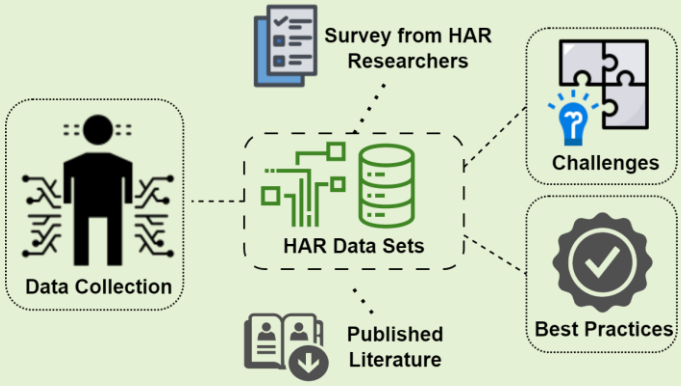
Open Data Sets in Human Activity Recognition Research - Issues and Challenges: A Review

Open Data Sets in Human Activity Recognition Research - Issues and Challenges: A Review

Gulzar Alam, Ian McChesney, Peter Nicholl, and Joseph Rafferty

***Abstract*— Huge amounts of data is generated with the emergence of new sensors technologies. Human Activity Recognition (HAR) data sets are generated from cameras, such as video or still images, capturing human behaviour through sensors like gyroscope, Bluetooth, sound sensors, and accelerometers. These generated data sources are collected by the researchers and formed into open data sets. However, these data sets often show issues during data set construction, sharing, searching, that could produce further challenges for reuse of the data by others. The main objective of this research is to explore the current issues and challenges faced by researchers in the HAR domain. A detail literature review was conducted to extract information from the published literature. Similarly, a questionnaire survey was sent to selected researchers having expertise in the HAR domain and who work with open data sets. The main issues and challenges were identified and classified into a hierarchical structure. This research will help HAR researchers to be aware of the current issues and challenges in the field of HAR open data sets. It will help to promote important attributes applicable to many open data sets, such as privacy, anonymity, platform maintenance, data sets descriptions, metadata, environmental conditions, resources, and training, while constructing and sharing new data sets.**



***Index Terms*— Human activity recognition (HAR), Open data set lifecycle, Data set quality, Data sets issues and challenges, Artificial intelligence**

I. Introduction

With the emergence of new computing technologies that are capable of capturing huge amounts of data on human activities [1], there is a need to store the information in a structured, meaningful and sharable way. Generated data are collected in the form of data sets. Across many areas of business and society, these data sets facilitate analysis, forecasting and decision making which can significantly impact quality of life, ranging from the optimization of supply chains to the transformation of healthcare systems. For example, within the realm of business, these data sets function as a dynamic tool for decision-makers, enabling them to develop targeted marketing plans, forecast future market patterns, reduce time to market and stimulate economic growth [2], [3], [4]. Within the domain of healthcare, these datasets serve as a catalyst for pioneering research, facilitating the development of tailored therapies, and accelerating advancements in the field of medicine [5], [6]. Other areas in which open data sets are having an impact include urban planning and public services [7], and the world of entertainment and virtual reality [8], [9]. Researchers and practitioners collect data sets for research objectives to understand the totality of an area of interest and to develop a basis for making decisions. Similarly, the researchers aim to investigate and tackle the challenges associated with ensuring the FAIRness (findability, accessibility, interoperability, reusability) of expanding biomedical data sets housed in diverse repositories. Their objective is to improve transparency, reproducibility, and the progress of research by promoting open science practices and the reuse of data [10]. The primary objective of constructing and sharing open data sets, metadata, related data set publications, and results is to encourage open data set benchmarking, replication, validation of research approaches, applied data analysis practices, detection of experimental errors and exploration of novel hypotheses [11], [12], [13].

Open data sets consist of those data that are freely available and accessible for use and sharing. Various public/private sector organizations and individuals use these data sets for health, social, environmental, and economic improvement. The significance of open data sets is due to the advances in data-driven systems and research across the globe. For

This paragraph of the first footnote will contain the date on which you submitted your paper for review. We acknowledge Department for the Economy (DfE) Ulster University for financial support.

Gulzar Alam, Ian McChesney, Peter Nicholl and Joseph Rafferty are with the School of Computing, Ulster University, York Street, Belfast, BT15 1AP, Northern Ireland, United Kingdom (e-mail: alam-g@ulster.ac.uk; ir.mcchesney@ulster.ac.uk; p.nicholl@ulster.ac.uk; j.rafferty@ulster.ac.uk).

example, data sets from healthcare can help to enrich personal care, speed up diagnosis, disease prediction and planning of treatment, as well as other benefits [14].

HAR systems can typically collect, store, process, and share data in any format over any network describing specific human activities and environments [15]. Researchers and data science practitioners are exploring new approaches to data visualization, building processes to link users with sensors/devices, understanding the contextual importance of sensor/device, data annotation, and data management during open data set creation and sharing [16]. Pervasive computing has made significant progress in providing insight into human activity in a range of natural settings through the data generated by intricate sensors and actuators, establishing a connected globe that produces massive amounts of data [17], [18], [19].

HAR has become an emerging research area due to the development of devices and sensors with minimum cost, low power consumption, and real-time streaming of data combined with sophisticated processing capabilities through technologies such as artificial intelligence (AI), machine learning (ML), and Internet of Things (IoT) [20]. HAR is the detection of movements/activities (walking, running, standing, sitting, drinking, talking, sleeping, etc.) of an individual based on sensor/device data [21]. HAR data collection can involve data from a camera such as video or images comprising human sentiments and from sensors like gyroscope, Bluetooth, sound sensors and accelerometers. The data received from devices/sensors are connected via a network or attached to the human body [22]. HAR can perform a significant role in human daily life activities by understanding people's behaviours from collected sensor data. Researchers apply machine learning and other statistical techniques for extracting important features, activities and patterns from video and sensor based HAR data sets [23], [24].

However, there are issues and challenges when using HAR open data sets such as recognizing similar and dissimilar actions, emotion recognition from video and images, detecting background noise in activity recording, missing values in a data set and data set annotation. The number of issues and challenges grows due to the numerous and complex activities being recorded, sensor placement and orientation, camera movement, and the increasing number of activity types being studied such as person only, person and object and person to person activities. Further issues and challenges related to HAR open data sets are the annotation/labelling ground truth, activity detection among multiple participants, sensor orientation and heterogeneity, different data formats, rogue sensors values, missing values, imbalanced data, and background environment [25].

The objective of this paper is to explore the current issues and challenges faced by researchers in the HAR domain when using open data sets. A dual approach has been taken involving a comprehensive literature review to extract themes from the published literature (January 2016 to February 2023) and a questionnaire survey of researchers having expertise in the use of open data sets in the HAR domain. The main contributions arising from this research work are:

- A conceptual framework of the open data set lifecycle.
- Greater exploration of the issues and challenges related to HAR open data sets through a questionnaire survey.
- Identification of HAR data sets issues and challenges through a comprehensive literature review.
- Derivation of useful insight from the analysis and comparisons of both survey and literature review results.
- Evidence-based classification of issues and challenges in the use of open data sets in HAR.

In contrast to previous published literature studies such as [23], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], this work specifically examines the issues and challenges faced by researchers in the field of HAR with regards to open data sets. Similarly, the creation of open data set lifecycle conceptual framework and evidence-based classification of issues and challenges. In addition to referring existing literature, we conducted a focussed questionnaire survey with experts in the HAR domain, thereby obtaining timely and firsthand information. Through the process of comparing literature findings with survey responses, we can gain a deeper understanding of the practical challenges that researchers endure. our research presents an opportunity for significant progress in this domain. This has the potential to bring about groundbreaking developments in the utilization and interpretation of open datasets within HAR research, fundamentally modifying our perspectives and methodologies.

The remainder of this paper is structured as follows: Section II describes the research background and Section III presents the research methodology. Section IV discusses the results from both the conducted literature review and the survey and then Section V presents an analysis and comparison of the survey and literature review results. Section VI illustrates the classification of open issues and challenges related to HAR open data sets and Section VII discusses the derived HAR data sets future scope from the research study. Finally, Section VIII describes the conclusion and future work of the proposed research.

A. Motivation and scope

HAR researchers are facing issues and challenges when interacting with open data sets. The main issues and challenges are discussed comprehensively in Section IV. By addressing these issues and challenges, the aim is to improve data set structure, quality, access and to facilitate analysis, forecasting and decision making that can then be more readily and reliably shared within the research community. Long term, good quality data sets can benefit personal healthcare, rehabilitation, early disease detection and globalization of data sets.

The scope of this paper is related to open data sets in the HAR domain, covering both video-based and wearable-based recognition systems. Similarly, the questionnaire survey was conducted by involving HAR researchers and a comprehensive literature review was performed considering HAR studies.

II. Research Background

Open data sets consist of those data that are freely available and accessible for use and sharing. Various public/private sector organizations and individuals use these open data sets for health, social, environmental, and economic purposes. The significance of data sets is due to the increase in data-driven systems across the globe [36], [37]. Hence, open data can allow a deeper understanding of worldwide trends and common problems. It can play a role, in combination with ML and statistics, to solve problems in the domains of healthcare, engineering and science. Similarly, it can encourage international collaboration and improve transparency [38]. They can modernize the development processes and systems built by governments, private organizations and societies. A useful characterization of open data sets is given below as described by the Organisation for Economic Co-operation and Development (OECD)[39], [40].

Redistribution and reusability: open data sets should be reusable and distributed for commercial and non-commercial use. To ensure reusability, they must be properly licensed, well-structured and in a machine-readable format [25]. The important concern of redistribution and reusability intersects with the fundamental principle of data ownership in open data sets. One noteworthy fact is that data owners frequently restrict redistribution, resulting in a complex system in which accessibility and sharing coexist with ownership rights. The dynamic interaction of these factors influences the extent to which open data sets can be freely shared and used for a variety of purposes [41]. The complexities of data ownership delicately thread themselves into the fabric of data transmission, influencing the possibilities for expanded reuse and collaboration. Recognizing the traditional constraints surrounding redistribution, it becomes critical to strike a delicate balance between facilitating open access and respecting ownership privileges, fostering a discourse that champions equitable data utilization while recognizing the critical role of data curators [42].

Open access and availability: the data sets should have transparent open access arrangements and should be available for download, reuse and sharing [43]. Various platforms such as World Bank Open Data, WHO open data repository, European Union Open Data Portal, UCI Machine Learning Repository, U.S. Census Bureau, Data.gov, DBpedia, UNICEF Dataset and Kaggle [44] provide open access to data sets for research use. They must also be available with full documentation, user guidelines, and a modifiable format.

General Participation: everybody should be able to utilize open data sets, reuse them and further redistribute them regardless of the domain. Every domain should use open data for its intended purpose with no discrimination and opposition to any person, team or field of work [45]. A good example is how non-commercial use of data sets prevents commercial use of it and keeps restrictions on data sets used only for certain purposes that are not permitted for general use.

Interoperability: this is the ability to combine diverse data sets from different systems and organisations. This permits various components to work together and merge several data sets to develop larger and more complex systems for solving complicated problems [40].

HAR data collection is mainly conducted based on two methods namely video data collection and sensor data collection [46]. It has been successfully applied to individual behaviour analysis [47], movement recognition [48], gait analysis [49] and video surveillance [50]. Researchers apply machine learning and other statistical techniques for extracting important features and activities from video and sensor based HAR data sets [24].

The utilization of HAR data sets specifically designed for individuals with disabilities has great importance. The utilization of these data sets holds the potential to greatly enhance the quality of life for those with disabilities, making them more independent and self-sufficient. The use of these data sets has facilitated the development of HAR systems by researchers, which in turn have the potential to assist those with disabilities in activities of daily living (ADLs). Developing these data sets has the capacity to enhance the quality of life for those with disabilities by fostering more independence and self-sufficiency. Kim et al. [51] collected the MyMove data set spanning a duration of 7 days, a group of 13 elderly individuals participated by gathering activity labels and wristwatch sensor data. Leving et al. [52] collected Activ8 data set by involving 16 able-bodied individuals who performed 16 various standardized 60s activities of daily living. Additional data sets have been created to cater to the needs of those with disabilities, encompassing KFall [53], CAUCAFall [54] and SisFall [55] .

