation. Their models, expressed in temporal logic formulas, verified the proper execution of sequential actions, interleaved actions, incomplete actions, and repeated actions. However, the introduced erroneous scenarios were fairly constrained (e.g., user receives a phone call and forgets to complete the task). In a subsequent work, Magherini et al. [21] employed a similar approach to recognize incorrect coffee preparation and incorrect medication intake.

Biswas et al. [68] addressed the erroneous behavior of elderly patients with mild dementia by extending the planning approach introduced in [82]. They defined a correct plan as an ADLs and an erroneous plan as a sequence of activities deviating from an ADLs. While the researchers achieved promising results in recognizing three erroneous executions of a meal-time scenario, they did not explicitly detail the types of errors made. However, they referenced another paper [83] that described simulated errors, including initiation error, completion error, and realization error.

In summary, our investigation into errors associated with kitchen tasks revealed a limited number of research papers. Moreover, the identified errors differed from the typical errors observed in individuals with neurodegenerative diseases. It is also noteworthy that datasets containing records of these erroneous behaviors, as observed in the aforementioned studies, are not openly shared.

Therefore, the exceptional nature of the Rostock KTA dataset [16], which stands as one of the few publicly accessible and unique resources specifically tailored for capturing erroneous behaviors exhibited by patients with neurodegenerative conditions and suitable for training and evaluating algorithms. Notably, this dataset provides structured semantic-based annotations.

V. COMPREHENSIVE EVALUATION AND BENCHMARKING OF THE ROSTOCK KTA DATASET

The section conducts a comprehensive evaluation of the Rostock KTA dataset. It covers the dataset's infrastructure, object extraction, pre-processing, feature extraction, observation model, and experimental assessment. This evaluation provides essential insights into the dataset's capabilities and suitability for various applications in kitchen AR.

A. SYSTEM OVERVIEW

Figure 2 provides an overview of our prototype system. In this study, we operate under the assumption that Rostock KTA offers a multi-modal infrastructure with the capability to continuously monitor inhabitants from various perspectives while engaging in kitchen tasks, ranging from a coarse-grained assessment to a fine-grained level. A comprehensive explanation of the employed infrastructure and the entire dataset is detailed in subsection V-B. Our system exhibits the capability to extract positional information not only from localization infrastructures, but also by leveraging the inhabitant's interaction with sensor-equipped objects and appliances, and even through the analysis of collected video data.

As previously stated, this paper emphasizes re-annotation derived from collected videos, the processing of embedded and wearable sensor data, and fine-grained AR. This encompasses the utilization of both conventional learning models and advanced techniques like BiLSTM, a widely recognized sequential RNN model in the field of AR.

B. ROSTOCK KTA MULTI-MODAL INFRASTRUCTURE

The primary objective of the KTA problem is to ascertain an individual's capacity to independently perform kitchen tasks by evaluating the manner in which the task is executed and identifying any associated errors [16]. Tracking the progression of dementia presents certain challenges, as error rates in performing kitchen activities tend to escalate until the individual is no longer capable of completing them autonomously.

To address this, the Rostock KTA multi-modal infrastructure continuously collects data as participants engage in kitchen tasks, utilizing a variety of data sources, including Wearable Sensors (WS) and Embedded Object (EO) sensors. The video data, recorded at a sampling rate of $25Hz$, is captured by two types of cameras: one mounted on the chest to capture hand interactions with objects, and another handheld camera to record the person's entire body.

Sensor data from manipulated objects are acquired through DIANA-boards from Bosch Sensortec, operating at a sampling rate of $25Hz$. Each sensor incorporates an accelerometer (Acc), gyroscope (Gyr), and magnetometer (Mag). These object sensors are affixed to 37 objects using tape. Each time a sensor is triggered, the platform transmits a **raw sensor event** ($rse = \langle t_s, t_{sys}, addr, acc_{x,y,z}, gyr_{x,y,z}, mag_{x,y,z} \rangle$) to the system. Here, t_s denotes the sensor's timestamp, t_{sys} is the system's timestamp, $addr$ corresponds to the MAC address of the sensor within the system (retrieved from "sensors information"), and $acc_{x,y,z}$, $gyr_{x,y,z}$, and $mag_{x,y,z}$ represent the generated values of Acr, Gyr, and Mag data along the x, y, and z axes.

Furthermore, acceleration data from a full-body motion capture suit (XSens MVN-Biomch) with 17 sensors, operating at a sampling rate of $120Hz$, is collected during kitchen tasks.

In addition to these sensors, data from the electrocardiogram and electrodermal activity, recorded at varying sampling rates (ECG: $1024Hz$, EDA: $64Hz$, Acc: $64Hz$, Temp: $1Hz$, Barometric Pressure: $8Hz$), are also included [16], [17].

The Rostock KTA dataset comprises a total of 24 runs involving 12 participants. This encompasses 12 runs exhibiting normal behavior, while the remaining 12 runs include simulated erroneous behavior to evaluate errors in kitchen task proficiency, as detailed in Table 1, and to track the

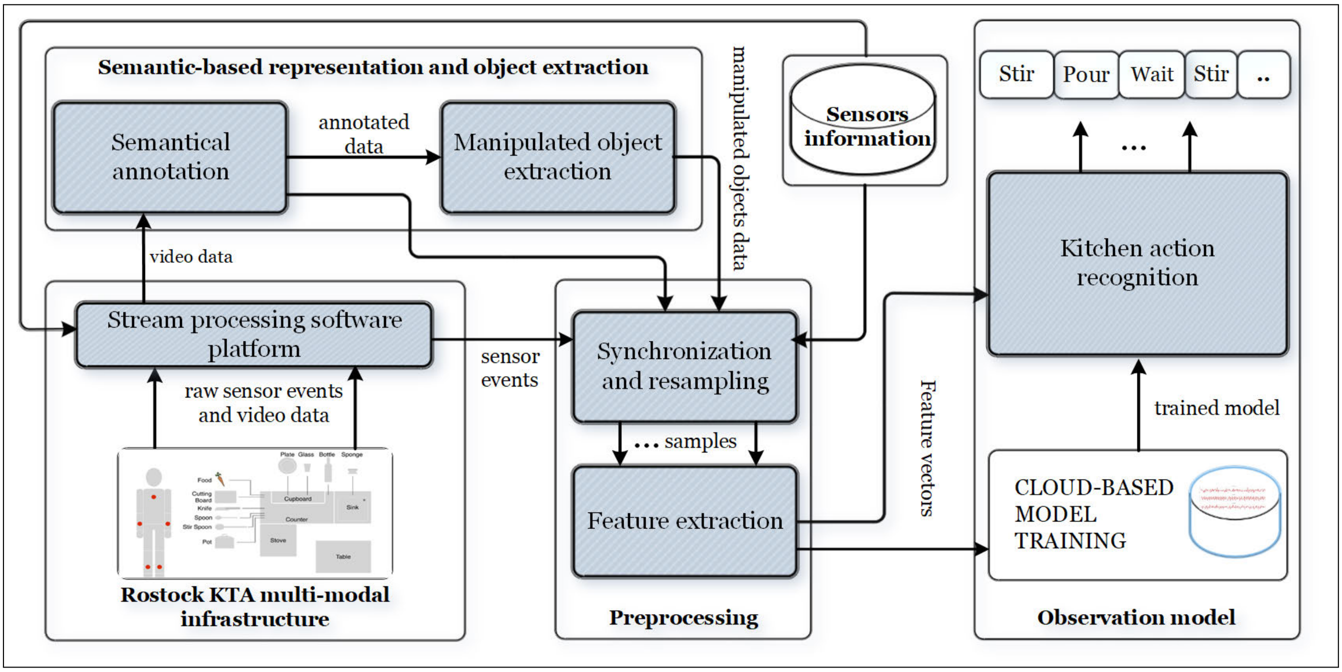


FIGURE 2. System overview.

progression of cognitive decline symptoms. For the purpose of this study, we exclusively considered subjects who executed the tasks without simulated errors. Furthermore, data for 3 subjects were excluded due to missing information and errors in actions and sensors recorded during kitchen tasks, a confirmation provided by both annotators upon review of the collected videos.

C. SEMANTIC-BASED REPRESENTATION AND OBJECT EXTRACTION

In smart homes, assistive systems must possess cognitive capabilities to understand dynamic situations in both temporal and spatial dimensions, compensating for potential cognitive limitations of inhabitants [84]. This necessitates data that is interpretable and processable by these systems. We advocate for the integration of semantic technologies, enriching sensor data with comprehensive metadata and precise meaning. This approach empowers automation and advanced processing, enabling assistive systems to make automated interpretations and informed decisions based on semantic situational data. This is particularly critical for achieving situation-aware assistance in kitchen ADLs [85], [86].

Hence, this module addresses both syntactic errors inherent to compilation and semantic errors related to validation [75]. Additionally, it determines how to formulate semantic relations in the context of kitchen AR.

Indeed, the creation of precise, high-quality ground truth is paramount for training intelligent systems in human behavior recognition and providing effective user support. It also plays a pivotal role in evaluating system performance [9]. In this application, contextual information is typically gathered through an array of sensors attached to objects. Each sensor monitors a specific aspect of a situation. Based on these observations, our context modeling and semantic annotation consider the intricate structure of kitchen tasks, encompassing information about actions, manipulated objects, and additional contextual details.

Expanding our system to encompass smart homes inhabited by individuals with chronic diseases, the integration of contextual insights pertaining to these conditions allows us to distinguish between normal and abnormal behaviors. This leads to a more profound understanding of the root causes behind errors in ADLs [85].

To accomplish this, we implement an offline annotation procedure on the video data captured during the kitchen tasks, allowing us to closely examine the participants' cooking activities. The proposed semantic annotation model, depicted in Figure 3, encompasses crucial situation and contextual details, including the specific "action" undertaken, the manipulated "object," the manipulating hand, as well as spatial and location information. Notably, certain annotations, such as "swap-milk-right-left" indicating the transition of the milk from the right hand to the left, or "walk-table-sink" signifying movement from the table to the sink, are notable examples of annotations utilized in this process.

The re-annotation process involved meticulously assigning causally accurate labels using the ELAN annotation tool [87]. This was carried out frame by frame, encompassing both instances of correct behavior and artificially induced erroneous behavior. Additionally, parallel actions were annotated, accounting for actions executed with the left hand, right

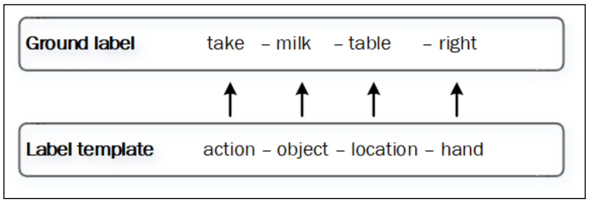

Ground label
take – milk – table – right
Label template
action – object – location – hand

FIGURE 3. Semantic annotation model with an example for sequential actions.

hand, or both hands during kitchen tasks. The format for annotating parallel actions followed the structure of "action-object-object-hand & action-object-object-hand".