Sing et al. [56] surveyed the use of different data sets in HAR for the research community. They discussed data set benchmarking, comparisons, and improvement of building data sets. Relevant features are explored such as participant classes and data source for data set construction, focus area of the data sets, modality, annotations and evaluation of HAR data sets. These are outlined below:

- ***Classes*** such as person only, person and object and person to person [57].
- ***Focus*** is the type of activity under observation such as sports, gaming, surveillance, healthcare, etc. [58].
- ***Modality*** is related to temporal based or spatial based data [59], [60].
- ***Data source*** concerns the origin of the data such as sensor placement, whether the data is recorded or scripted, data labelling procedure and finally whether it is a generated data set or crowdsourced [61], [62].
- ***Annotation*** is concerned with the correct annotation of the data set and confirmation of labelling actions/activities [63].
- ***Evaluation*** is checking the accuracy of the data set, such as identifying imbalanced data and missing values [64].

Established data set repositories having HAR data sets are UCI Machine Learning Repository [65], Harvard Dataverse [66], Dataset search from Google [67], IEEE Dataport [68], Zonedo [69] and Figshare [70]. Some of the popular HAR data sets available include Hollywood [71], Action Similarity LAbeliNg (ASLAN) [72], YouCook [73], ActivityNet [57], CASAS [74] and Opportunity [75]. Data

sets in HAR are illustrated in Fig. 1 in the form of a word cloud generated through an online tool[1] from the published SLR research studies. Largest fonts of words show the highest frequency of a data set to the lowest fonts with low words frequency. Some data sets that are commonly used by the researchers are EmotiW [76], CASAS [74], WISDM [77], ARUBA [74] , PAMAP [78], SHL [79], Opportunity [75], and WHARF [80].

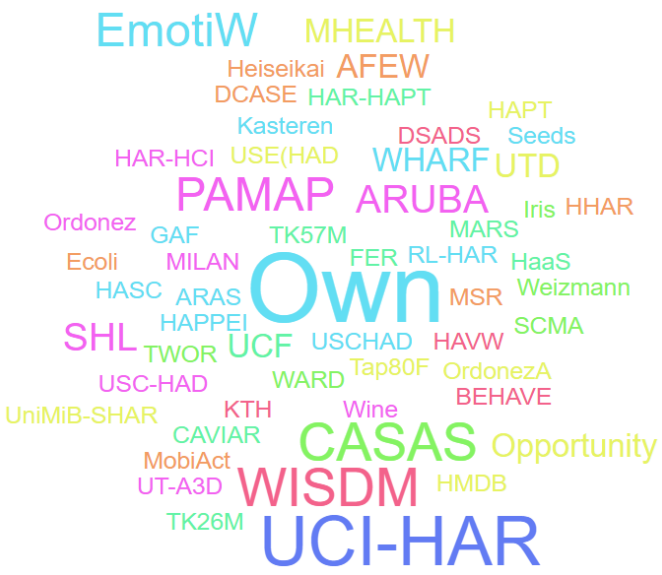


Fig. 1. Word cloud of HAR data sets explored in literature review.

However, HAR open data sets present researchers with issues and challenges such as recognizing similar and dissimilar actions, emotion recognition from video and images, background noise in activity recording, missing values in a data set, data set annotation and so on see in Section IV. These issues and challenges are growing due to the numerous and complex activities involved, sensor placement and orientation, camera movement and the increasing number of classes and activities being studied.

There has been a rise in the use of network analysis as a method for exploring the interconnections between various concepts and phrases in the academic literature [81]. Unit analysis as a keyword and type analysis as co-occurrence allow researchers to find important phrases that are regularly linked with the unit of study, allowing them to begin to construct a more comprehensive knowledge of the issue at hand [82]. The constructed network analysis as shown in Fig. 2 consist of the relationship between keyword and co-occurrence, developed via VOSviewer[2]. The colours represent the keyword categories.

As shown in Fig. 2, the terms machine learning (ML), convolutional neural network, classification, support vector machine, feature representation, feature extraction, random forest, learning systems, and training were received in the graph are related to artificial intelligence and ML. In the context of ML and artificial intelligence, each of these concepts is fundamental to the study of unit analysis. ML is a subfield of computer science that uses statistical models and algorithms to enable computer systems to learn from data without being explicitly programmed [83]. The graph depicts the most prevalent ML methods, including convolutional neural networks, support vector machines, and random forests. Convolutional neural networks are a form of artificial neural network that is frequently used for image categorization and identification applications [84]. Support vector machines are a form of machine learning technique utilised often for classification problems [85]. They are particularly efficient when the data is high-dimensional and have the ability to effectively handle data sets with a large number of features is a notable benefit of SVMs, making them very capable in complex scenarios [86]. Random forests, on the other hand, are an ensemble learning method for classification, regression, and other tasks that function by constructing a large number of decision trees at training time and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees [87]. Both feature representation and feature extraction are essential topics in machine learning. Feature representation is the manner in which data is represented, whereas feature extraction is the process of extracting features from data. In machine learning, learning systems and training are both fundamental ideas. Learning systems are the techniques and models used to train a machine learning system, whereas training is the act of adjusting the model's parameters to the input [88].

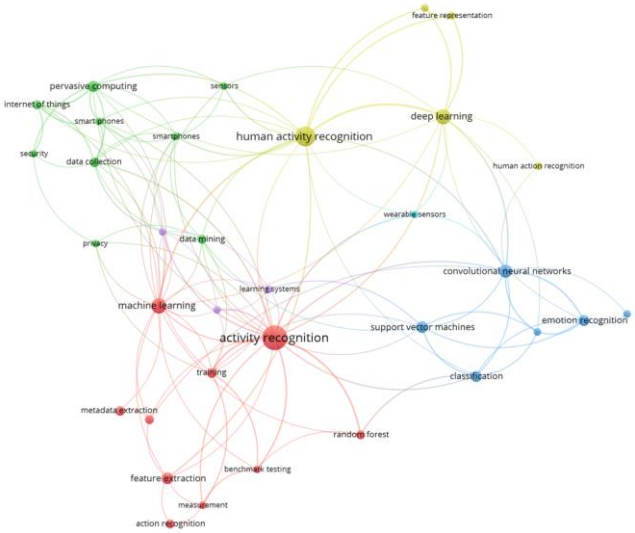


Fig. 2. Network analysis of the conducted literature review studies

Similarly, Human activity recognition (HAR), data collection, pervasive computing, activity action recognition, emotion recognition, benchmark testing, action recognition, measurement, metadata extraction, data mining, privacy, sensors, smartphones, security, the Internet of Things, and wearable sensors are all terms gathered from the network analysis of the graph in Fig. 2 In the context of recognising human activities, unit analysis is important to each of these words.

HAR is a field of activity recognition that employs sensor data and other information sources to recognise and

[1] https://worditout.com/word-cloud/create
[2] https://www.vosviewer.com/

comprehend human actions. This is possible in several settings, including healthcare and fitness. Activity recognition is a similar term that applies to the identification and categorization of various activity types [89]. This may be accomplished using a variety of methods, such as ML algorithms and statistical models. Pervasive computing refers to the use of technology to build highly interconnected and interactive environments. This may be utilised in the context of human activity identification to produce precise and dependable systems [90]. Emotion recognition uses sensors and other data sources to recognise and comprehend human emotions. This is applicable in several settings, including healthcare and personal fitness [91].

Data collection is a vital stage in the process of recognising human activities. This involves collecting data from a number of sources, such as sensors, cameras, and other devices, and then analysing that data to derive actionable insights[92]. In the realm of HAR, benchmark testing and action recognition are both essential. Benchmark testing is the process of evaluating the accuracy and dependability of various algorithms and models, whereas action recognition is the identification of particular actions or gestures [93]. Important phases in the process of evaluating and interpreting sensor data include measurement, extraction of information, and data mining. The extraction of important characteristics from the data followed by the application of ML algorithms to find patterns and trends [28]. In the context of HAR, privacy and security are crucial issues, especially when data is collected from a growing number of sources and participants. Wearable sensors, smartphones, and the internet of things are all key data sources that may be used in this context; nevertheless, it is crucial that this data be gathered and utilised in an ethical manner [94].

A. Open Data Set Lifecycle

An open data set lifecycle is presented in the form of a conceptual framework consisting of four main phases namely construction, sharing, finding, and using a data set as shown in Fig. 2. We believe this is the first open data set lifecycle framework in computing and HAR domain [25], and this framework was influenced from Stampers et al. [95] research work. We use this lifecycle to organize the conducted questionnaire survey. As shown in Fig. 3, our proposed life cycle consists of the following four phases.

1) Construction: In this phase researchers design a protocol for data collection according to the main goal of their research. This phase involves the detailed documentation and description of the data set. Description includes specific elements such as data set title, creation date, subject description, sensor and device information, actions and activities, supported data format and so on. Data is collected in various formats such as video, audio, images, and text. During the data set construction phase, a significant emphasis is placed on carefully establishing the fundamental aspects of the dataset's integrity and usefulness. The annotation protocol assumes an important role in this context since it establishes explicit criteria and procedures for the systematic labelling and categorization of data points. Concurrently, it is crucial to prioritize the pursuit of validity, ensuring that every annotation accurately aligns with the facts of ground truth. The dataset's dependability is supported by the implementation of rigorous validation methods and quality assurance mechanisms, which effectively mitigate the risk of mistakes or inconsistencies. As the data set undergoes transformation via the use of these techniques, its creation becomes a careful undertaking characterized by precision and diligence. This process lays the foundation for the following stages of data sharing, finding, and using.

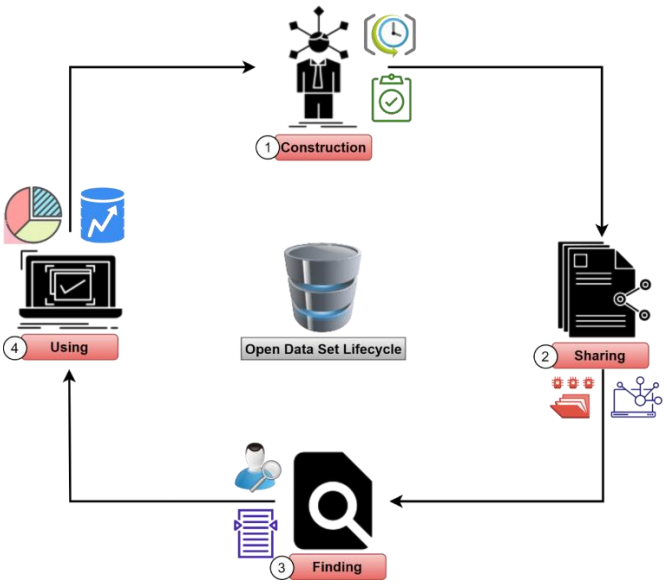


Fig. 3. Open data sets life cycle

2) Sharing: This phase is related to sharing the data set and making it available to the research community. The data set owner grants access for using the data set for research purposes. The data sets are stored in a centralized or local repository depending on the data storage and access arrangements as agreed in the research proposal and also the distribution license determined for the data set.

3) Finding: This phase is where data set users (researchers) search and find a data set for achieving their research objective. Researchers use different search terms and keywords to retrieve a data set from a central repository or to retrieve it from a collaborative research group via a local repository.

4) Using: In this phase, users/researchers usually apply various machine learning techniques, statistical analysis tools and frameworks to visualize the data and produce meaningful results for solving a problem. Users typically perform data set pre-processing to make the data set informative and to improve data set quality before use.

III. Research Methodology

A. Literature Review

The extracted information was collected from various well-known digital libraries such as ACM, IEEE, Science Direct, Springer and Google scholar. The following research questions were addressed in the conducted literature review.

- *RQ1 - What are the unresolved challenges in the open data set lifecycle in HAR research?*
- *RQ2 - What are the best practices in the open data set lifecycle in HAR research?*
- *RQ3 – What machine learning techniques have been used for the analysis of open data sets?*

RQ1- (What are the unresolved challenges in the open data sets lifecycle in HAR research?) This concerns issues and challenges faced by the researcher during data set construction, sharing, finding and using. For example, challenges related to the data set itself such as metadata, data representation, contents and structure, annotations and documentation. Also, challenges related to the data set's context such as reusability, privacy, societal concern , usage policies and so on [96].

RQ2- (What are the best practices in the open data sets lifecycle in HAR research?) This might be practices such as what criteria and quality measures were used by the researchers while finding and using a data set? Searching best practices relating to internet search engines, social media platforms and other publicly available online websites and resources for finding and uploading data sets. The best practices used for data set documentation and annotations for real world objects throughout the development process of data collection. Also, what criteria and practices are used by HAR researchers for data set sharing in a data set repository.

RQ3 – (What machine learning techniques have been used for the analysis of open data sets?) The importance of data sets for ML cannot be overlooked and ML largely depends on the data sets and training algorithm/techniques to make decisions. This research question explores the applied algorithms and techniques on different data sets in the HAR domain. ML often relies on huge-size data sets at the center of model development and evaluation, and it depends heavily on data sources.