Furthermore, particular attention was given to rectifying any discrepancies in identifying the manipulating hands (i.e., left, right, or both) in relation to the objects.

Subsequently, the semantic-based annotation was augmented with additional details, including the manipulated objects (termed as "objects"), the hand involved, and the corresponding start and end times of object usage. To validate the causal correctness of our annotations, we employed a plan-validation procedure akin to the methodology outlined in [76]. Since our annotations essentially represent a sequence of consecutive actions involving manipulated objects, each experimental run can be viewed as a plan, transitioning from an initial state to a specific goal state. This validation process encompassed the following steps:

1) We manually created a comprehensive domain definition using Planning Domain Definition Language (PDDL) to model the kitchen task.
2) Individual problem files were developed for each experimental run, defining the initial and goal states.
3) Our annotations were transformed into executable plans.
4) We employed a PDDL plan validator to ensure the plans were indeed executable.

It's important to note that the annotations underwent a rigorous evaluation process to gauge the level of agreement between two independent annotators tasked with annotating data from 12 individuals, encompassing both normal and simulated erroneous runs. This process resulted in a commendable average Cohen's kappa of 0.9694. The average interrater reliability per subject is further detailed in Table 3.

Furthermore, Table 4 presents a comprehensive list of annotated objects, their respective locations, and corresponding action classes. Additionally, Table 5 provides an estimated distribution of activity classes among participants based on the annotations of two independent annotators. It is worth noting that actions such as "loosen", "drop", "hold", and "release" were omitted from consideration due to their infrequent occurrence and limited representation across participants.

As evident, there exists an imbalance in the distribution of samples among subjects and action classes. Specifically, Table 5 highlights that the predominant class, namely 'stir', accounts for over 53% of occurrences within each subject's dataset.

TABLE 3. Cohen's kappa score per subject.

Subjects	Cohen's kappa
s001-kta	0.94386
s002-kta	0.95786
s003-kta	0.98850
s004-kta	0.98157
s005-kta	0.97658
s006-kta	0.98768
s007-kta	0.96455
s008-kta	0.94328
s009-kta	0.96563
s010-kta	0.96931
s011-kta	0.96091
s012-kta	0.97200

TABLE 4. Objects and actions used in the annotation and their types.

Type	Including
Placeable objects	saucepan, hotplate, measuring_cup, paper_cups, tool_jar, cutting_board, saucepan_lid, clipboard
Other objects	milk_lid, milk_seal,pudding_seal, hotplate_dial, wooden_spoon, rubber_scraper, milk, pudding_mix, plastic_spoon
Actions	stir, take, put, hold, shake, walk, wait, scrape, pour, turn, turn_on, turn_off, screw, unscrew, tear, swap, open, close, release, drop, loosen
Location	table, sink, fridge, wastebasket, floor, middle_of_room
Hand	left, right, both

It's worth noting that in addition to the collected data and semantic-based representation, we have a valuable information source referred to as "sensors information" (depicted in Figure 2). This encompasses details about sensor offsets in relation to annotations, as well as the mapping between sensor addresses and their respective objects. This data will play a crucial role in the synchronization, resampling, and amalgamation of diverse information collected from multimodal sensors. It will also facilitate the identification of the MAC address of the manipulated sensor during kitchen tasks.

D. PRE-PROCESSING AND FEATURE EXTRACTION

To preprocess the data, we initiated by removing values corresponding to sensor downtime, which were labeled as 'unknown'. Following this, for synchronization and resampling, given the clock drift in the wireless sensor nodes, we employed the cubic spline method. This enabled us to rectify the effective output data rate of each node by resampling.

For sampling and feature extraction, we utilized a sliding window approach. Features were computed using the R

TABLE 5. Distribution of actions per subject.

Subjects	Actions																
	wait	take	swap	walk	open	close	put	turn	unscrew	tear	pour	screw	shake	stir	turn_on	turn_off	scrape
s001-kta	9.4%	6.3%	2.9%	3.2%	0.1%	0.3%	4.5%	0.4%	0.1%	1.0%	10.1%	0.4%	0.2%	58.6%	0.4%	0.4%	1.7%
s002-kta	20.2%	8.3%	3.3%	4.1%	0.2%	0.2%	8.1%	0.5%	0.1%	0.8%	7.1%	0.2%	0.6%	44.3%	0.4%	0.4%	1.5%
s003-kta	23.9%	3.8%	1.2%	2.9%	0.1%	0.1%	3.8%			0.6%	12.4%	0.3%	4.3%	42.0%	0.5%	0.3%	3.9%
s004-kta	15.6%	4.0%	2.1%	3.7%	0.2%	0.2%	3.8%	0.2%	0.2%	0.6%	7.6%	0.4%	1.8%	56.1%	0.6%	0.2%	2.7%
s005-kta	37.8%	4.2%	1.3%	3.4%	0.2%	0.2%	4.1%	0.3%	0.2%	1.2%	11.4%	0.6%	0.2%	31.2%	0.6%	0.5%	2.9%
s006-kta	25.9%	5.9%	3.7%	2.5%	0.2%	0.3%	6.1%	0.1%	0.4%	1.4%	8.2%	0.5%	1.0%	40.0%	0.4%	0.2%	3.4%
s007-kta	6.5%	6.8%	2.5%	3.4%	0.2%	0.2%	6.7%	0.0%	0.2%	0.8%	8.9%	0.3%	0.2%	57.0%	0.6%	0.6%	5.2%
s008-kta	7.9%	4.0%	1.7%	3.5%	0.1%	0.2%	4.4%	0.4%	0.3%	0.6%	6.5%	0.2%	0.5%	65.4%	0.3%	0.3%	3.7%
s009-kta	7.7%	4.4%	1.3%	3.0%	0.2%	0.1%	3.5%	0.1%	0.3%	0.8%	6.9%	0.5%	0.5%	67.5%	0.4%	0.1%	2.8%
s010-kta	13.5%	3.2%	1.8%	3.7%	0.2%	0.2%	4.4%		0.3%	1.7%	7.6%	0.9%	0.7%	57.6%	0.5%	0.4%	3.5%
s011-kta	15.2%	3.5%	0.9%	1.9%	0.2%	0.2%	4.5%	0.4%	0.3%	1.1%	5.1%	0.3%	1.6%	59.7%	0.4%	0.3%	4.4%
s012-kta	13.8%	3.1%	1.3%	2.0%	0.2%	0.1%	3.3%	0.1%	0.3%	0.7%	5.7%	0.2%	0.6%	65.1%	0.9%	0.9%	1.8%

library ‘roll apply’,[3] employing a window size of 50 samples with a step of 1. Any missing values were subsequently filled with zeros.

In order to determine the correlation between wearable and embedded object sensors, we utilized an adapted similarity measure for vectors (Equation 1), akin to the Pearson correlation coefficient and cosine similarity.

$$ccf\_[hand - sensor] - [object]_{w_i} = \frac{(x - mean(x)).(y - mean(y))}{Var(|x - y|)} \tag{1}$$

where $ccf\_[hand - sensor] - [object]_{w_i}$ denotes the correlation between wearable sensors and the specific hand used, as well as the manipulated object for the current window of samples. The variables x and y pertain to the data from different sensor axes.

Additionally, for the Acc, Gyr, and Mag, we computed seven additional features: mean, variance, skewness, kurtosis, and standard deviation, as well as the magnitude value.

The resulting features underwent all the preprocessing steps and were subsequently integrated into the dataset, with each column appropriately labeled to denote the extracted feature. As a final step, we downsampled the data to a rate of 100Hz. This was a pragmatic choice, considering the somewhat inconsistent sampling frequency of the sensors.

E. OBSERVATION MODEL

The observation model establishes the connection between low-level sensor-based observations and the actual performed activities.

In this context, the objective of the cloud-based model training module is to instruct the designated models–whether they be conventional, modern, or foundational sequential learners–to categorize each sensory observation into distinct kitchen-based actions.

To achieve this, we adopted a collaborative approach. The module routinely receives a training set comprising feature vectors extracted from both WS and EO sensors during kitchen tasks. Each vector is tagged with the corresponding action executed by the participant (e.g., “wait” “take” “put” and so forth). These feature vectors are locally computed on the edge, where various instances of the designated models assimilate the provided training sets. The cloud-based module is responsible for training these instances of the designated model for action detection based on the extracted features. The action recognition model accesses the trained model of the same type from the cloud to classify the actions performed by the subjects. Indeed, for the sake of privacy, feature vectors are processed only by the trusted cloud module.

VI. EXPERIMENTAL EVALUATION

In this section, we present the results of our experimental evaluation conducted with the Rostock KTA dataset, gathered from an instrumented kitchen with the participation of 12 individuals. This assessment encompasses various feature categories and sensor types. In the following subsections, we detail our experimental setup, outline the obtained results, and provide a comparative analysis of the performance against state-of-the-art methods using the extracted features.

A. STATE–OF–THE–ART OBSERVATION MODELS

1) CONVENTIONAL NON-SEQUENTIAL MODELS

We evaluated nine conventional machine learning classifiers. They are presented as follows as well as their settings. It should be mentioned that they are mostly implemented using the “sklearn” python library and the settings are as follows.

- Naive Bayes (NB) [88]: which is based on the Gaussian function and prior probabilities are adjusted based on classes;
- SVM [89]: The regularization parameter is set to 1.0 and the kernel is ‘rbf’ one;
- RF [90]: The number of trees in the forest is set to 200, all the features are considered for the best split, and the maximum depth is set to 10;
- kNN [91]: with $k = 3$ as the number of neighbors;
- DT [92]: with the maximum depth 20;

[3] https://www.rdocumentation.org/packages/zoo/versions/1.8-11/topics/rollapply

- Multilayer perceptron (MLP) [93]: composed of 100 neurons in a hidden layer, ReLU activation function, Adam optimizer, and trained for 1000 epochs, with batch size equal to 200;
- Linear Discriminant Analysis (LDA) [94]: with Singular value decomposition which is suitable for data with a large number of features;
- Logistic Regression (LR) [95]: with inverse of regularization strength equal to 1.0.

2) CONVENTIONAL SEQUENTIAL MODELS

We opted to implement a HMM as a straightforward sequential baseline model. Following a similar approach to Albert et al. [96], we employed a classifier as the emission probability model for our HMM. The states within our HMM corresponded to the actions we sought to classify, augmented by additional start and end states. To derive the transition matrix, we computed the frequencies of action sequences from the training set across each fold in our cross-validation procedure.

In our experiments, we selected the best-performing classifier as the emission model. We subsequently conducted forward filtering to assess the probability of a state at each time step within the test sequence, taking into account the historical estimates of preceding states.

3) LONG SHORT TERM MEMORY BACKGROUND AND MODEL ARCHITECTURE

Recently, advanced variations of RNNs, such as Long Short Term Memory (LSTM) networks, have demonstrated their effectiveness in achieving state-of-the-art performance on demanding AR tasks [97], [98].

A distinguishing feature of LSTM networks lies in their cell memories, responsible for retaining states over varying time intervals, whether short or long. These memory cells facilitate the exchange of information between different layers of the LSTM network [97].

Among the RNN variants, the BiLSTM [99] stands out. It employs two parallel LSTM layers, operating through both forward and backward loops. By considering patterns from both past and future contexts, the BiLSTM excels in capturing temporal dependencies effectively. The architectural representation of this model is illustrated in Figure 4.

In the model architecture, the inputs are channeled into the Bidirectional layer, encompassing an LSTM layer equipped with ‘n’ LSTM cells for class prediction. The forward layer processes input data ‘X’ from left to right (as denoted by the green arrows), while the backward layer processes it from right to left (as indicated by the red arrows). The output prediction takes the form of a hidden state, which subsequently undergoes dropout regularization to effectively mitigate overfitting. This is followed by a dense layer employing a softmax activation function, resulting in a probability distribution denoted as $\{P(a_{m-1}), P(a_m), P(a_{m+1}), \ldots\}$ across all activities in ‘A’. The final output of the model is the activity ‘a’ from set ‘A’ with the highest probability.

Given that the number of cells (‘n’) and the learning rate serve as common hyper-parameters for all LSTM-based and neural network approaches, their selection is integrated into the validation process.

To fine-tune the hyper-parameters of the BiLSTM model, the training data is partitioned, allocating 10% for validation and reserving 90% for actual training. The Adam optimizer is employed for network training, minimizing the categorical cross-entropy loss function. This loss function, based on the maximum likelihood estimate approach, is tailored for single-label categorization [100]. Additionally, a dropout layer is incorporated in the final layer to substantially curb overfitting, with the fraction of dropped units empirically set at 0.5. The training dataset is subdivided into small batches of 32 samples, utilized in each epoch for error calculation and model coefficient updates. In this specified configuration, the number of epochs is established at 64.

B. EXPERIMENTAL SETUP

To ensure scalability, our system’s general architecture envisions the utilization of a cloud-based system for training both sequential and non-sequential models. However, due to the relatively modest size of our training set in these experiments, we conducted the training on a departmental server. This server was equipped with four NVIDIA Tesla P6 graphic boards, a single NVIDIA Pascal GP104 graphics processing unit, and 16 GB of GDDR5 memory.

Our semantic-based representation, object extraction, and all algorithms were developed in Python. For preprocessing and feature extraction, we employed an adapted version of the R script coding scheme provided by Yordanova et al. [16], customizing it to suit our specific use case. Additionally, we utilized the Python Keras neural network library[4] to implement the BiLSTM network.

All experiments employed the “Leave-One-Person-Out” (LOPO) cross-validation approach, which is particularly effective in preserving the sequential nature of the data. This methodology involves reserving the data of one participant for the test set, while the data from the remaining subjects are used for training and validation. This ensures that data from the same participant is never used for both training/validation and testing simultaneously.

For the purposes of this study, we focused exclusively on subjects who performed kitchen tasks without any simulated errors. Furthermore, we excluded three subjects due to missing information and errors in the actions and sensors during data collection, a verification that was corroborated by both annotators upon reviewing the collected videos. To address class imbalance, we initially selected a subset of the most common actions, specifically six actions: ‘pour’, ‘put’, ‘scrape’, ‘stir’, ‘take’, and ‘wait’.

[4] https://keras.io/

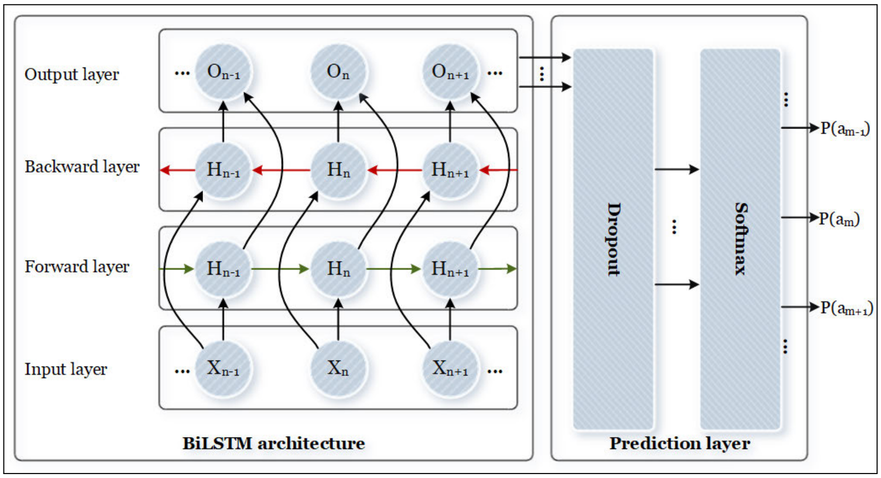


FIGURE 4. Model architecture.

In evaluating the overall performance of all models, given the pronounced class imbalances, we relied on the metrics of precision, recall, and F_1-score. We reported both weighted and macro-averaged values of these measures. The former is an average over all instances, while the latter is an average over the action classes.

Notably, the macro average F_1-score provides a reliable metric by assigning equal importance to different classes, irrespective of their size. Conversely, the accuracy metric is unsuitable for imbalanced problems like the one we addressed. Therefore, we excluded it from our evaluations. Depending on the severity of class imbalance, the accuracy value of the majority class could overshadow that of the minority classes.

In this context, the term ‘weighted’ denotes the calculation of metrics for each class, with the average weighted by support (i.e., the number of true instances for each class). This modification accommodates class imbalance and can yield an F_1-score that does not fall between precision and recall.

From an evaluation perspective, having two distinct sets of sensors affords three potential input features for the classifiers: WS only, EO sensors only, or a combination of both WS and EO sensors. It is of interest to compare the performance of selecting either one set of sensors or using both concurrently. Consequently, we conducted a significance assessment of the results by comparing the various combinations of sensor-based extracted features.

C. EXPERIMENTAL RESULTS

In order to assess the impact of the re-annotation process and the utilization of sensor-based extracted features and evaluation the developed system, we conducted experiments employing a variety of both traditional machine learning models and state-of-the-art sequential models, encompassing BiLSTM and HMM. The outcomes of these experiments are detailed in the subsequent subsections, presented in the form of tables and plots for clarity and comprehensive analysis.

1) CONVENTIONAL MODELS RESULTS

The results for all 16 actions, obtained using the most conventional machine learning models, are illustrated in Figures 5a and 5b. Although MLP performed reasonably well, RF demonstrated slightly superior results across various classification models. Particularly, the MLP achieved its highest scores in terms of macro average F_1-score with the Gyr-Mag (WS+EO) features, which were more than 5% lower than RF’s best result. In terms of weighted average F_1-score, MLP’s best performance with Acc-Gyr (WS+EO) features was approximately 3% lower than RF’s top-weighted result.

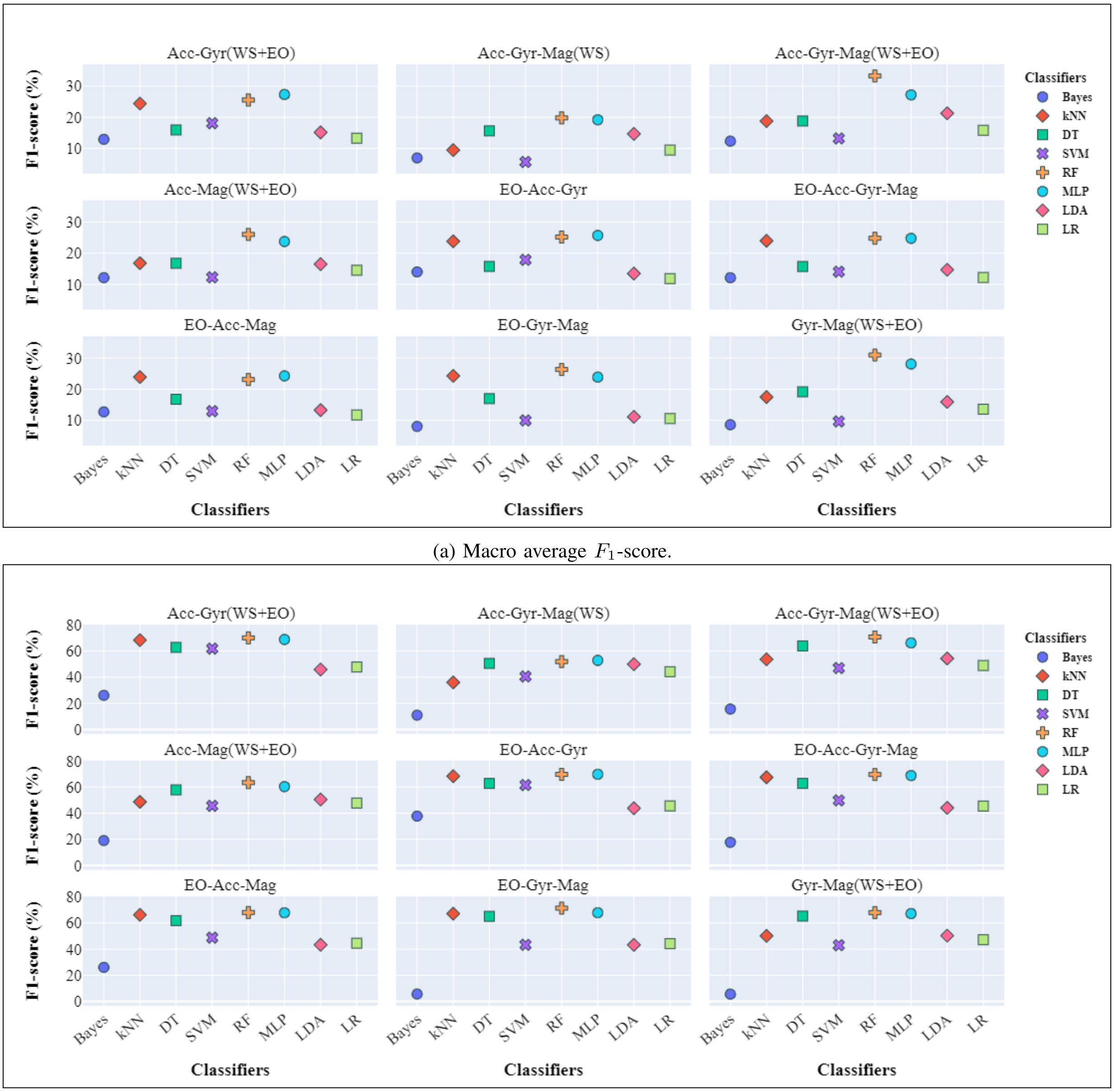
Analyzing the experimental results based on different feature categories, RF excelled with Acc-Gyr-Mag (WS+EO) features (33.25%) for macro average F_1-score, and with EO-Gyr-Mag features (71.33%) for weighted average F_1-score. Consequently, the comprehensive RF results, considering all 16 actions, are presented in Tables 6 and 7.