B. Questionnaire Survey

A questionnaire survey was chosen for exploring the current issues and challenges faced by HAR researchers. This would enable a direct response from researchers on their current approach to working with open data sets. A comprehensive questionnaire survey was conducted consisting of three main phases of planning, performing and reporting as shown in Fig. 4.

1) *Planning*

The questionnaire consisted of 40 questions as shown in appendix section - these included both open and closed questions.

Open questions: an opportunity was given to the participants to express their expertise and opinions related to a particular question.

Closed questions: the respondents were restricted to choose an answer from the given options in the form of yes/no, multiple choice, ranking and rating scale questions.

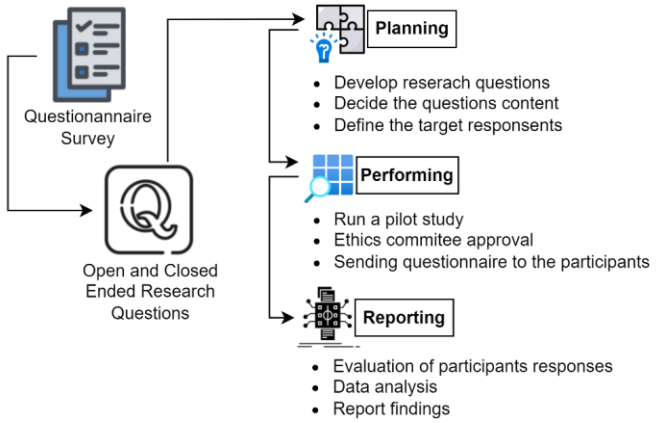


Fig. 4. Protocol for the conducted questionnaire survey

The overall aim of the survey was to elicit open issues and challenges, best practices and processes used by the researchers and their views on the quality of existing data sets in the HAR domain.

After a successful pilot study involving four HAR researchers, the survey was sent to 98 participants who had some research experience in the HAR domain. The participants were selected from academia and research institutes and included researchers, faculty, PhD students, and practitioners from industry.

2) *Performing*

Ethical approval for the survey was granted from the Faculty Research Ethics Committee at Ulster University to ensure the integrity, trustworthiness, and authenticity of the research outcomes. The questionnaire survey was administered using the [3]Jisc Online Survey Tool and the link sent to all participants through email.

3) *Reporting*

In this phase, the participants' responses were evaluated. The information was extracted from the open-ended questions through structural and thematic coding analysis [98], [99]. The findings of the survey are reported in section IV (B) below.

IV. Result and Discussion

This research aims to provide a comprehensive analysis of available data sets in the field of HAR by integrating literature review with conducted questionnaire survey. This study conducts a comprehensive analysis of the existing literature to explore the complexities associated with the issues and challenges prevalent in the field of HAR.

[3] https://www.onlinesurveys.ac.uk/

A. From Literature Review

RQ1 - What are the unresolved challenges in the open data sets lifecycle in HAR research?

Table 1 shows the issues and challenges associated with HAR open data sets that were identified through the conducted literature review. The groups and sub-groups were created from the authors's brainstorming and from the published work of Sing and Vishwakarma [56] . The authors identified challenges related to Red Green Blue (RGB) and Red Green Blue-Depth (RGB-D) data sets and then they divided into five groups based on application domain, environmental condition, occlusion, viewpoints variations and similarity of actions.

However, these challenges are only related to the human actions and external factors related to the data sets. The conducted literature review considered both external factors such as activities and actions, background condition, resources training, device and subject heterogeneity, and data set sharing and the internal factors of a data set such as annotation, noisy and imbalanced data sets, missing values, data privacy, features selection and data set size. Following are the brief description of the identified challenges.

Activities/Action Recognition: During activity recognition it can be challenging to differentiate similar and dissimilar activities such as walking and running, stairs up and down movement. Complex activities consist of more than two activities and actions for example exercise activities such as jump, extending legs and bending down. Complex activity recognition can be achieved by incorporating a model which addresses, for example, proper posture monitoring. 15.8% of papers in the conducted literature review reported activity recognition to be a challenge.

Annotations: Annotation is the labelling of data (to label an activity/action) in different formats like audio, video, text and images. Annotated data sets are important for supervised ML for pattern interpretation and accurate outcomes Named Entity Recognition (NER) and sentiment analysis require annotated training data to detect emotions and opinions [101].

Noise/ Imbalanced Data sets: Noisy and imbalanced data sets are created when participants tend to make mistakes, sensors/devices record incorrect data or insert incorrect values to attributes while collecting data. Also, the collection of additional and irrelevant information can create noise in the data, which can impact the prediction and recognition of activities and affect the overall quality of a data set [102]. 11.8% of the papers reported such issues as challenging.

Background condition: Background condition refers to the different objects of the observed environment which can hinder activity recognition – objects such as trees, rain, water and waves. [100]. Similarly, moving objects in the background can also impact activity recognition. The data sets that are collected from social media and YouTube also contain challenges arising from background conditions and moving objects. Light condition refers to brightness and darkness. Image quality is also affected by camera movement. Object shadows are also an issue when detecting human activity from images. 9.2% of papers in the literature review reported such background conditions to be a challenge when working with an open HAR data set.

Resource/Training: Training is where the recruited participants/users/subjects for the experiments are instructed on how to use devices correctly according to the defined protocol guidelines. Resources issues and challenges include the battery life of devices and reliable internet connection for all experimental wearable devices. Also, the installation of different devices producing data in various formats can be a resource intensive activity. Finally, data sets created from video recordings are large and require enough GPUs for model training. Human complexity is an issue when constructing an HAR data set because humans are highly variable and unpredictable in their movements and actions. This can make it difficult to accurately capture and classify a wide range of different activities. Additionally, factors such as lighting, camera angle, and background can also affect the quality of the data and make it difficult to generalize the data set to different environments. In the literature review, 7.9% of papers reported such challenges.

Missing Values: Missing values are situations where data has not been recorded in relation to a significant or meaningful event which should have been observed. For example, missing an activity, incomplete information or feature due to a network problem, where participants forget to record an activity or the low energy characteristics of a wearable device. This is challenging when recognizing elderly people's activities, because it is difficult to predict the activities associated with missing data, such as falling down or any other serious activity. 7.2% of papers reported this to be a challenge in HAR data set use.

Privacy: Secure storing and proper distribution of human obtained data is a crucial part of data ethics [154]. Each data set involves people who both collect data and provide data. It is important to consider that in both cases we have human beings [155]. International and cultural context should be respected where privacy concerns may have less governance in some areas, and the possibility for data manipulation is a real threat to the protection of data participants. This can be a challenge and a risk in very sensitive contexts such as personal information related to finance, medicine and biometrics [156]. Using sensors and devices for elderly assistance and for patient monitoring presents privacy issues. Installation of devices in the home for tracking could be considered as a violation of intimacy and privacy [157], [158]. 5.9% of papers reported data handling and privacy to be a challenge.

Feature Selection: Feature selection is the process of choosing the most relevant features that are essential to perform machine and deep learning and to eliminate unnecessary features. This process is useful because it will reduce noise in a data set. However, it is difficult to select the optimal features due to high correlation among features and to select the most appropriate features that contribute to prediction outcomes.

Table 1. Open issues and challenges in HAR open data sets

Group	Subgroup	In percent	Reference
Activities/Actions Recognition	• Similar actions • Dissimilar actions • Complex activities • Emotion recognition • Monitoring activities • Multiple occupants	15.8%	[103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126]
Annotations	• Data labelling problems	14.5%	[127],[103], [105],[128], [129], [110], [112], [113], [114], [130], [115], [116], [131], [132], [117], [133], [134], [120], [135], [122], [136], [125]
Noise/Imbalanced data set	• Data interpolation • Temporal relationship (Video frames) • Irrelevant information	11.8%	[127], [137], [138], [128], [106], [107], [139], [115], [115], [140], [132], [133], [119], [120], [141], [142], [143], [123]
Background Condition	• Other objects • Light condition • Image quality • Movement/posture pattern • Shadow • Dynamic background	9.2%	[103], [104], [127], [137], [107], [109], [112], [139], [113], [140], [141], [144], [145], [126]
Resource/Training	• Subject/user training • Battery condition of device • Internet connection • Device variety • Sensor orientation • Camera variation	7.9%	[137], [129], [109], [146], [147], [148], [139], [116], [149], [120], [125], [126]
Missing values	• Lack of efficient data representation	7.2%	[150], [103], [104], [105], [106], [129], [146], [147], [148], [134], [136]
Privacy	• Data sensitivity • Ethics concern	5.9%	[108], [151], [112], [114], [140], [117], [152], [119], [125]
Feature Selection	• Auto extraction of important features • Selection of active features • Removal of irrelevant features • Appropriate preprocessing techniques	5.9%	[138], [147], [151], [112], [113], [116], [118], [120], [145]
Heterogeneity (Device/Data/subject	• Various data format • Various devices installed • Subject heterogeneity	5.9%	[138], [108], [147], [148], [113],[153], [117], [134], [142]
Data set size	• Less sample of data	4.6%	[151], [113], [153], [118], [144], [122], [145]
Other Issues and Challenges	• Lack of sharing data • No data collection guidelines • Subject observation and monitoring • Subject annoying to wear device • Support new participants in data collection • Data set benchmarking	11.2%	[132], [149], [120], [121], [144], [122], [126]

Other issues related to feature selection is how to choose the appropriate statistical and ML approach and technique. 5.9% of papers reported this process as a challenge.

Heterogeneity (Devices/Data/Subjects): The use of sensors, RFID tags and wearable devices for collecting activity data in HAR research area is growing. The variations in sensors and the frequencies at which data has been collected creates data in different formats. Subjects' heterogeneity due to different lifestyles and culture and different mechanisms for their recruitment is also a potential challenge in HAR domain. 5.9% of papers reported such issues.

Data Set Size: When the size of activity data is small, the HAR model developed through training leads to anomalies and random noise. Therefore, it negatively affects the model's generalization capability. Furthermore, limited data means that HAR models are unable to model new data and to generalize to unseen (new) data leading to low model performance.

The activities/actions recognition and annotations are the highest mentioned groups of the identified issues and challenges. Similarly, noise/imbalanced data sets were mentioned by the researchers up to 11.8% and other issues and challenges to 11.2%. The remaining issues and challenges are 9.2% and below. The data set issues and challenge as described above are important to address and to create data sets which are meaningful for the HAR research community.

Other Issues and Challenges: At a fundamental level, the lack of data sharing amongst HAR researchers is itself an issue. This may be due to constraints on data set sharing from within the researcher's organization, the confidential nature of data, or a data set having low quality such that sharing and reuse is not possible. This might arise when open data sets are collected without proper data collection guidelines and protocols. Poor quality data might arise because it is difficult to fully control the participants during data collection. Sometimes, it might be annoying for the subject to wear a device for recording activity or they might have a privacy or other social concern. Similarly, introducing a new participant into an experiment during data collection or an existing participant leaving is a potential issue for consistent data collection. A benchmark data set is a set of data that is widely used in a particular field to evaluate the performance of different models or algorithms. These data sets are often used as a standard for comparing the performance of different approaches and for measuring progress in the field. Benchmark data sets are taken from several various sources, each with a distinct composition of copyright owners and agreements for their usage in training and evaluation in ML models [159]. Open data set benchmarking is good for validating data sets and for evaluation of a model or an experiment against internal or external standards related to HAR data sets.

RQ2 - What best practices are used by researchers when constructing, sharing, finding, and using data sets in HAR research?

Researchers developed their own practices according to the needs of their experiments. However, most of the researchers used data set pre-processing techniques in the construction and using phase of a data set. They extracted the prominent features and improved accuracy through using different ML classifiers. Some researchers replaced missing values from a data set with a mean value.

Researchers also observed that not all extracted features are good for classification because some of them have a negative impact on classification performance. To reduce overfitting in a data set and to make a data set more generalisable, data augmentation was applied by some researchers [150], [103], [105], [138], [128].

Researchers are working to improve metrics for data set benchmarking by including the full empirical evaluations, including negative results during evaluation and full sharing of additional experimental details [160]. For data labelling, researchers used data transformation and segmentation algorithms to distribute the data into different length windows and then human experts assign the labels by marking a time stamp for each activity[1], [161]. To overcome the problems of data scarcity and also privacy concerns, researchers are developing synthetic data sets which make it easy to share data and increase model robustness. Synthetic data sets can be generated by learning the statistical properties of actual data sets [162], [163]. Detailed best practices are shown in Fig. 24 in section V.

RQ3 – What machine learning techniques have been used for the analysis of open data sets?