Figures 6a and 6b showcase the results for models focusing on the most common actions. As anticipated, there was a noticeable improvement–approximately 25% for macro average F_1-score and 8% for weighted average F_1-score. RF and MLP demonstrated superior performance, with RF emerging as the top performer. In this subset of experiments, EO-Gyr-Mag feature groups delivered the best results for both macro and weighted average F_1-score, outperforming other feature categories.

Moreover, MLP attained its best results using EO-Gyr-Mag features (which coincides with RF’s feature category), albeit approximately 5% lower for macro average F_1-score and 3% lower for weighted average F_1-score compared to RF. Refer to Tables 8 and 9 for detailed RF results. It is worth noting that among all experiments, the lowest results for all feature categories and actions were obtained using WS features.

2) HMM WITH RF EMISSIONS RESULTS

As previously discussed, we employ an HMM approach that leverages another classifier for the emission probabilities



(a) Macro average F_1-score.

(b) Weighted average F_1-score.

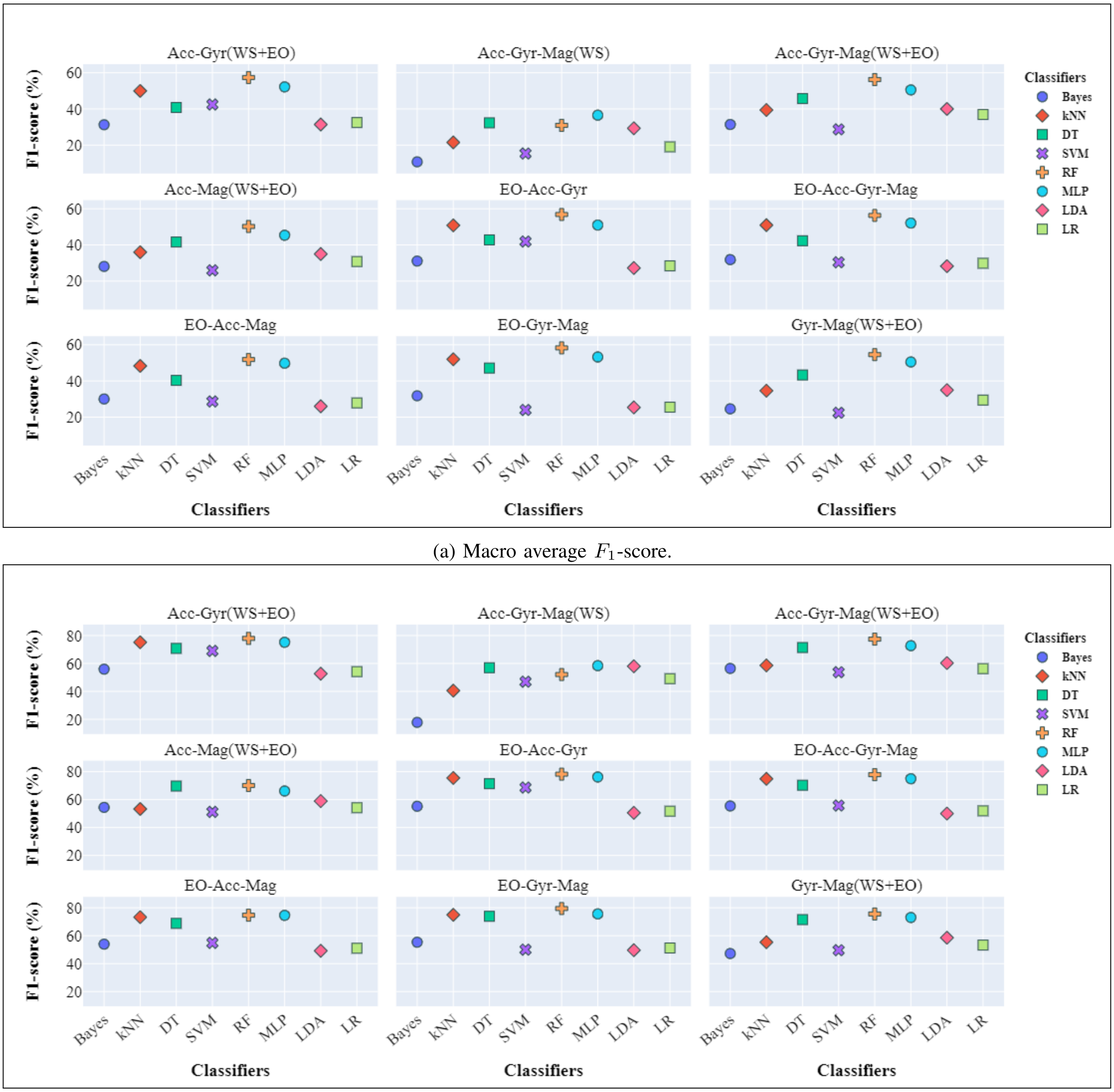
FIGURE 5. Action recognition F_1-score results utilizing conventional models and LOPO cross-validation considering all 16 actions.

associated with each action. Given the consistent performance of RF, which demonstrated the highest F$_1$-score in the majority of our experiments with conventional classifiers, we opted to integrate it with the HMM in our experimental setup.

The results of the HMM-RF classifier, encompassing all actions and the most prevalent actions, are presented in Figures 10, 11, 12, and 13.

In the context of predicting all actions, the HMM-RF exhibited a modest performance in terms of macro F$_1$-score. The highest achieved score was 24.89% with the Gyr-Mag(WS+EO) features. Interestingly, this result was surpassed by the RF classifier operating independently on the same task, attaining its peak macro F_1-score of 33.25% with the Acc-Gyr-Mag(WS+EO) features. However, in terms of weighted F_1-score, the HMM-RF displayed improved performance at 74.22%, in contrast to the 71.33% achieved when using RF alone to classify all actions.

For the prediction of the six most common actions in our dataset, the HMM-RF model achieved its highest macro F_1-score of 57.22% with the Gyr-Mag(WS+EO) features. In terms of weighted F_1-score, the optimal performance was observed with the EO-Gyr-Mag features, amounting to 81.22%. Notably, according to the weighted F_1-score, the HMM-RF classifier outperforms RF in this context, albeit with a marginal increase of approximately 1.76%.



(a) Macro average F_1-score.

(b) Weighted average F_1-score.

FIGURE 6. Action recognition F_1-score results utilizing conventional models and LOPO cross-validation considering 6 most common actions.

In summary, our findings indicate that the integration of HMM with RF leads to only marginal performance enhancements. This aligns with the results obtained by Albert et al. [96], who conducted similar experiments in a comparable experimental setting. One potential explanation lies in our dataset, where certain actions may occur in prolonged sequences of consecutive timesteps–such as ''stir'' which typically spans more than 20 seconds. This results in a substantial number of annotated frames with the same label, potentially obscuring information about preceding distinct actions. This suggests that future studies should explore the use of larger time windows. However, it's important to bear in mind that our dataset also includes actions occurring in very brief timeframes, such as ''turn'' and expanding the time window may pose challenges in detecting them.

3) BILSTM RESULTS

In prior studies [97], [98], it has been posited that the capacity of networks like BiLSTM to model long-term dependencies and learn non-linear feature representations contributes to superior modeling of ADLs patterns, particularly in datasets characterized by imbalances.

TABLE 6. RF action recognition macro average results regarding LOPO cross-validation considering all 16 actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	25.54	30.90	24.79	Macro
Acc-Mag(WS+EO)	26.03	30.50	25.64	
Gyr-Mag(WS+EO)	30.97	34.20	30.23	
EO-Acc-Gyr	25.19	30.71	24.31	
EO-Acc-Gyr-Mag	24.80	29.62	24.07	
EO-Acc-Mag	23.10	29.35	22.31	
EO-Gyr-Mag	26.33	30.09	25.41	
Acc-Gyr-Mag(WS)	19.80	22.01	19.95	
Acc-Gyr-Mag(WS+EO)	33.25	37.55	31.56	

TABLE 7. RF action recognition weighted average results regarding LOPO cross-validation considering all 16 actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	70.08	70.46	71.34	Weighted
Acc-Mag(WS+EO)	63.34	66.25	62.62	
Gyr-Mag(WS+EO)	67.96	70.03	67.05	
EO-Acc-Gyr	69.64	69.82	71.09	
EO-Acc-Gyr-Mag	69.53	69.77	71.10	
EO-Acc-Mag	68.01	68.33	69.62	
EO-Gyr-Mag	71.33	70.50	73.06	
Acc-Gyr-Mag(WS)	51.94	57.46	49.88	
Acc-Gyr-Mag(WS+EO)	70.76	71.85	70.91	

TABLE 8. RF action recognition macro average results regarding LOPO cross-validation considering 6 most common actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	57.35	57.04	59.31	Macro
Acc-Mag(WS+EO)	50.21	50.27	53.13	
Gyr-Mag(WS+EO)	54.54	55.93	53.98	
EO-Acc-Gyr	56.83	56.45	58.46	
EO-Acc-Gyr-Mag	56.31	56.00	57.88	
EO-Acc-Mag	51.85	51.36	54.28	
EO-Gyr-Mag	58.29	57.29	60.05	
Acc-Gyr-Mag(WS)	30.92	34.57	30.69	
Acc-Gyr-Mag(WS+EO)	56.24	56.41	57.51	