In addition to investigating issues and challenges, the conducted literature review identified the range of ML techniques which have been used in HAR analysis as it has been applied to open data sets. Currently, researchers are using ML techniques to identify and interpret activities in HAR in various domains such as sports, healthcare, and falls of elderly people. The limitation of the conventional approaches such as basis transform coding, statistics of raw signals and symbolic representation was the shallow learning that needed feature engineering from data, and it was largely dependent on human knowledge of the specific domain [164]**.** The conventional approaches of heuristic nature and human understanding make it difficult to capture complex actions having micro activities. Higher-level activities include more semantic and contextual information, making it harder to discern their hierarchical structure. Existing approaches often overlook signal correlation, which limits their ability to provide satisfactory outcomes. Hence, the popularity of ML due to the successful learning from complex movements/actions using Convolutional Neural Networks (CNN) [165].

For large scale data sets, deeply learned feature methods is suitable for HAR domain because deep learned features, also known as deep features, are representations of data that are learned by a deep neural network. Deep learning-based solutions consists of feature extraction and classification in video-based data sets. In addition to the development of high computational power and growing volume of the video-based data sets, a deep learning-based solution is useful for real-life

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317645

applications [100]. Feature extraction is an important aspect in HAR to identify the most relevant and significant features from the data to reduce errors in classification and computational complexity. The efficient performance of HAR activity recognition depends on suitable feature representation. Achieving maximum performance depends on the selection of suitable techniques for features extraction [120]. Researchers already work and apply ML to automatically extract features from raw data generated through sensors. ML techniques and classifiers that are used in HAR for automatic feature extraction, pre-processing, and classification of data sets are Support Vector Machine (SVM), k Nearest Neighbour (KNN), Decision Tree (DT), CNN, long short-term memory (LSTM) and Random Forest (RF) [24], [166].

Researchers are applying new trends of ML such as transfer learning which is the ability to transfer knowledge from one model to another to train it with a minimum amount of data and to reduce computational complexity and effort [167]. Similarly, the other emerging ML method is active learning with an objective to reduce learning complexity and computational cost. It seeks to choose the relevant information from unlabelled data and ask the annotator for labelling information. The main advantages of active ML in HAR are to reduce annotation efforts and to increase forecasting accuracy [168]. As shown in Fig. 5, ML techniques were used in the literature review papers for prediction, classification, feature extraction, and data set labelling purposes.

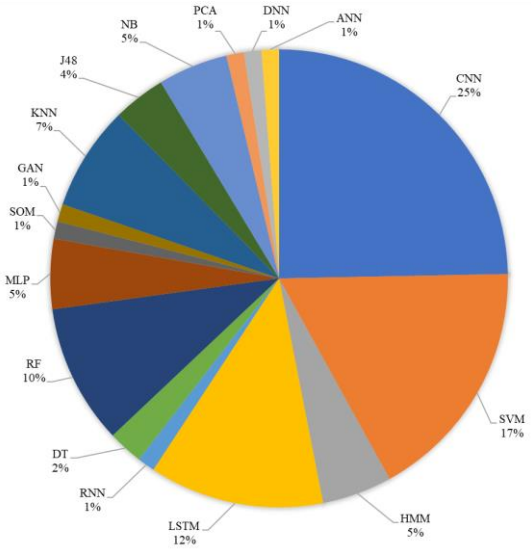


Fig. 5. Weightage machine learning techniques explored in SLR applied on open data sets Convolutional neural network (CNN), support vector machine (SVM), hidden Markov model (HMM), long short-term memory (LSTM), recurrent neural network (RNN), decision tree (DT), random forest (RF), multi-layer perceptron (MLP), self-organizing map(SOM), generative adversarial network (GAN), k-nearest neighbours (KNN), Naive bayes (NB), principal component analysis (PCA), deep neural network (DNN), artificial neural network (ANN).

B. Survey results

The purpose of the survey of HAR researchers was to elicit their views on the use of open data sets. The open data set lifecycle (Fig. 3) was used as a conceptual framework for designing this survey. The survey was sent to 98 HAR researchers and 32 (32%) responses were received. The experience of the participants with respect to each phase of data set lifecycle is shown in Fig. 6 and Fig. 7 illustrates the range of participant occupations.

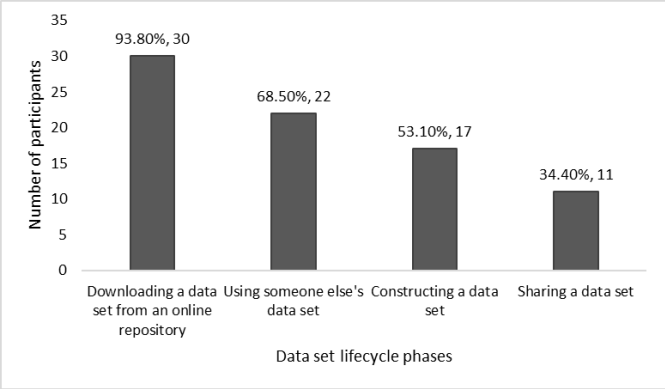


Fig. 6. Experience of the participant with respect to each phase of data set lifecycle

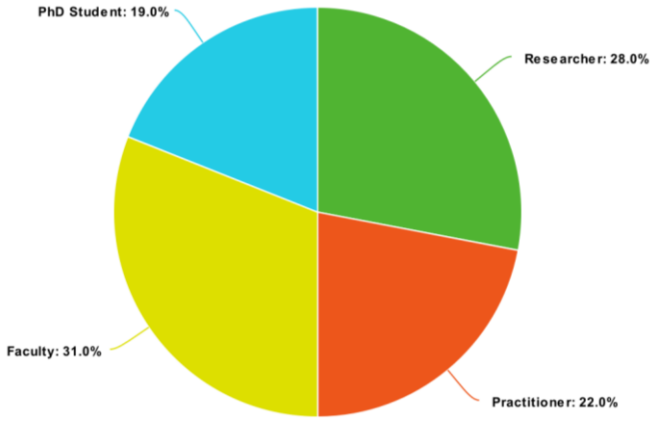


Fig. 7. Participant occupation

Similarly, Fig. 8 shows the participants experience by years and Fig. 9 presents the location of the participants.

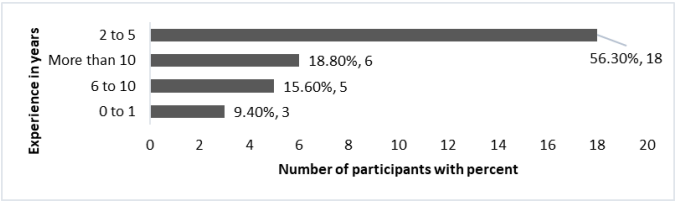


Fig. 8. Participants experience in years

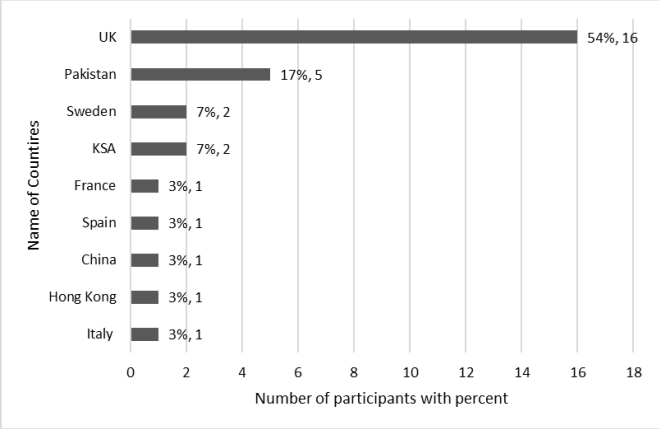


Fig. 9: location of the participants

1) *Construction*

This is the most significant phase of the open data set lifecycle as it involves the identification of what type of data is needed for analysis to achieve the research objectives. Further, what type of data collection protocol will be used to collect data systematically and to build the overall guidelines for data collection. A total of 53.1% participants who took part had experience of constructing a data set.

From an initial analysis of the question "What is the main issue or challenge you have faced when generating a new data set in human activity recognition?", the major issues and challenges identified are shown in Table 2.

Table. 2. Issues and challenges in open data sets construction phase

Main issues/challenges	**Description**	**Frequency of participants**
Data set annotation	Generating sufficiently precise annotation of the data. This is a very time-consuming task if labelling the data set after collection.	4
Time synchronization with multiple sensors	Synchronization of timings from multiple sensors and heterogeneous sources	2
Ethical approval	Obtaining Ethical approval from the host organization	1
Data set size	Data set size is important in determining the performance of a machine learning model. The consequence of using a test set with a limited sample size might lead to a large variation as well as a high error rate. Further, a small set of data set tend leads to overfitting and creates incorrect results.	1
Privacy	Collecting video data involves enough with sufficient contextual information can leads to privacy issues.	1
Training of users and actors	Providing proper training to the users/actors who help in generating new data by performing different activities.	1
Missing values	Missing data due to sensor malfunctioning or poor user engagement when collecting data with multiple users in free-living. The missing data could bias the final results because missing data adds ambiguity to the data.	1
Finding volunteers	Finding sufficient volunteers for collecting the data set while ensuring diversity of participants with respect to privacy and social concerns.	1
Lack of funding	Lack of funding specifically for data collection process.	1

Data sets sharing after construction, the responses were taken from the participants regarding sharing restrictions of the constructed data sets from HAR research community are shown in Fig. 10.

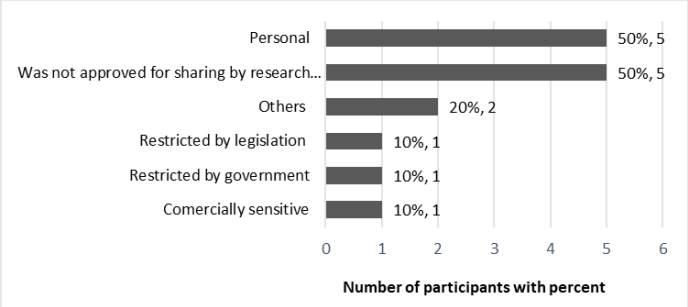


Fig. 10: Constructed data sets sharing constraints.

Finally, participants were asked what is the main piece of advice they would give to another researcher when generating a new data set in human activity recognition? From the 32 survey participants, 18 responses were received as summarised in Table 3, with similar advice paraphrased and grouped as shown.

Table. 3. The main piece of advice participants gives to another researcher when generating a new data set in human activity recognition.

Participants responses	**Frequency**
Be thorough with annotation of the dataset.	3
Keeping participants anonymity	2
Give information about the process of data filtering and cleaning.	2
Plan well the formatting and synchronisation of data for timestamps. Using the same format facilitates cross-validation of approaches	2
Be prepared to spend quite a bit of time trying to make it clear what is in the data and documenting it, if you want other research groups to be able to use it.	2
Carefully clean the data prior to sharing	1
Provide context and ground truths if possible.	1
What kinds of activities were going on in the time between the collection and the sensor placement?	1
Ensure ethical approval before data collection	1
Include or reference supporting papers with result	1

discussions	
Data should remain open if shared online, links must not be deactivated.	1
It is important to design a protocol for generating new dataset in HAR	1

Responses related to the normal practice for sharing constructed data sets in which the statement "always seeks permission to share" up to 50% and "shared only if required by the research sponsor" equal to 33%. This indicates that data sets are often not open sourced but used only for their original research purpose. When asked about the normal practice for data set sharing, the responses are shown in Fig. 11.

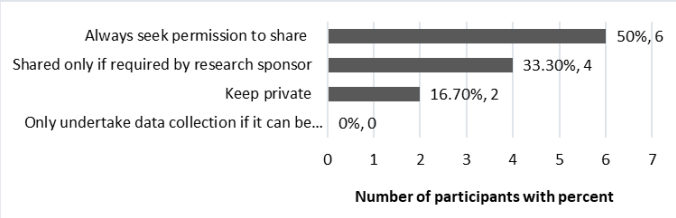


Fig. 11. Normal practice for constructing experimental data sets

2) *Sharing*

This phase of the data set life cycle is concerned with making the data set available for researchers and users to use it for solving their HAR problems. Data set owners must grant access to their shared data sets for research purposes. Researchers and data set owners store their data sets in a range of storage repositories. Different categories of data set repository are used based on the nature of the data and the researcher's licensing requirements. The main categories of open data set repository are discussed below.