TABLE 9. RF action recognition weighted average results regarding LOPO cross-validation considering 6 most common actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	77.92	78.69	78.08	Weighted
Acc-Mag(WS+EO)	70.08	72.83	68.56	
Gyr-Mag(WS+EO)	75.58	77.52	74.28	
EO-Acc-Gyr	78.20	78.33	78.70	
EO-Acc-Gyr-Mag	77.84	78.10	78.25	
EO-Acc-Mag	74.67	75.32	75.14	
EO-Gyr-Mag	79.46	79.15	80.07	
Acc-Gyr-Mag(WS)	52.07	61.23	48.84	
Acc-Gyr-Mag(WS+EO)	77.39	78.10	77.15	

TABLE 10. HMM-RF action recognition macro average results regarding LOPO cross-validation considering all 16 actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	21.35	24.41	22.14	Macro
Acc-Mag(WS+EO)	20.44	24.12	21.35	
Gyr-Mag(WS+EO)	25.00	29.09	25.82	
EO-Acc-Gyr	20.40	23.24	21.52	
EO-Acc-Gyr-Mag	20.51	24.38	21.39	
EO-Acc-Mag	19.26	23.83	19.91	
EO-Gyr-Mag	21.67	25.47	22.15	
Acc-Gyr-Mag(WS)	16.87	20.91	17.57	
Acc-Gyr-Mag(WS+EO)	23.62	26.84	24.54	

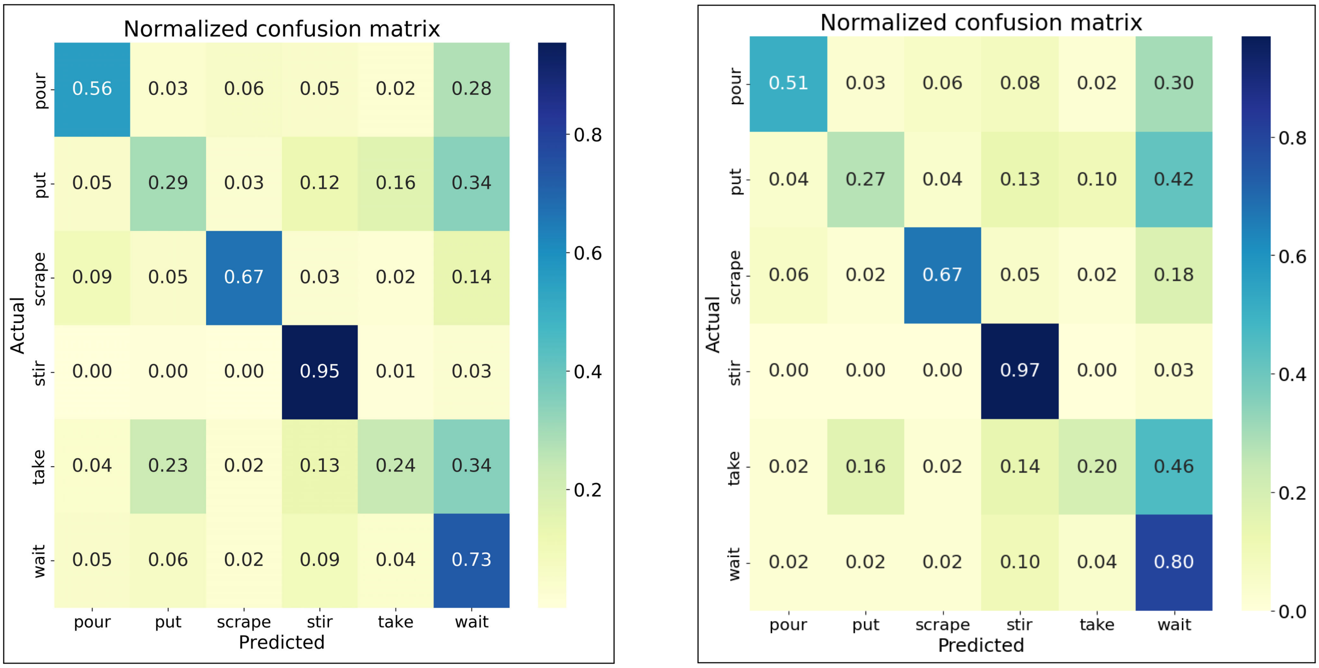
TABLE 11. HMM-RF action recognition weighted average results regarding LOPO cross-validation considering all 16 actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	69.68	72.14	72.49	Weighted
Acc-Mag(WS+EO)	67.92	70.75	71.82	
Gyr-Mag(WS+EO)	74.10	74.92	76.67	
EO-Acc-Gyr	69.90	71.64	72.69	
EO-Acc-Gyr-Mag	68.39	71.76	71.31	
EO-Acc-Mag	67.38	70.45	70.66	
EO-Gyr-Mag	72.68	72.24	76.45	
Acc-Gyr-Mag(WS)	56.43	59.90	59.46	
Acc-Gyr-Mag(WS+EO)	70.35	73.58	72.87	

The outcomes generated by our BiLSTM implementation for both classification tasks (with 6 and 16 actions) are presented in Tables 14, 15, 16, and 17.

In our experimental setup, the performance of the BiLSTM network closely parallels that of the RF model. Specifically, with regards to the F$_1$-macro score, the top-performing RF classifier (employing ACC-Gyr-Mag(WS+EO) features) surpasses its leading BiLSTM counterpart (using Gyr-Mag(WS+EO) features) by a marginal 1.0% when considering all 16 actions (33.25% vs. 32.25%, respectively). In terms of the weighted F_1-score for the same classification task, the highest-performing BiLSTM model achieves 72.36% (with Gyr-Mag(WS+EO) features), slightly surpassing the score of the best RF model (71.33% with EO-Gyr-Mag features) in this regard.

Furthermore, even in the classification task involving only 6 actions, RF demonstrates superior macro F_1-scores with varying combinations of features. However, based on the weighted F_1-score, the BiLSTM exhibits a slight edge in detecting these specific actions.



(a) The RF confusion matrix regarding EO-Gyr-Mag features and LOPO cross-validation considering 6 most common actions.

(b) The HMM-RF confusion matrix regarding EO-Gyr-Mag features and LOPO cross-validation considering 6 most common actions.

FIGURE 7. Examples of confusion matrices.

TABLE 12. HMM-RF action recognition macro average results regarding LOPO cross-validation considering 6 most common actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	55.93	63.67	56.56	**Macro**
Acc-Mag(WS+EO)	47.39	61.84	47.88	
Gyr-Mag(WS+EO)	57.20	65.85	57.31	
EO-Acc-Gyr	54.16	59.81	55.29	
EO-Acc-Gyr-Mag	54.82	62.64	55.71	
EO-Acc-Mag	49.42	61.16	50.06	
EO-Gyr-Mag	56.89	63.31	57.27	
Acc-Gyr-Mag(WS)	35.60	45.52	36.85	
Acc-Gyr-Mag(WS+EO)	55.79	64.90	56.06	

TABLE 13. HMM-RF action recognition weighted average results regarding LOPO cross-validation considering 6 most common actions.

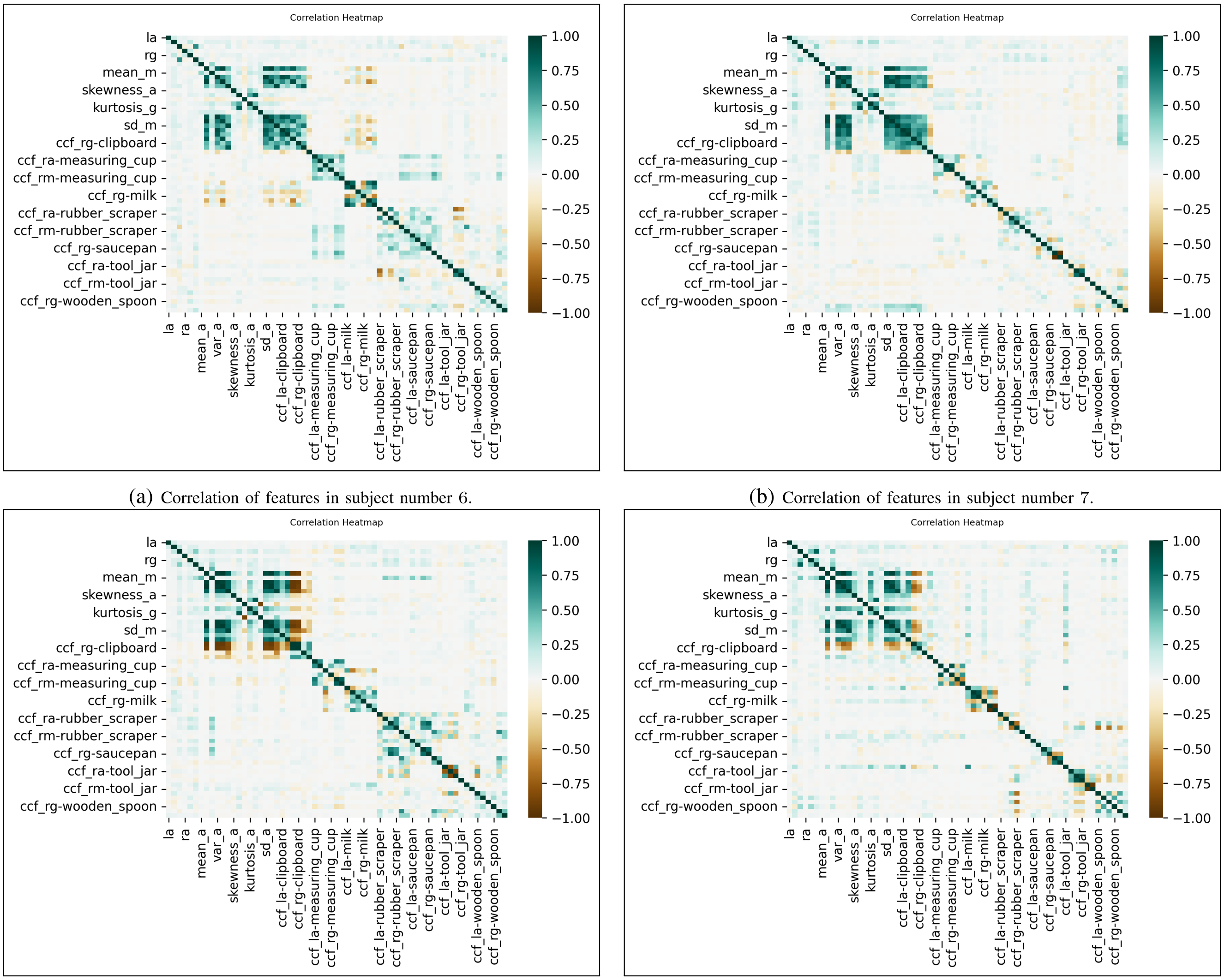
Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	77.86	81.16	78.78	**Weighted**
Acc-Mag(WS+EO)	74.16	80.93	75.57	
Gyr-Mag(WS+EO)	80.78	83.20	81.87	
EO-Acc-Gyr	77.16	80.85	78.19	
EO-Acc-Gyr-Mag	77.45	81.81	78.44	
EO-Acc-Mag	74.55	79.82	76.30	
EO-Gyr-Mag	81.40	82.14	83.37	
Acc-Gyr-Mag(WS)	60.93	65.72	62.74	
Acc-Gyr-Mag(WS+EO)	77.32	82.78	78.09	

The results from our experiments underscore that this type of network may still be influenced by class imbalances, emphasizing the critical role of the chosen validation method in such scenarios. Similar to the employed HMM, the network could potentially lose information about preceding actions due to the limited number of time frames per observation. Here, one potential solution could be to consider larger time windows. While this might result in a reduction of samples, it could also lead to shorter intervals between annotations for different actions (for example, reducing the number of repetitions from 100 to 10 before the next action occurs), which could have a positive impact on overall performance.

VII. DISCUSSION

In this study, we treated the Rostock KTA dataset as temporal data, transforming the kitchen task recognition problem into a time series classification challenge. Extensive experiments were conducted to assess various models, particularly those with sequential architectures capable of processing parallel sequences from different sensor axes, such as Acc and Gyr data. These models demonstrated their ability to extract features from observational sequences and map them to distinct kitchen activities.

Our results revealed that relying solely on WS features produced the least favorable outcomes across diverse feature categories and model types. To better understand this,



(a) Correlation of features in subject number 6.

(b) Correlation of features in subject number 7.

(c) Correlation of features in subject number 9.

(d) Correlation of features in subject number 12.

FIGURE 8. Examples of randomly selected subjects and representation of their feature correlations.

TABLE 14. BiLSTM action recognition macro average results regarding LOPO cross-validation considering all 16 actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	27.49	30.47	26.49	**Macro**
Acc-Mag(WS+EO)	28.21	30.89	27.39	
Gyr-Mag(WS+EO)	32.25	36.62	31.10	
EO-Acc-Gyr	26.02	28.00	25.31	
EO-Acc-Gyr-Mag	26.98	29.10	26.53	
EO-Acc-Mag	25.73	27.76	25.32	
EO-Gyr-Mag	25.57	28.82	24.70	
Acc-Gyr-Mag(WS)	15.10	18.88	15.78	
Acc-Gyr-Mag(WS+EO)	30.68	34.23	29.38	

TABLE 15. BiLSTM action recognition weighted average results regarding LOPO cross-validation considering all 16 actions.

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	69.97	70.60	70.76	**Weighted**
Acc-Mag(WS+EO)	69.58	70.4	69.79	
Gyr-Mag(WS+EO)	72.36	73.15	72.92	
EO-Acc-Gyr	69.83	69.81	70.84	
EO-Acc-Gyr-Mag	69.80	70.48	70.33	
EO-Acc-Mag	67.44	68.46	67.64	
EO-Gyr-Mag	68.34	69.28	68.73	
Acc-Gyr-Mag(WS)	51.88	55.63	51.97	
Acc-Gyr-Mag(WS+EO)	71.43	71.97	72.23	

we calculated the Pearson correlation coefficient [101], exposing strong correlations, particularly among WS features (illustrated in Figure 8), leading to collinearity that may adversely affect model performance.

TABLE 16. **BiLSTM action recognition macro average results regarding LOPO cross-validation considering 6 most common actions.**

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	53.73	56.18	52.84	**Macro**
Acc-Mag(WS+EO)	51.65	52.20	51.46	
Gyr-Mag(WS+EO)	55.40	57.84	54.40	
EO-Acc-Gyr	52.25	54.85	51.33	
EO-Acc-Gyr-Mag	54.58	56.30	53.93	
EO-Acc-Mag	50.49	51.58	50.38	
EO-Gyr-Mag	52.99	54.30	52.80	
Acc-Gyr-Mag(WS)	29.98	37.22	31.29	
Acc-Gyr-Mag(WS+EO)	55.91	57.77	55.15	

TABLE 17. **BiLSTM action recognition weighted average results regarding LOPO cross-validation considering 6 most common actions.**

Features	Measures (%)			
	F_1-score	Precision	Recall	
Acc-Gyr(WS+EO)	77.65	77.97	78.30	**Weighted**
Acc-Mag(WS+EO)	76.32	76.14	76.71	
Gyr-Mag(WS+EO)	78.69	79.02	79.23	
EO-Acc-Gyr	77.26	77.41	78.11	
EO-Acc-Gyr-Mag	78.84	78.83	79.49	
EO-Acc-Mag	75.55	76.02	75.75	
EO-Gyr-Mag	76.91	77.47	77.05	
Acc-Gyr-Mag(WS)	59.00	60.96	61.24	
Acc-Gyr-Mag(WS+EO)	79.34	79.36	79.97	

The confusion matrices depicted in Figure 7 generated when using EO-Gyr-Mag features with RF and HMM-RF clearly show that while these setups yield the highest F_1-score for the 6 most common actions, reliably recognizing certain actions, such as "put" and "take" still poses a challenge.

Imbalances in data distribution were identified as a contributing factor to moderate performance, with classifiers achieving significantly higher scores when trained on fewer actions, e.g. the highest weighted-averaged F_1-scores slightly above 80% with different feature combinations and macro F_1-scores around 57%. In comparison, our best models trained on all 16 actions achieved an average macro F_1-score of about 32% and a weighted F_1-score of about 74%.

To tackle this challenge, our future work will delve into hierarchical classification, distinguishing between coarse- and fine-grained actions. This approach aims to improve the detection of all executed actions, achieving higher macro-averaged F_1-scores and effectively handling imbalanced situations.

Another research direction involves refining sequential models, particularly the BiLSTM architecture, by integrating a symbolic model. This integration aims to leverage the strengths of deep learning while drawing contextual information from symbolic models, thereby enhancing model explainability and reasoning capabilities. The planned symbolic model employs a compact PDDL syntax and automatically generates a dynamic Bayesian network, facilitating probabilistic reasoning about actions, goals, contexts, and causes of behavior [14]. By developing hybrid machine learning models for AR, we anticipate overcoming challenges associated with sequential learning, imbalanced data, and enhancing behavior analysis in terms of actions, goals, and causes recognition.

VIII. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we tackled the challenge of discerning cooking actions executed by 12 participants within a simulated kitchen environment at the University of Rostock in Germany. Recognizing the limited availability of publicly accessible data depicting the behavior of individuals with dementia, the MMIS group at this university gathered data from actors simulating kitchen scenarios and associated activities, including dementia-induced errors in behavior. This dataset is now accessible to the public for extended use.

Our contribution involved a systematic examination of existing literature, coupled with the design and implementation of a functional prototype for our envisioned system, and extending and refining the Rostock KTA dataset which annotated semantically using the ELAN annotation tool [87]. Through precise frame-by-frame analysis of collected videos, we enriched annotations with details like hand usage (left, right, or both), additional actions during kitchen tasks, and typo corrections. Subsequently, we explored a spectrum of classical and modern sequential and non-sequential data-driven approaches for action classification. Among the experimented models RF, HMM-RF, and BiLSTM exhibited comparable performance, demonstrating promising outcomes.

Generally, in KTA challenges, executed actions and discerning underlying goals are pivotal. Addressing cognitive impairment concerns necessitates error detection in behavior and causal reasoning. Recognizing the limitations of classical machine learning and standalone deep learning for these challenges, our future work will showcase how a knowledge-driven probabilistic model can enhance action recognition and goal detection. We aim to enhance the reliability of sequential models. Learning from a sequence of observations can introduce errors that propagate, potentially leading to inaccurate results. Our focus will be on identifying anomalies in predictions and rectifying them using a knowledge-based probabilistic model. Fine-tuning parameters, refining features, and employing data augmentation techniques to create additional training data for a more balanced dataset are steps we'll take to boost model performance. Additionally, exploring a hybrid approach that integrates domain knowledge through a symbolic model with the capabilities of machine learning is a promising avenue. This hybrid strategy can effectively address challenges in sequential learning, provide explainability, handle imbalanced data, and enhance behavior analysis, encompassing actions, goals, and causes recognition. These directions signify promising paths for ongoing research.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions for improving the technical content and presentation of this article.

REFERENCES

[1] Y. Hou, X. Dan, M. Babbar, Y. Wei, S. G. Hasselbalch, D. L. Croteau, and V. A. Bohr, "Ageing as a risk factor for neurodegenerative disease," *Nature Rev. Neurol.*, vol. 15, no. 10, pp. 565–581, Oct. 2019.

[2] *Dementia: A Public Health Priority*, World Health Organization, Geneva, Switzerland, 2012.

[3] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 3, pp. 579–590, May 2013.

[4] L. Chen, C. Nugent, M. Mulvenna, D. Finlay, and X. Hong, "Semantic smart homes: Towards knowledge rich assisted living environments," in *Intelligent Patient Management*. Berlin, Germany: Springer, 2009, pp. 279–296.

[5] E. Castillejo, A. Almeida, D. López-de Ipina, and L. Chen, "Modeling users, context and devices for ambient assisted living environments," *Sensors*, vol. 14, no. 3, pp. 5354–5391, 2014.

[6] G. Okeyo, L. Chen, H. Wang, and R. Sterritt, "Ontology-based learning framework for activity assistance in an adaptive smart home," in *Activity Recognition in Pervasive Intelligent Environments*. Paris, France: Atlantis Press, 2011, pp. 237–263.

[7] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Trans. Syst. Man, Cybern., C, Appl. Rev.*, vol. 42, no. 6, pp. 790–808, Nov. 2012.

[8] D. Triboan, L. Chen, F. Chen, and Z. Wang, "A semantics-based approach to sensor data segmentation in real-time activity recognition," *Future Gener. Comput. Syst.*, vol. 93, pp. 224–236, Apr. 2019.

[9] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 961–974, Jun. 2012.

[10] M. Amiribesheli, A. Benmansour, and A. Bouchachia, "A review of smart homes in healthcare," *J. Ambient Intell. Humanized Comput.*, vol. 6, no. 4, pp. 495–517, Aug. 2015.

[11] J. Wang, H. Mahajan, P. Toto, A. McKeon, M. McCue, and D. Ding, "Comparison of two prompting methods in guiding people with traumatic brain injury in cooking tasks," in *Proc. Int. Conf. Smart Homes Health Telematics*. Cham, Switzerland: Springer, 2014, pp. 83–92.