Institutional Data Set Repositories: Institutional-based repositories are the collections of data sets that are created and maintained by a specific institution, such as a university or research organization. These repositories may contain data from a wide range of research projects and disciplines and may be used to share data across the institution or with external researchers. They may also have more robust data management and preservation capabilities, as they are typically maintained by dedicated staff. They distribute and manage data sets, research outcomes such as data analysis, source code, tools and frameworks that are developed by the research community and postgraduate students. Examples include [4]DataShare from the University of Edinburgh, [5]University of Cambridge Data repository and [6]UCL Research Data Repository.

Hosted Repositories: Hosted repositories refer to datasets that are stored and maintained on a remote server and can be accessed by the public over the internet. These data sets can be easily discovered and downloaded by anyone with an internet connection. UCI Machine Learning Repository, Figshare, DataPort, Zonedo, Kaggle, etc. also engage in the management and distribution of open data sets from educational organizations and industries.

Government Data Set Repositories: These repositories are managed by government and store sensitive data for health care management, surveillance, and administrative purposes. These data sets are not shared commonly due to privacy concerns and intellectual property protection of the user data [169]. Examples of government data set repositories are "[7]data.europa.eu" and "[8]data.gov.uk".

Specific Domain Data Set Repositories: These repositories consist of data sets and metadata from a specific domain such as health care, sports, engineering, and social sciences. The main advantages of these data sets are related to specific research domains. Examples are geriatric [9]healthcare and [10]physiotherapy.

Project-Based Repositories: Project-based repositories are collections of data sets that are created and maintained by a specific research project or group. These repositories are typically focused on a specific topic or area of research and may contain data that is collected and analysed by the project team. For example, the centre for data and visualization sciences at Duke university [170], Managing research data by the University of Bristol [171] and machine learning and AI data set managing by Carnegie Mellon university [172].

The percentage of participants who took part in data set sharing was 46%. This clearly shows that most of the researchers are not taking part in data set sharing.

Further, the participants were also questioned whether the sharing of data was restricted because it contained information about an organisation or participants. On a Likert scale that included strongly agree, agree, neutral, disagree, and strongly disagree, participants' opinions were measured. Overall, 13% of respondents were neutral, 46% agreed and strongly agreed with the statement, and the remaining respondents disagreed.

Furthermore, the participants were asked the question "Sharing the data set was not possible because of problems with the data?" In response, most of the participants were agreed up to 93% and consent the problem with data is the main barriers of data sets sharing.

Moreover, participants were asked to select reasons for not sharing their data set. The top reasons were the missing values, errors during measurement, and small sample size of the data. Also, 28% participants were agreed to the large size, the data is not in presentable form, and the lack of time and resources for data sets sharing are the main reasons. The detail

[4] https://www.ed.ac.uk/information-services/research-support/research-data-service/after/data-repository
[5] https://www.data.cam.ac.uk/data-repository
[6] https://www.ucl.ac.uk/library/open-science-research-support/research-data-management/ucl-research-data-repository
[7] https://data.europa.eu/en
[8] https://www.data.gov.uk/
[9] https://mira.mcmaster.ca/research/open-access-datasets-from-aging-studies
[10] https://pedro.org.au/

responses are mentioned in Table 4 with participants frequency and percentage.

Table. 4. why the sharing of data was limited.

Asked questions related to data set sharing	Participants responses in numbers	Participants responses in percentage
The sharing of data was limited because it contained identifying information about an organization or participants.	-Strongly disagree: 0 -Disagree:6 -Neutral:2 -Agree:6 -Strongly agree: 1	-Strongly disagree: 0% -Disagree: 40% -Neutral: 13.3% -Agree: 40% -Strongly agree: 6.7%
Sharing the data set was not possible because of problems with the data?	-Yes: 1 -No: 14	-Yes: 6.7% -No: 93.3%
Please review the following list of reasons and select all that apply for your data set:	-Missing values: 1 -Outliers: 0 -Measurement errors: 1 -Overfitting is harder to avoid: 0 -Sample size too small: 1	-Missing values: 100% -Outliers: 0% -Measurement errors: 100% -Overfitting is harder to avoid: 0% -Sample size too small: 100%
Sharing the data set was not possible because it was too large?	-Yes: 4 -No: 10	-Yes: 28.6% -No: 71.4%
Please select all that apply	-Too large for selected - repository: 1 -Raw data not in a presentable form: 1 -Lack of time and resources to share large data set: 4 -Lack of experience in data management: 0	-Too large for selected repository: 25% -Raw data not in a presentable form: 25% -Lack of time and resources to share large data set: 100% -Lack of experience in data management: 0%
I did not share the data set because I was unsure of the best approach?	Yes: 2 No: 13	Yes: 13.3% No: 86.7%

The other potential reasons that were asked from the participants for not sharing data sets are shown in Fig. 12.

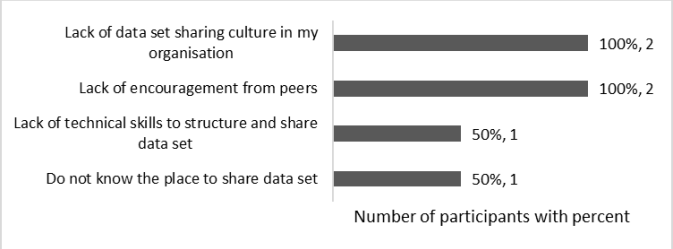


Fig. 12. Other reasons for not sharing data sets.

When asked "What is the main issue or challenge you have faced during data set sharing?", participants responded as shown in Table 5.

Table. 5. Issues and challenges in open data sets sharing phase.

Main issues/challenges	Description	Frequency of participants
Privacy	Even when participant anonymity is a challenge, the collection and exchange of massive quantities of data may expose individuals to significant privacy violations.	4
License	License awareness and how to select the most appropriate license for data sharing	2
Proper data collection	Expensive to do a proper data collection by buying good quality resources, recruiting, training and monitoring of the participants	1
Getting data visibility	Difficult to interpret and to get the visibility of the data from the existing data sets.	1
Ethics approval	Obtaining ethics approval from the organization before data sharing	1
Data format	There is no standardized data format for data sharing	1
Data presentation	Presenting data in an understandable and easy way	1
Resolving data set queries	Resolving data set queries by checking access to the server, adding new data records and version control of the data set.	1
Lack of time	Need time to clean and share data	1
Lack of resources	Need resources for data sharing and to perform maintenance or version control of the data set.	1
Data set size	Low sample of data leads to bias in the results and effect's reliability due to higher variability in the	1

	data.	
Data quality	Hard to maintain standard data format and quality before sharing by removing data bias and to deal with missing data.	1
Platform for sharing	No proper platform for researchers and practitioners to share data set, collaborate, benchmark, validate and improve data set.	1

It motivates us to find insights from the human behaviour.	1
Having standard repository for storage and sharing	1
To see how different researchers manage and analyse the data.	1
Collecting a dataset requires is a huge effort. Making it available to other researchers maximize the return	1
The desire to benefit humankind by improving data analysis and algorithms.	

Responses related to the license used for data set sharing are shown in Fig. 13. Most of the participants did not use any license while sharing their data sets.

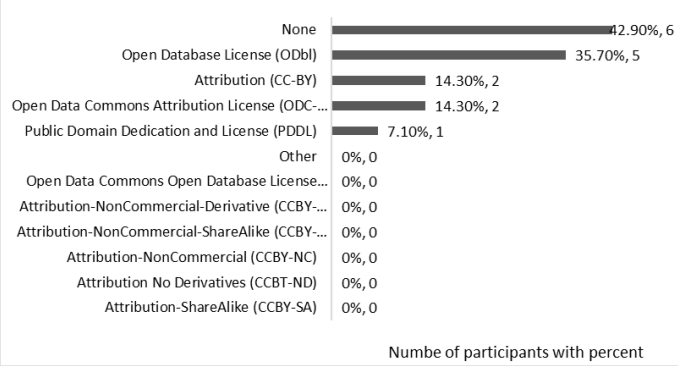


Fig. 13. License used by the researchers while sharing data sets.

Participants were asked if they found the process of depositing and sharing a data set time consuming. As shown in Fig. 14, most of the participants strongly agreed that this was the case.

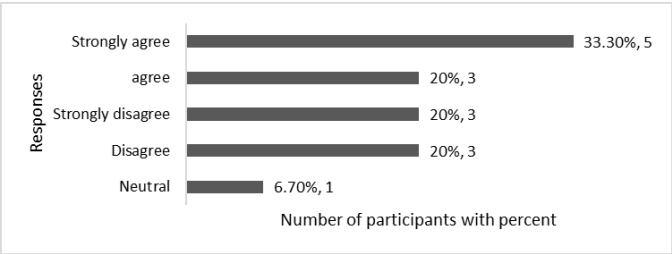


Fig. 14. The participants' response regarding the process of depositing and sharing data sets are time consuming.

Participants were also asked "what motivates you to share your data set with the human activity recognition research community?" Details of participants' responses with frequency are given in Table 6.

Table. 6. motivating factors for sharing your data set.

Participants responses	**Frequency**
Further citations.	3
To allow for study replication, transparency of method and verification of results and contribution.	2
Serve the scientific community and industry.	2
Open research and funding grants.	2
To get more insights from the data and extract more useful information.	1

The final question in relation to data set sharing is "what is the main piece of advice you would give to another researcher when sharing a new data set in human activity recognition?". Table 7 shows the exact response of participants with frequency.

Table. 7. main piece of advice to another researcher when sharing a new data set in HAR.

Participants responses	**Frequency**
Ensure data set is shared on a repository that has longevity/permanence.	2
Be prepared to spend quite a bit of time making the dataset understandable to others.	2
To be aware of the size and privacy of data	2
Carefully clean the data	1
Publish your research to increase visibility of the dataset	1
Establishing clear and simple dataset guidance	1
Explore for datasets relevant to your research. This will facilitate future research.	1
Share in a standard repository	1
Ensure you capture the methods used to collect the dataset at the time of recording the data.	1
Research must be open.	1

3) *Finding*

This phase of the open data set life cycle is related to searching and finding a data set for a specific problem domain. Responses received from the participants related to the questions such as "Have you searched for and downloaded an open data set from an open data online repository for experimental / research purposes?" A total of 96.9% of the participants obtained an open data set for the purposes of study or experimentation by downloading it from an online open data repository. The next question related to the criteria that are used by the researcher when searching a data set – the results are as shown in Fig. 15.

Participants were asked about the amount of pre-processing (none, a little, some, a lot) required after downloading a data set. 29% participants indicated a lot, 38.7% some, 32% a little and 0% with none. For the respondents, every data set required pre-processing after downloading to make it usable and informative.

As with previous stages of the open data set lifecycle, participants were asked "What is the main issue or challenge you have faced when trying to find a suitable data set?" The responses and their frequency are shown in Table 8.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317645

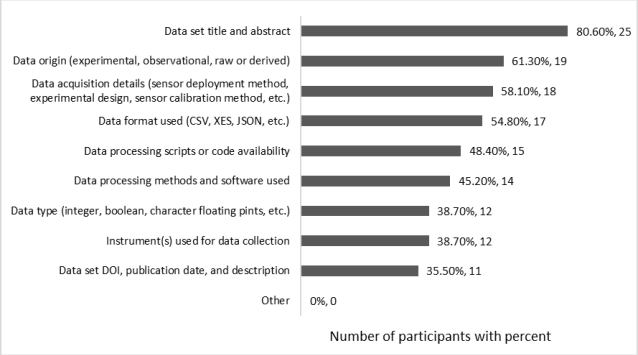


Fig. 15. Searching criteria for data set searching.

Table. 8: Issues and challenges in open data sets Finding phase.

Main issue/challenge	Description	Frequency of participants
Metadata	Lack of data set description such as activities observed, devices, sensors, participants, etc.	5
Data set completeness	Missing data and missing annotations	5
Specific domain data set	Data sets are only limited to a specific domain	4
Access restriction	Data set access and use restriction from the data set owner	3
Data set size	Insufficient data for meaningful use	2
Data set format and structure	Lack of standard data format and structure	2
Repository search optimization	Hard to search relevant data set in the repository	2
Data set authenticity	Lack of trust in data collection methods and protocol used	1
Time and effort	Time and effort needed to search a data set and understand it	1

Similarly, participants were asked about their motivations for using an existing data set instead of creating the new one. Responses are shown in Table 9.

Table. 9. Motivation factors for using the existing data sets.

Motivating factors	Description	Frequency
Save time/effort/resource	Cleansing data necessitates more time and resources before it can be utilised for experimental purposes. Already cleaned data will save more time/effort/resource.	14
Benchmarking	Data set able to be used as a benchmark	3
Meta data	Availability of a clear data description	2
Problem domain	Provide a description of the data together with the applicable problem domain.	1
Use already published data set	Data set already published in a research platform	1
Authentic data	Trusted data	1
Reproducibility	Ease of reproducibility of results and application of novel algorithms	1
Research topic	The topic and the challenges of the research are relevant?	1
Challenge	Creating new dataset is challenging! Existing public datasets are easy to use	1
Data set size	Data set provides enough amount of sufficient data for meaningful analysis.	1

The participants responses regarding the data sets repository or directory used when searching are shown in Table 10.

Table. 10. Data sets repository/directory used for finding a data set.

Data set	Frequency
UCI	11
Google search	9
Kaggle	8
GitHub	5
Publisher website	5
Data search	3
IEEE Dataport	3
Direct contact	1
Sparkbankan	1
Data portal	1
Open ML	1
Planet lab	1
Zonedo	1

Finally, the main piece of advice from the participants for other researchers while searching for a data set in HAR are shown in Table 11.

Table. 11. Main piece of advice for other researchers while searching a data set.

Main piece of advice	Description	Frequency
Documentation	Look for supported documentation of a data set such as published papers, meta data, data collection protocol, sensors used, participants, format etc.	7
Data set check	Before using a data set, check for certain feature such as timestamps are being in order, missing values, no mismatch of columns, noise in data, data normalization etc.	4

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3317645

Search terms and keywords	Identify the most relevant keywords and search terms for searching a data set according to the domain problem	3
Goal and objective	The main advice for searching HAR dataset is to first identify the goals you want achieve and also to arrange resources for processing the data in Advance.	3
Generic data set	Make datasets generic for a specific filed field of research so that several research questions can be answered using the datasets	3
Data quality	Look for data set completeness such as meta data and description	2
Libraries identification	Exploring available libraries for data set cleansing and pre-processing	1
Check data set license	Check licensing to ensure permission to use is within remit of experiment.	1
Available resources	Try every available resource (repository?) for searching and finding a data set	1
Benchmarked data set	Search for an already used and benchmarked data set	1
Contact	To contact related parties that may provide the data set	1

4) *Using*

This is the final stage of the open data set lifecycle. After finding a data set, researchers typically apply different machine learning and statistical analysis tools and frameworks to visualize the data and produce meaningful results. Researchers and practitioners use open data sets to solve their domain problems and make a decision for future solutions.

Data set pre-processing is a key step to make the data set informative and to improve data set quality before using. It is a data extracting procedure that includes converting raw data into a comprehensible and informative format. Data sets collected within the natural environment are often inadequate and inconsistent due to missing values and activities/actions labelling that leads to errors [173]. Data pre-processing is performed by the researchers to resolve challenges such as missing values, data annotations, handling errors and outliers, codes and naming discrepancies [174]. The major steps involved while performing data set pre-processing are: 1) data cleaning, 2) handling null values, 3) standardization, 4) handling categorical values, and 5) feature scaling and dependencies [175].

In the survey, the question was asked "Have you used and evaluated someone else's open data set for experimental / research purposes?" 78% of participants responded yes.

Two questions related to metadata description – was the metadata description easy to understand and was the metadata description accurate? Responses are shown in Fig. 16 and 17 respectively.

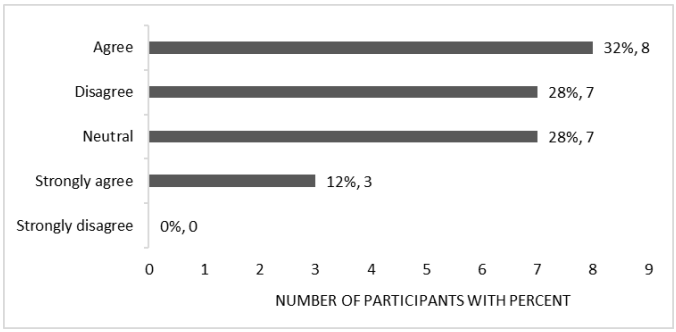


Fig. 16. Metadata description make it easy to understand a data set

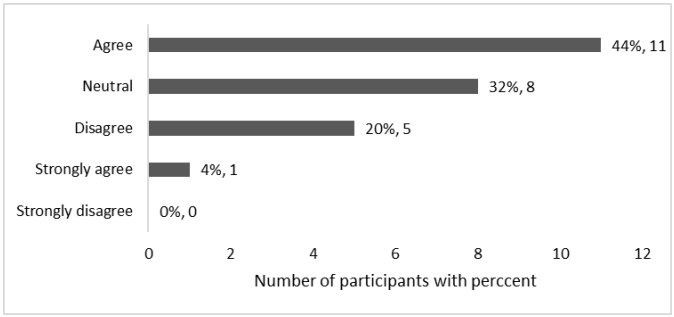


Fig. 17: Metadata description accuracy

Respondents were asked if they encountered a data set update issue after the initial sharing, with response as shown in Fig. 18.

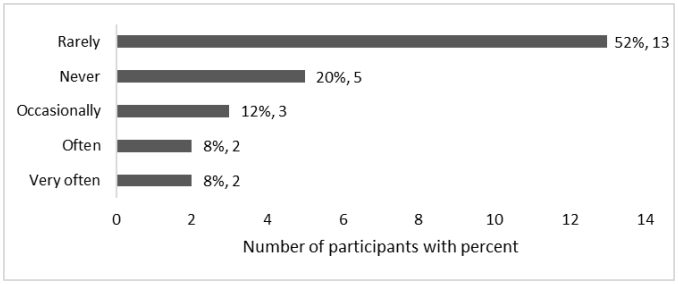


Fig. 18. Agreement of data set update issue after the initial sharing

The main issues and challenges identified by respondents while using someone else's data set are shown in Table 12.

Table. 12. Issues and challenges in open data sets using phase.

Main issue/challenge	**Description**	**Frequency of participants**
Documentation	A complete description of data set such as data collection protocol, devices used, guidelines for using the data set, etc.	6
Annotation	Understanding how data was captured, what the data shows (annotation/labelling	5

	of activity)	
Format	Various formats of data sets	5
Missing values	Missing values and empty columns of a data set	4
Data noise and imbalanced data	Missing columns, irrelevant information, anomalies in a data set	3
Inaccurate assumption	The An inaccurate assumption from the researcher about the experiment relevance of the data set??.	2
Sensor	No information about sensors and data collection with limited sensors	1
Data restriction	Limited/restricted amount of data. For example, 7 days' work or a few months' work or lab settings only, etc.	1
Trust	Lack of trust and proper ground truth about a data set	1

Finally, the main piece of advice from the participants to other researchers while using someone else's data set in HAR are illustrated in Table 13.

Table. 13. Main piece of advice while using someone else's data set.

Main piece of advice	Description	Frequency
Data set analysis	Conduct a Complete analysis of data set and meta data for understanding data set	5
Time and effort	Provide more Set aside sufficient time and effort for understanding the data set and matching it to the requirements of problem domain	4
Documentation	Read supporting documentation provided with data set and also published papers on a data set	3
Data set pre-processing	Pre-process and cleaning of the data set to make it usable and informative	2
Issues reporting	During data set analysis, if found any issues or problems are found, report it them to the data set owner/research community	1
Trust	Be careful. If you don't trust the data/labels, then everything built on top of it cannot be trusted.	1
Open source	Priority should be to find open-source dataset first.	1
Validation	Validate your research experiment on multiple datasets and in different scenarios to increase achieve generalization.	1

Other relevant questions that were asked "In your view, what factors improve data set quality (one per line)?" The responses associated with the important factors of a data set quality are shown in Table 14.

Table 14: Important factors for data set quality

Improvement factors	Description	Frequency
Annotation	Annotating a data set involves adding metadata, such as descriptions or tags. This can be done in a variety of ways, for as by transcribing audio recordings, labelling certain elements in videos, or providing textual descriptions to images.	6
Standard format and structure	The term "standard format and structure data set" is used to describe a collection of information that has been organised in a way that makes it simple to read, analyse, and disseminate.	6
Data cleaning and pre-processing	Preparing a dataset for analysis necessitates cleaning and pre-processing it to remove or rectify errors, inconsistencies, and missing data. Data cleansing include eliminating duplicates, fixing typos, completing blanks, and standardising file formats.	4
Data set noise	The term "dataset noise" is used to describe the occurrence of meaningless, contradictory, or incorrect information inside a dataset. This may arise as a result of typos, bad data gathering, or the addition of extraneous information. The quality of any studies or conclusions made from a dataset might be diminished by the presence of noise in the dataset.	4
Documentation	The term "dataset documentation" is used to describe the information and documentation supplied about a dataset, such as the dataset's goal, origin, structure, format, relevant metadata, annotations, etc.	3
Experimental setup	An experimental setup is the arrangement of equipment, materials, and circumstances utilised in an experiment. It covers experiment design, data collection, and quality control. The experiment's validity and reliability depend on its setup.	3
Typos errors and duplication	A dataset is a collection of data that is structured and organised in a certain way.	2

	Typographical errors are mistakes produced while typing data, such as misspellings or grammatical errors. The occurrence of duplicate data inside a dataset, where the same information is repeated many times, is referred to as duplication.	
Activity description	The dataset often contains sensor data obtained from wearable devices, such as accelerometer and gyroscope measurements, as well as annotations or labels indicating the activity being done at a given moment. The possible actions include walking, running, leaping, sitting, and standing, among others. The dataset may additionally contain extra information, such as participant demographics and environmental characteristics.	2
Quality metrics	A dataset's quality metrics are a set of standards by which its reliability and accuracy may be judged. Completeness, correctness, consistency, timeliness, and relevance are some examples of metrics that may be used. The authenticity and trustworthiness of the sources and the suitability of the data for its intended purpose are other crucial considerations when assessing the quality of a dataset. Quality metrics for datasets may also incorporate indicators of data governance, such as data lineage, data provenance, and data security.	1
Accuracy	Dataset accuracy refers to the degree to which the data inside a dataset are correct or reliable. It assesses how accurately the dataset's data matches the actual values or attributes of the items or individuals it represents. A dataset with high accuracy has few mistakes or inaccuracies, whereas a dataset with low accuracy contains a significant number of errors or inaccuracies.	1
Meta data	Dataset metadata includes the title, author, date generated, format, and other attributes about a dataset. The dataset or its metadata file normally contains this information. Metadata aids data discovery, reuse, and analysis by explaining the data's context and purpose.	1
Feature engineering	Making new features or modifying existing features in a dataset is called "feature engineering," and it may be used to enhance a machine learning model's ability to accurately represent the data. Dimensionality reduction, feature scaling, and feature extraction are all examples of such methods. The purpose of feature engineering is to enhance the performance of a machine learning model by identifying and utilising the most informative and pertinent features within the available data.	1
Missing data	A dataset with missing data is a collection of information in which certain values are absent or have not been captured. This can arise for a variety of reasons, including mistakes in data collection, data input, or data source constraints. Missing data may have a substantial influence on the accuracy and use of a dataset, since it might result in findings that are skewed or insufficient.	1
Data set size	Dataset size is the number of observations or records. It can also mean a dataset's variables or properties. A dataset might comprise hundreds or millions of observations. Data processing, analysis, and tools depend on a dataset's size.	1

Participants were asked “Have you ever pre-registered an experiment in any the following online repository domains?” 59% indicated that they never used any online repository, 21% had pre-registered on project/programme-based repository and 18.8% indicated an institutional data set repository as shown in Fig. 19.

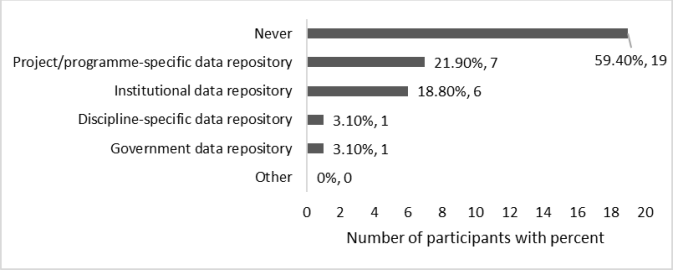


Fig. 19. Data sets repository used for online registration.

The question was asked “how important to you is the replication of a human activity recognition experiment using an open data set?”. Responses are as shown in Fig. 20.

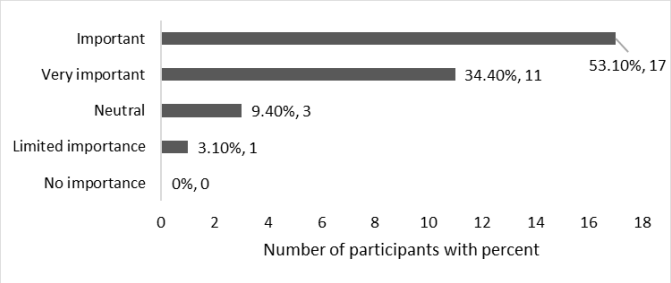


Fig. 20. Importance of data set replication

Finally, the participants were asked about the importance of data set benchmarking with responses as shown in Fig. 21.