[12] N. Nijhof, *eHealth for People With Dementia in Home-Based and Residential Care*. Enschede, The Netherlands: Universiteit Twente, 2013.

[13] A. Serna, H. Pigot, and V. Rialle, "Modeling the progression of Alzheimer's disease for cognitive assistance in smart homes," *User Model. User-Adapted Interact.*, vol. 17, no. 4, pp. 415–438, Aug. 2007.

[14] K. Yordanova, S. Lüdtke, S. Whitehouse, F. Krüger, A. Paiement, M. Mirmehdi, I. Craddock, and T. Kirste, "Analysing cooking behaviour in home settings: Towards health monitoring," *Sensors*, vol. 19, no. 3, p. 646, Feb. 2019.

[15] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[16] K. Yordanova, A. Hein, and T. Kirste, "Kitchen task assessment dataset for measuring errors due to cognitive impairments," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2020, pp. 1–6.

[17] K. Yordanova, A. Hein, and T. Kirste, "Artifact abstract: Kitchen task assessment dataset for measuring errors due to cognitive impairments," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2020, pp. 1–2.

[18] C. Ajlen, "*Allen Cognitive Level (ACL) Test in Assessfng Adults Functional Measu. Res and Successful Outcome (Section I)*. Rockville, MD, USA: American Occupational Therapy Foundation, 1991.

[19] C. Baum and D. F. Edwards, "Cognitive performance in senile dementia of the Alzheimer's type: The kitchen task assessment," *Amer. J. Occupational Therapy*, vol. 47, no. 5, pp. 431–436, 1993.

[20] S. Zolfaghari, S. Suravee, D. Riboni, and K. Yordanova, "Sensor-based locomotion data mining for supporting the diagnosis of neurodegenerative disorders: A survey," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–36, Aug. 2023, Art. no. 10.

[21] T. Magherini, A. Fantechi, C. D. Nugent, and E. Vicario, "Using temporal logic and model checking in automated recognition of human activities for ambient-assisted living," *IEEE Trans. Human-Mach. Syst.*, vol. 43, no. 6, pp. 509–521, Nov. 2013.

[22] F. Luo, S. Poslad, and E. Bodanese, "Kitchen activity detection for healthcare using a low-power radar-enabled sensor network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[23] M. Garcia-Constantino, A. Konios, M. A. Mustafa, C. Nugent, and G. Morrison, "Ambient and wearable sensor fusion for abnormal behaviour detection in activities of daily living," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2020, pp. 1–6.

[24] Y. Nakauchi, T. Suzuki, A. Tokumasu, and S. Murakami, "Cooking procedure recognition and inference in sensor embedded kitchen," in *Proc. RO-MAN-18th IEEE Int. Symp. Robot Hum. Interact. Commun.*, Sep. 2009, pp. 593–600.

[25] M. Ogawa, S. Ochia, K. Otsuka, and T. Togawa, "Remote monitoring of daily activities and behaviors at home," in *Proc. 23rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 4, Oct. 2001, pp. 3973–3976.

[26] M. Garcia-Constantino, A. Konios, and C. Nugent, "Modelling activities of daily living with Petri nets," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 866–871.

[27] S. Zhang, S. McClean, B. Scotney, P. Chaurasia, and C. Nugent, "Using duration to learn activities of daily living in a smart home environment," in *Proc. 4th Int. Conf. Pervasive Comput. Technol. Healthcare*, Mar. 2010, pp. 1–8.

[28] J. Wagner, A. van Halteren, J. Hoonhout, T. Ploetz, C. Pham, P. Moynihan, D. Jackson, C. Ladha, K. Ladha, and P. Olivier, "Towards a pervasive kitchen infrastructure for measuring cooking competence," in *Proc. 5th Int. Conf. Pervasive Comput. Technol. Healthcare (PervasiveHealth) Workshops*, May 2011, pp. 107–114.

[29] M. Schroth, A. Ilg, L. Kohout, and W. Stork, "A method for designing an embedded human activity recognition system for a kitchen use case based on machine learning," in *Proc. IEEE Int. Conf. Ind. 4.0, Artif. Intell., Commun. Technol. (IAICT)*, Jul. 2022, pp. 249–254.

[30] P. Urwyler, R. Stucki, R. Müri, U. P. Mosimann, and T. Nef, "Passive wireless sensor systems can recognize activites of daily living," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 8042–8045.

[31] S. Chatterjee, B. Mitra, and S. Chakraborty, "AmicroN: Framework for generating micro-activity annotations for human activity recognition," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2022, pp. 26–31.

[32] O. Mur, M. Frigola, and A. Casals, "Modelling daily actions through hand-based spatio-temporal features," in *Proc. Int. Conf. Adv. Robot. (ICAR)*, Jul. 2015, pp. 478–483.

[33] N. Žarić, M. Radonjić, N. Pavlićević, and S. P. Žarić, "Design of a kitchen-monitoring and decision-making system to support AAL applications," *Sensors*, vol. 21, no. 13, p. 4449, 2021.

[34] A. Hein, S. Winkelbach, B. Martens, O. Wilken, M. Eichelberg, J. Spehr, M. Gietzelt, K.-H. Wolf, F. Büsching, and M. Hülsken-Giesler, "Monitoring systems for the support of home care," *Inform. Health Social Care*, vol. 35, nos. 3–4, pp. 157–176, 2010.

[35] G. Bouaziz, D. Brulin, H. Pigot, and E. Campo, "Detection of social isolation based on meal-taking activity and mobility of elderly people living alone," *IRBM*, vol. 44, no. 4, Aug. 2023, Art. no. 100770.

[36] M. Plananamente, C. Plizzari, and B. Caputo, "Test-time adaptation for egocentric action recognition," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2022, pp. 206–218.

[37] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo, "Domain generalization through audio-visual relative norm alignment in first person action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 1807–1818.

[38] R. Granada, J. Monteiro, N. Gavenski, and F. Meneguzzi, "Object-based goal recognition using real-world data," in *Proc. Mex. Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2020, pp. 325–337.

[39] Y.-P. Huang, H. Basanta, H.-C. Kuo, and H.-T. Chiao, "Sensor-based detection of abnormal events for elderly people using deep belief networks," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 33, no. 1, pp. 36–47, 2020.

[40] Z. Moore, C. Sifferman, S. Tullis, M. Ma, R. Proffitt, and M. Skubic, "Depth sensor-based in-home daily activity recognition and assessment system for stroke rehabilitation," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1051–1056.

[41] N. E. Tabbakha, W.-H. Tan, and C.-P. Ooi, "Elderly action recognition system with location and motion data," in *Proc. 7th Int. Conf. Inf. Commun. Technol. (ICoICT)*, Jul. 2019, pp. 1–5.

[42] M. Ma, B. J. Meyer, L. Lin, R. Proffitt, and M. Skubic, "VicoVR-based wireless daily activity recognition and assessment system for stroke rehabilitation," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 1117–1121.

[43] J. D'Agostini, L. Bonetti, A. Salem, L. Passerini, G. Fiacco, P. Lavanda, E. Motti, M. Stocco, K. Gashay, and E. Abebe, "An augmented reality virtual assistant to help mild cognitive impaired users in cooking a system able to recognize the user status and personalize the support," in *Proc. Workshop Metrol. Ind. 4.0 IoT*, Apr. 2018, pp. 12–17.

[44] G. Azkune and A. Almeida, "A scalable hybrid activity recognition approach for intelligent environments," *IEEE Access*, vol. 6, pp. 41745–41759, 2018.

[45] E. Mavroudi, D. Bhaskara, S. Sefati, H. Ali, and R. Vidal, "End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1558–1567.

[46] J. Korpela and T. Maekawa, "Privacy preserving recognition of object-based activities using near-infrared reflective markers," *Pers. Ubiquitous Comput.*, vol. 22, no. 2, pp. 365–377, Apr. 2018.

[47] I. M. Pires et al., "Android library for recognition of activities of daily living: Implementation considerations, challenges, and solutions," *Open Bioinf. J.*, vol. 11, pp. 61–88, 2018.

[48] J. Collins, J. Warren, M. Ma, R. Proffitt, and M. Skubic, "Stroke patient daily activity observation system," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 844–848.

[49] Y. Mohammad, K. Matsumoto, and K. Hoashi, "A dataset for activity recognition in an unmodified kitchen using smart-watch accelerometers," in *Proc. 16th Int. Conf. Mobile Ubiquitous Multimedia*, Nov. 2017, pp. 63–68.

[50] J. Monteiro, R. Granada, R. C. Barros, and F. Meneguzzi, "Deep neural networks for kitchen activity recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2048–2055.

[51] K. Yordanova, S. Whitehouse, A. Paiement, M. Mirmehdi, T. Kirste, and I. Craddock, "What's cooking and why? Behaviour recognition during unscripted cooking tasks for health monitoring," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2017, pp. 18–21.

[52] R. L. Granada, R. F. Pereira, J. Monteiro, D. D. A. Ruiz, R. C. Barros, and F. R. Meneguzzi, "Hybrid activity and plan recognition for video streams," in *Proc. 31st AAAI Conf., Plan, Activity Intent Recognit. Workshop*, 2017, pp. 819–825.

[53] Z. Lu, Y. Y. Chung, H. W. F. Yeung, S. M. Zandavi, W. Zhi, and W.-C. Yeh, "Using hidden Markov model to predict human actions with swarm intelligence," in *Proc. 24th Int. Conf. Neural Inf. Process. (ICONIP)*. Guangzhou, China: Springer, Nov. 2017, pp. 21–30.

[54] R. L. Granada, J. Monteiro, R. C. Barros, and F. R. Meneguzzi, "A deep neural architecture for kitchen activity recognition," in *Proc. 30th Florida Artif. Intell. Res. Soc. Conf.*, 2017, pp. 2048–2055.

[55] T. Giannakopoulos and S. Konstantopoulos, "Daily activity recognition based on meta-classification of low-level audio events," in *Proc. ICT4AgeingWell*, 2017, pp. 220–227.

[56] C. Flores-Vázquez and J. Aranda, "Human activity recognition from object interaction in domestic scenarios," in *Proc. IEEE Ecuador Tech. Chapters Meeting (ETCM)*, Oct. 2016, pp. 1–6.

[57] S. Awwad and M. Piccardi, "Local depth patterns for fine-grained activity recognition in depth videos," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2016, pp. 1–6.

[58] Y. Xu, D. Bull, and D. Damen, "Unsupervised daily routine modelling from a depth sensor using top-down and bottom-up hierarchies," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 56–60.

[59] S. Suzuuchi and M. Kudo, "Location-associated indoor behavior analysis of multiple persons," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2079–2084.

[60] V. Magnanimo, M. Saveriano, S. Rossi, and D. Lee, "A Bayesian approach for task recognition and future human activity prediction," in *Proc. 23rd IEEE Int. Symp. Robot Human Interact. Commun.*, Aug. 2014, pp. 726–731.

[61] F.-G. Wu and T.-H. Tsai, "Building a recognition process of cooking actions for smart kitchen system," in *Proc. 8th Int. Conf. Universal Access Hum.-Comput. Interact. (UAHCI)*. Heraklion, Greece: Springer, Jun. 2014, pp. 575–586.

[62] A. Shimada, K. Kondo, D. Deguchi, G. Morin, and H. Stern, "Kitchen scene context based gesture recognition: A contest in ICPR2012," in *Proc. Int. Workshop Depth Image Anal. Appl.* Berlin, Germany: Springer, 2012, pp. 168–185.

[63] J. Ortiz, A. García-Olaya, and D. Borrajo, "Using activity recognition for building planning action models," *Int. J. Distrib. Sensor Netw.*, vol. 9, no. 6, Jun. 2013, Art. no. 942347.

[64] I. Duque, K. Dautenhahn, K. L. Koay, L. Willcock, and B. Christianson, "Knowledge-driven user activity recognition for a smart house. Development and validation of a generic and low-cost, resource-efficient system," in *Proc. 6th Int. Conf. Adv. Comput.-Hum. Interact. (ACHI)*. Nice, France: International Academy, Research and Industry Association, 2013, pp. 141–146.

[65] J. Hoey, T. Plötz, D. Jackson, A. Monk, C. Pham, and P. Olivier, "Rapid specification and automated generation of prompting systems to assist people with dementia," *Pervas. Mobile Comput.*, vol. 7, no. 3, pp. 299–318, Jun. 2011.

[66] C. Pham, T. Plötz, and P. Olivier, "A dynamic time warping approach to real-time activity recognition for food preparation," in *Proc. Int. Joint Conf. Ambient Intell.* Berlin, Germany: Springer, 2010, pp. 21–30.

[67] T. Magherini, G. Parente, C. D. Nugent, M. P. Donnelly, E. Vicario, F. Cruciani, and C. Paggetti, "Temporal logic bounded model-checking for recognition of activities of daily living," in *Proc. 10th IEEE Int. Conf. Inf. Technol. Appl. Biomed.*, Nov. 2010, pp. 1–4.

[68] J. Biswas, K. Sim, W. Huang, A. Tolstikov, A. Aung, M. Jayachandran, V. Foo, and P. Yap, "Sensor based micro context for mild dementia assistance," in *Proc. 3rd Int. Conf. Pervasive Technol. Rel. Assistive Environ.*, Jun. 2010, pp. 1–4.

[69] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops, ICCV Workshops*, Sep. 2009, pp. 1089–1096.

[70] R. Aipperspach, E. Cohen, and J. Canny, "Modeling human behavior from simple sensors in the home," in *Proc. 4th Int. Conf. Pervasive Comput.* Dublin, Ireland: Springer, May 2006, pp. 337–348.

[71] C. Lin and J. Y. Hsu, "IPARS: Intelligent portable activity recognition system via everyday objects, human movements, and activity duration," in *Proc. AAAI Workshop Modeling Others Observ.*, 2006, pp. 1–9.

[72] A. Vafeiadis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Audio content analysis for unobtrusive event detection in smart homes," *Eng. Appl. Artif. Intell.*, vol. 89, Mar. 2020, Art. no. 103226.

[73] E. Thomaz, V. Bettadapura, G. Reyes, M. Sandesh, G. Schindler, T. Plötz, G. D. Abowd, and I. Essa, "Recognizing water-based activities in the home through infrastructure-mediated sensing," in *Proc. ACM Conf. Ubiquitous Comput.*, Sep. 2012, pp. 85–94.

[74] P. Lago, S. Takeda, S. Shamma Alia, K. Adachi, B. Bennai, F. Charpillet, and S. Inoue, "A dataset for complex activity recognition withmicro and macro activities in a cooking scenario," 2020, *arXiv:2006.10681*.

[75] K. Yordanova and T. Kirste, "A process for systematic development of symbolic models for activity recognition," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–35, 2015.

[76] K. Yordanova, F. Krüger, and T. Kirste, "Providing semantic annotation for the CMU grand challenge dataset," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 579–584.

[77] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-08-22, Apr. 2009.

[78] C. Pham and P. Olivier, "Slice&Dice: Recognizing food preparation activities using embedded accelerometers," in *Proc. Eur. Conf. Ambient Intell.* Berlin, Germany: Springer, 2009, pp. 34–43.

[79] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Ishiguro, K. Kamei, Y. Sakurai, and T. Okadome, "Object-based activity recognition with heterogeneous sensors on wrist," in *Proc. Int. Conf. Pervasive Comput.* Berlin, Germany: Springer, 2010, pp. 246–264.

[80] T. Stoev and K. Yordanova, "BehavE: Behaviour understanding through automated generation of situation models," in *Proc. German Conf. Artif. Intell. (Künstliche Intelligenz)*. Cham, Switzerland: Springer, 2021, pp. 362–369.

[81] I. Psychoula, L. Chen, and O. Amft, "Privacy risk awareness in wearables and the Internet of Things," *IEEE Pervasive Comput.*, vol. 19, no. 3, pp. 60–66, Jul. 2020.

[82] C. Phua, V. S.-F. Foo, J. Biswas, A. Tolstikov, A.-P.-W. Aung, J. Maniyeri, W. Huang, M.-H. That, D. Xu, and A. K.-W. Chu, "2-layer erroneous-plan recognition for dementia patients in smart homes," in *Proc. 11th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Dec. 2009, pp. 21–28.

[83] K. Sim, G.-E. Yap, C. Phua, J. Biswas, A. A. Phyo Wai, A. Tolstikov, W. Huang, and P. Yap, "Improving the accuracy of erroneous-plan recognition system for activities of daily living," in *Proc. 12th IEEE Int. Conf. e-Health Netw., Appl. Services*, Jul. 2010, pp. 28–35.

[84] S. Zolfaghari, M. R. Keyvanpour, and R. Zall, "Analytical review on ontological human activity recognition approaches," *Int. J. E-Bus. Res.*, vol. 13, no. 2, pp. 58–78, Apr. 2017.

[85] L. Chen, C. Nugent, and A. Al-Bashrawi, "Semantic data management for situation-aware assistance in ambient assisted living," in *Proc. 11th Int. Conf. Inf. Integr. Web-Based Appl. Services*, Dec. 2009, pp. 298–305.

[86] L. Chen, N. R. Shadbolt, F. Tao, C. Goble, C. Puleston, and S. J. Cox, "Semantics-assisted problem solving on the semantic grid," *Comput. Intell.*, vol. 21, no. 2, pp. 157–176, May 2005.

[87] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A professional framework for multimodality research," in *Proc. 5th Int. Conf. Lang. Resour. Eval. (LREC)*, 2006, pp. 1556–1559.

[88] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. Workshop Empirical Methods Artif. Intell. (IJCAI)*, vol. 3, no. 22, 2001, pp. 41–46.

[89] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Aug. 1998.

[90] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[91] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, Jan. 1991.

[92] Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, 2015.

[93] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classifiaction," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 683–697, 1992.

[94] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*. New York, NY, USA: Springer, 2013, pp. 237–280.

[95] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic Regression*. Totowa, NJ, USA: Springer, 2002.

[96] M. V. Albert, A. Sugianto, K. Nickele, P. Zavos, P. Sindu, M. Ali, and S. Kwon, "Hidden Markov model-based activity recognition for toddlers," *Physiol. Meas.*, vol. 41, no. 2, Feb. 2020, Art. no. 025003.

[97] D. Liciotti, M. Bernardini, L. Romeo, and E. Frontoni, "A sequential deep learning application for recognising human activities in smart homes," *Neurocomputing*, vol. 396, pp. 501–513, Jul. 2020.

[98] D. Das, Y. Nishimura, R. P. Vivek, N. Takeda, S. T. Fish, T. Ploetz, and S. Chernova, "Explainable activity recognition for smart home systems," 2021, *arXiv:2105.09787*.

[99] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[100] R. Das and S. Chaudhuri, "On the separability of classes with the cross-entropy loss function," 2019, *arXiv:1909.06930*.

[101] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009, pp. 1–4.

SAMANEH ZOLFAGHARI received the B.S. degree in computer engineering with a specialization in software from Shahid Rajaee University, Iran, in 2013, the M.S. degree in computer engineering with a specialization in software from Alzahra University in 2016, and the Ph.D. degree in computer science from the University of Cagliari, Italy.

She is currently a Postdoctoral Researcher with Mälardalen University, Sweden. In addition to her academic achievements, she was a Visiting Scholar with the Junior Research Group "Cognitive Methods for Situation-Aware Assistive Systems (CoMSA^{2}t)," University of Rostock, from July 2021 to December 2022. From January 2023 to March 2023, she continued her research endeavors with the Institute for Data Science, University of Greifswald, Germany. Her research interests include information retrieval, machine learning, big data, mobile and pervasive systems, and human factors in pervasive computing applications.

TEODOR STOEV received the B.Sc. and M.Sc. degrees in computer science from the University of Rostock, Germany. He is currently pursuing the Ph.D. degree with the Institute for Data Science, University of Greifswald, Germany. He was a Researcher with the University of Rostock, from September 2020 to September 2022. His research interests include natural language processing, machine learning, data fusion, behavioral models, and ontologies. From 2017 to 2020, he was a Student Trainee and a Research Assistant at one of the biggest German direct banks—The Comdirect Bank. In May 2020, he was a Researcher with the Junior Research Group, University of Rostock "Cognitive Methods for Situation-Aware Assistive Systems" (led by Dr.-Ing. Kristina Yordanova). He was also the Co-Chair of the Fifth and Sixth ARDUOUS Workshops on Data Annotation, which are affiliated with PerCom.

KRISTINA YORDANOVA received the bachelor's degree in computer engineering from the University of Duisburg-Essen, Germany, in 2008, the master's degree in artificial intelligence from Maastricht University, The Netherlands, in 2009, and the Ph.D. degree in ubiquitous computing from the University of Rostock, Germany, in 2014. Until 2022, she was leading the Junior Research Group "Cognitive Methods for Situation-Aware Assistive Systems," University of Rostock. She is currently a Full Professor in data science and the Head of the Institute for Data Science, University of Greifswald. Her research interests include natural language processing, machine learning, and symbolic and probabilistic modelling, with applications in assistive systems, social sciences, healthcare, and medicine.