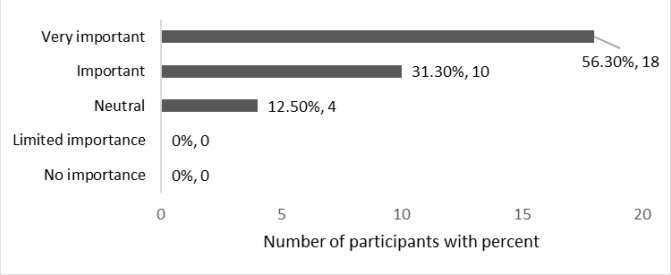


Fig. 21: Importance of benchmarking

V. Analysis and Comparison of Literature Review and Questionnaire Survey

This section presents an analysis and comparison of the outcomes of both literature review and questionnaire survey. Section V-(A) assesses the open issues and challenges identified Section V-(B) outlines the best practices identified from the literature review and mentioned by the survey participants.

A. Issues and challenges

Fig. 22 shows the main issues and challenges identified from the literature review and survey. Some issues and challenges intersect the questionnaire survey and literature review are:

- missing values
- privacy
- annotations
- data set size
- resource and training
- data noise and imbalanced data sets
- others (data set documentation, volunteer participants recruitment, lack of sharing data sets, subject annoying to wear device and lack of protocol for data collection).

Some issues and challenges that are identified only in the literature review are:

- Background condition,
- Activities/actions recognition,
- Device/data/subjects' heterogeneity
- Feature selection.

Those from the questionnaire survey only are:

- Sensor's information
- Data restriction from organization and research groups
- Data set documentation
- Standard data format
- Inaccurate assumption during data set construction
- Metadata
- Trust on the data set authenticity
- Ethical approval from the hosted organizations
- Sharing platform
- Data quality

As a result, all the issues and challenges that are obtained as an intersection form both literature review and questionnaire survey or separated from each needs focus to address by proposing approaches, frameworks and tools to overcomes.

Getting over the issues and challenges and coming up with solutions for the future might be beneficial in the following areas.

- *To improve data set structure, quality and access improvement:* new knowledge and approaches can be used to improve the structure of the data, quality of the data by removing irrelevant information and to increase data access due to open-source technology. Open-source technology improves availability and transparency of the data [176], [177].
- *Good quality open data sets can improve personal fitness:* good quality open datasets may contain data on levels of physical activity, diet, sleep habits, and other health-related information. This data may be used to develop individualised workout routines, establish and track objectives, and measure progress over time. By leveraging technology and data sets with personal fitness can improve health related quality of life [178], [179].

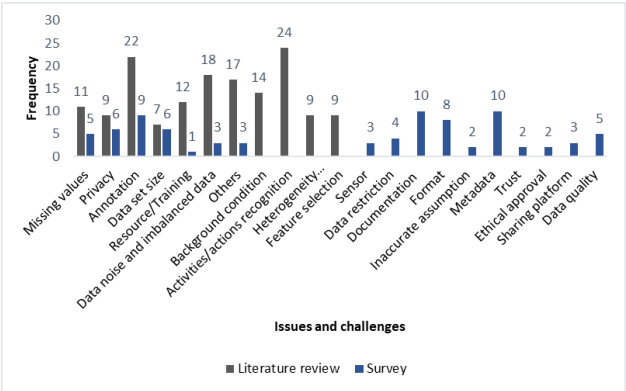


Fig. 22. Open data sets issues and challenges from both literature review and questionnaire survey in HAR domain

- *Good quality open datasets can enhance decision making by providing a broad and diverse range of information that can be used to inform decisions:* decision making can be improved and optimize by the access of accurate and up to date information and data. This will make easy to allow decision faster. HAHAR and medicine related data sets assist health care to get more informed decision [176],[180], [181].

the best approaches from the literature review are mostly related to the using phase of the data set lifecycle due to their use in experiments and subsequent pre-processing and improvement in order to fit the reported experiments.

The best practices identified from the literature review on the right side of Fig. 23 are data transformation and segmentation to divide the data into various length window and then assign labels manually based on marking timestamp for each activity; to improve the metrics for data sets

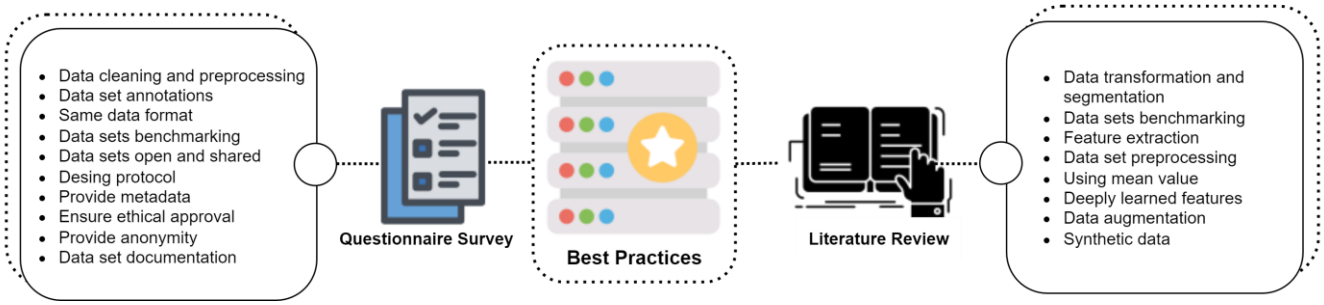


Fig. 23. Best practices explored from literature review and recommended by survey participants

- *Globalization of data sets:* good data sets play a critical role in the globalization of data sets by providing accurate, reliable, and relevant information that can be shared and used across different countries and cultures. The distribution of data sets paly an essential role in leveraging the practices and knowledge globally. Accurate and good quality data sets are widely accessible, and the contributors may retrieve new information from all regions globally [176], [182].
- *Good data sets can play a crucial role in early disease detection by providing the information needed to identify patterns and trends that can indicate the presence of a disease:* HAR and health care related data sets developed from wearable sensors allows to detect disease early to assist medical experts to improve treatment and patient health care [180], [181], [182].
- *Rehabilitation:* by providing valuable information that can inform the design, implementation, and evaluation of rehabilitation programs for example identifying patient needs, tracking progress, evaluating outcomes, identifying risk factors and disparities. properly collected data sets can play an important role in rehabilitation to assess, treat and manage an individual to leverage their social, physical, cognitive and physiological functions and get back to normal condition [183], [184].

B. Best practices

Fig. 23 presents the best practices used and recommended by HAR researchers from both the literature review published literature and survey. The best practices arising from the survey are related to all the data set lifecycle phases. However,

benchmarking. Also, feature extraction and data set pre-processing to make data sets fit for their experiments. Researchers used mean values when dealing with missing values in a data set. Deeply learned features are frequently more robust and accurate than hand-engineered features because the model can learn from the data itself, as instead of relying on human-designed features. Further, researchers used data augmentation to reduce overfitting and deeply learned features for dealing with large scale video-based data sets. Finally, using synthetic data to overcome the problem of privacy and data scarcity to deal with sensitive and more personal information.

The recommended best practices from survey participants on the left side of Fig. 23 are data cleaning and pre-processing of data set before sharing and provide data annotation while constructing a data set. Also, they recommended using standardized data formats during data collection, performing data set benchmarking to validate and improve data sets, data sets should be open source and should be easily accessible to all researchers. Researchers should design a protocol for data collection during data set construction and provide complete metadata related to a data set to improve data quality and reusability. Importantly, researchers should ensure ethical approval before data collection from the host organization and provide anonymity to participants preserve privacy. Finally, providing complete documentation with shared data sets will assist other researchers prior performing their experiments.

Regarding the data collection methods such as controlled and uncontrolled in HAR domain, the distinction between both methods plays a pivotal role in shaping the landscape of dataset quality and applicability [185]. Controlled data collection involves particularly structured environments and predefined activities, enabling precise data labelling and facilitating algorithm training. This method ensures a high degree of consistency and reproducibility that can be used for benchmarking and fine-tuning algorithms [26]. On the other

hand, uncontrolled data collection unfolds in real-world contexts, capturing natural human behaviour. While generating data sets that represent real scenarios, this approach introduces challenges such as data noise and annotation ambiguity [30]. Both methodologies offer unique insights and bear relevance in addressing diverse research problems. Exploring the interaction between these methods is crucial, as it drives advancements in HAR methodologies, enriches data set diversity, and augments the field's overall robustness.

VI. Classification of Issues and Challenges With Open Data Sets

The main objective of this section is to organize and classify all of the identified issues and challenges in a systematic manner to assist researchers in their understanding and interpretation of the results. From the conducted comprehensive literature review and survey, key issues and challenges relating to open data sets in HAR have been identified in terms of all phases of the open data set lifecycle. In addition, we also make reference to external factors relevant to these issues and challenges, as shown in Fig. 24. These external factors were derived through authors' brainstorming on the results of the literature review, survey, and related published research work [100].

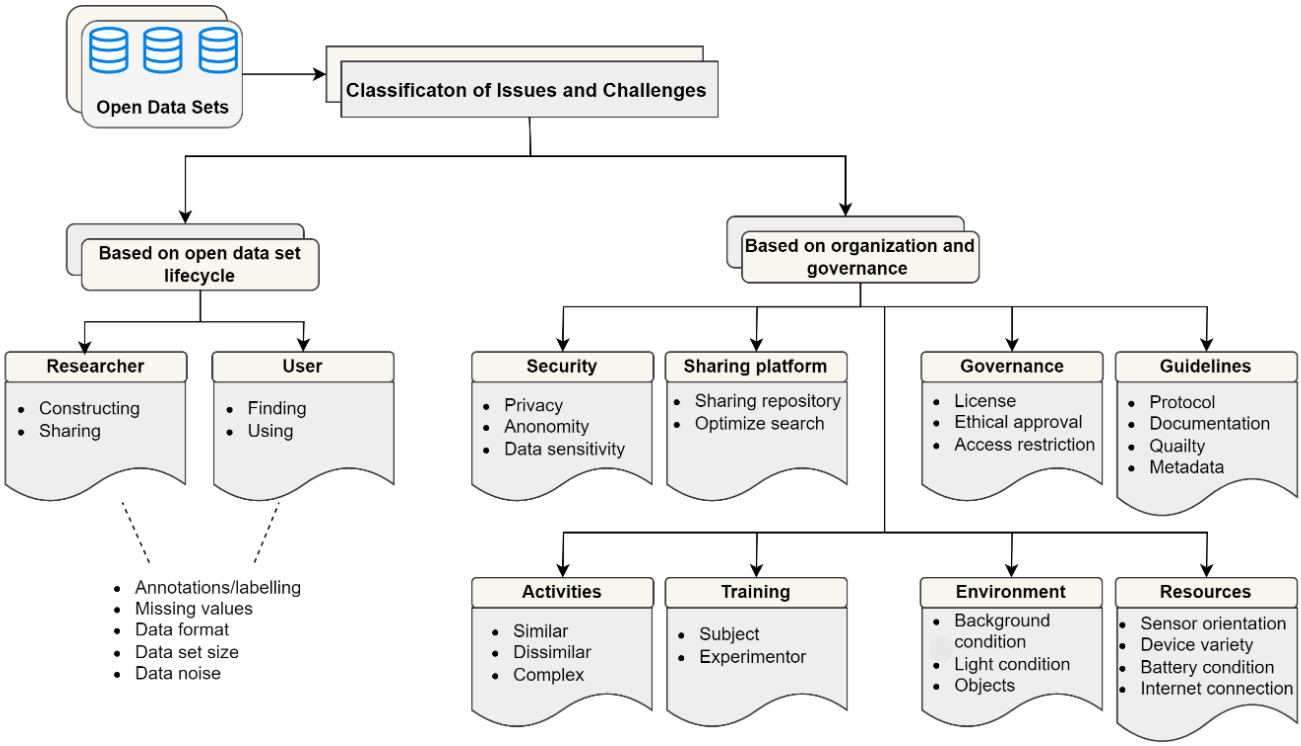


Fig. 24: Classification of open data sets issues and challenges

1) *Based on data set lifecycle*

This is divided into researcher and user perspectives. Researchers construct data sets and then share and deposit them into data set repositories. Users find data sets by applying relevant keywords and search strings to download and use them according to their research purposes. The issues and challenges in this category have been discussed above and include data set annotations/labelling, missing values, data format, data set size and data noise.

2) *Based on organization and governance*

This is further divided into 8 major categories: security, sharing platform, governance, guidelines, activities, training, environment, and resources.

Security: This category is concerned with the security aspect of open data sets. Privacy concerns are the main issue while constructing and sharing data sets. Data set owners must ensure anonymity during data set construction and sharing to protect the personal data of the participants. Because some people in the data set may not want their data to be made public. Additionally, certain sorts of data, such as medical information or financial records, may be particularly sensitive and should be secured properly. Data must be appropriately anonymized and de-identified, and access restrictions and monitoring must be in place to avoid illegal access and exploitation of open datasets.

Sharing Platform: A sharing platform is an online community where people and groups may pool their resources and knowledge for the greater good. A sharing repository is a database or online storage space where users may deposit and retrieve files, documents, and other media. Making it simpler for users to access and share open datasets, a sharing repository on a platform might be invaluable. Open datasets are those that may be accessed, shared, and augmented by anybody who wants to do so. All sorts of studies, analyses, and machine learning projects might benefit from these data sets.

Several methods may be used by a sharing platform to make the most of available public datasets.

- Facilitating the discovery of useful datasets by providing search and filtering tools.
- Facilitating data analysis and interpretation by providing visual representations of data.
- facilitating users' ability to work together and share their platform-based discoveries.
- Pushing users to add their own open datasets to the service as a means of contributing to it.

Governance: These are the emerging issues related to open data sets such as proper licensing of the data while sharing. Similarly, ethical approval for the host organization and access restrictions on the collected data sets for sharing in a repository. Proper licencing, ethical approval, and access limits are all components of open dataset governance. Data sharing and commercial usage are only two examples of the kinds of restrictions that might be stated in a licence. Legal and ethical considerations, such as the need to preserve individuals' privacy, are considered throughout the ethical review process. Data can be made available to the public or kept private, depending on the access settings in place. Good governance of open datasets guarantees that data is utilised in a way that doesn't break the law or compromise ethical principles.

Guidelines: A common issue to arise during data set construction is the need to design a protocol for data collection in a systematic way. Similarly, data consistency and mitigation of missing values, insufficient data samples, and metadata description of the data set are common problems faced by the research community. Open data set guidelines involve a protocol for data collection and sharing that ensures data is collected ethically and legally and shared in a way that is accessible and understandable to others. Documentation is also important, as it provides information on the data's origin, context, and any limitations or biases present. Quality is crucial, as inaccurate, or unreliable data can negatively impact research and decision-making. Metadata, or data about the data, is also important as it provides information on the data's structure, format, and any relevant information for understanding and using the data. Overall, open dataset guidelines aim to promote transparency, accessibility, and reliability in data collection and sharing.

Activities: Open HAR datasets are collections of data that are publicly available for researchers and developers to use in their work. These data sets are often used in activities related to recognizing similar, dissimilar, and complex activities. The main technical concern in HAR data sets is the recognition of dissimilar activities such as eating, drinking, and making tea. For example, researchers may use open HAR data sets to train machine learning models to recognize similar activities, such as walking or running, based on sensor data. Similarly, the recognition of complex activities having sub-activities and activities involving multiple participants. Additionally, open HAR data sets can be used to analyse complex activities, such as multi-tasking or multitasking while performing different activities.

Training: HAR open data sets are collections of data collected from a variety of subjects and individuals, generally via wearable devices like smartwatches or activity trackers. These data sets are used to train machine learning models to identify and categorise various human activities, such as walking, running, and cycling. Data set annotation is the main challenge for constructing a data set and it is often unnoticed due to insufficient or no training of the recruited participants/users for experimental study. Awareness and training of the experimenter related to environmental factors involved are also needed while conducting an experiment.

Environment: Environmental and background conditions is the most challenging factor during data set construction. The open HAR dataset collecting environment consists of a range of background conditions, lighting conditions, and other potential scene objects. This includes indoor and outdoor settings, varying levels of illumination, and diverse things such as furniture, people, and vehicles. The data set is gathered in a naturalistic context, which means that the individuals are conducting activities of everyday life in their natural surroundings, unrestricted by any external factors. Researchers must be able to control and avoid irrelevant and noisy information and irrelevant objects during data collection.

Resources: This category includes the challenges related to sensor placement and orientation during data collection. Similarly, connectivity of the Internet with sensors, devices, and software applications. Also, a variety of different devices producing various data output formats and camera variations, and the battery condition of wearable devices. The data set resources contain information on sensor orientation, devices employed, battery health of devices and sensors, and internet connectivity. This information is essential for comprehending the performance and limitations of various wearable devices and sensors, and it may be utilised to enhance the design and functioning of future devices. Additionally, the data set is continually updated to reflect new technologies and trends in the industry, making it an excellent resource for keeping up with the latest developments in wearable technology.

VII. HAR Data Evolution

Data sets are not always static artefacts and the data set lifecycle can be iterative in nature. Adopting this perspective, a number of promising avenues come to the forefront regarding data set evolution.

Enhancing Real-World Diversity: Subsequent iterations of HAR data sets have the potential to incorporate a broader range of real-world settings, therefore reflecting the complex composition of human actions across varied surroundings. By broadening its scope to include uncontrolled environments, data sets have the potential to accurately represent the intricacies of everyday life, resulting in models that possess more resilience, adaptability, and alignment with real-world user encounters.

Cross-Domain Merging: The integration of HAR data with data sets originating from other domains, such as healthcare, environmental monitoring, or social interactions, presents a promising opportunity for cross-domain merging. The

utilization of an interdisciplinary approach has the potential to reveal new and unique perspectives and associations, facilitating a more profound comprehension of human behaviour within various settings.

Multimodal Fusion: The integration of data from diverse sensors and modalities, including accelerometers, gyroscopes, audio, and video, has the potential to enhance the informative richness of the data set. The integration of these two components has the potential to facilitate the development of comprehensive and precise activity recognition models, hence expanding the present limitations in this field.

Longitudinal and Contextual Data: The inclusion of longitudinal data, which tracks the progression of activities over time, and contextual information such as environmental conditions and user emotions, has the potential to increase the data set's temporal and situational value. This technological development has the potential to provide more sophisticated and contextually sensitive identification of human activities.

Privacy-Preservation: Privacy-preservation refers to methods that aim to protect individuals' privacy during data collection and analysis. Evolving HAR data sets have the potential to investigate novel approaches to preserving user privacy, all the while providing important insights. Methods such as differential privacy and secure multi-party computing have the potential to be included in data-collecting protocols.

Customization Focused on User Needs: The adaptation of HAR data sets to provide the distinctive actions and preferences of individual users has the potential to generate personalized models that more effectively correspond to users' distinct behavioural patterns, hence improving accuracy and usability.

Benchmarking Standards: The establishment of defined standards and assessment criteria for HAR data sets has the potential to facilitate a more uniform and comparative evaluation of various models and algorithms, hence expediting advancements in the area.

Collaborative Environment: The growth of open data sharing and collaboration among researchers has the potential to facilitate the generation of extensive and diverse data sets, hence facilitating the advancement of more robust and generalizable models for HAR.

VIII. Conclusion and Future Work

The emergence of new computing technologies with the ability to collect more detailed and accurate data has led to the generation of huge amounts of data, and its relevance to understanding and impact on decision making is increasing. Generated data are typically collected in the form of data sets. Researchers and practitioners use data sets for research objectives to understand the totality of an area of interest and to develop a basis for making decisions. The primary objective of constructing a data set and making it available and open to others is to allow benchmarking, replication and validation of research approaches as well as exploration of novel hypotheses.

The main objective of this research study is to identify current issues and challenges faced by researchers in the HAR domain with respect to open data sets. A literature review and survey were conducted to identify these issues and challenges from the published literature and the research community. The identified issues and challenges were classified for ease of understanding and interpretation. This classification of issues and challenges will help HAR researchers to be aware of the open issues and challenges in HAR open data sets. This research has helped to identify and promote important attributes such as privacy, anonymity, platform maintenance, data sets descriptions and metadata, environmental condition, resources, and training while constructing and sharing new data sets. In future work, our own data sets will be shared using the recommendations and good practices identified. An evaluation workshop will be conducted involving other HAR researchers to explore the above issues and challenges along with other outcomes of our work in curating open data sets in HAR, including the analysis of data sets to automatically extract metadata and to assess data set quality.

Appendix A

Questions collection that was asked in the conducted questionnaire survey.

- **Demographic information (to help us understand the type of respondents to our survey)**

1. Occupation: Required
2. Organization/research institute:

- **Your experience using data sets.**

3. Approximately how many years' experience do you have in working with open data sets in human activity recognition?
4. What is your experience in using open data sets in human activity recognition?

Constructing?

5. Have you taken part in constructing an open data set in human activity recognition?

- ***Data Set Construction***

6. The dataset I constructed may not be shared because it contains information that is (select all that apply); If you selected Other, please specify:
7. For the research group / organization in which I work, normal practice in relation to experimental data sets is to (please indicate the statement that best describes your situation).
8. What are your preferred data formats for data set construction and sharing?
9. Are there other data sharing formats not on the above list which you use?
10. What is the main piece of advice you would give to another researcher when generating a new data set in human activity recognition?
11. What is the main issue or challenge you have faced when generating a new data set in human activity recognition?

Sharing?

12. Have you taken part in sharing open data set in human activity recognition?

- ***Data Set Sharing***

13. The sharing of data was limited because it contained identifying information about an organization or participants.
14. Sharing the data set was not possible because of problems with the data?
15. Sharing the data set was not possible because it was too large?
16. I did not share the data set because I was unsure of the best approach.
17. When I shared the data set I used the following license.
18. I found the process of depositing and sharing the data set time consuming.
19. As a researcher, what motivates you to share your data set with the human activity recognition research community?
20. What is the main issue or challenge you have faced during data set sharing?
21. What is the main piece of advice you would give to another researcher when sharing a new data set in human activity recognition?

Finding?

22. Have you searched for and downloaded an open data set from an open data online repository for experimental / research purposes?

- ***Data Set Finding***

23. In searching for a data set, I make a selection based on the following.
24. How much pre-processing did you need to perform on the data set after downloading?
25. In your view, what factors improve data set quality (one per line)?
26. As a researcher, what motivates you to gain access to an existing data set instead of creating a new data set?
27. What open data set repositories / directories do you typically search when looking for a data set?
28. What is the main issue or challenge you have faced when trying to find a suitable data set?
29. What is the main piece of advice you would give to another researcher when searching for a data set in human activity recognition?

Using?

30. Have you used and evaluated someone else's open data set for experimental / research purposes?

- ***Data Set Using***

31. The metadata describing the data set was easy to understand?
32. The metadata describing the data set was accurate?
33. I encountered a data set update issue after the initial sharing
34. What is the main issue or challenge you have faced when trying to use someone else's data set in your research?
35. What is the main piece of advice you would give to another researcher when using someone else's data set in human activity recognition?
36. Have you ever pre-registered an experiment in any the following online repository domains?
37. As a researcher, how important to you is the replication of a human activity recognition experiment using an open data set?
38. As a researcher, how important to you is the benchmarking of a human activity recognition approach using an open data set?
39. In your view, what are the major issues and challenges in relation to using open data sets in human activity recognition research?
40. Have you any other comments regarding the survey?

APPENDIX B

Table 15: Paper metrics for Fig. 5

Digital Libraries	Paper Title	ML technique
ACM	Classical Machine Learning Approach for Human Activity Recognition Using Location Data	RF
	Emotion Recognition in the Wild from Videos using Images	CNN, SVM
	Emotion Recognition in the Wild using Deep Neural Networks and Bayesian Classifiers	CNN, BN
	Ensemble Approach for Sensor-Based Human Activity Recognition	SVM, KNN, RF
	Face Recognition via Active Annotation and Learning	DNN
	Feature Based Random Forest Nurse Care Activity Recognition Using Accelerometer Data	RF
	From Individual to Group-Level Emotion Recognition: EmotiW 5.0	SVM
	UPIC: User and Position Independent Classical Approach for Locomotion and Transportation Modes Recognition	RF
	Modeling Multimodal Cues in a Deep Learning-Based Framework for Emotion Recognition in the Wild	CNN, LSTM
	Multi-modal Emotion Recognition using Semi-supervised Learning and Multiple Neural Networks in the Wild	CNN, LSTM
	Multi-view Common Space Learning for Emotion Recognition in the Wild	CNN
	Summary of the 2*nd* Nurse Care Activity Recognition Challenge Using Lab and Field Data	kNN
	HoloNet: Towards Robust Emotion Recognition in the Wild	CNN
IEEE	Deep Triplet Networks with Attention for Sensor-based Human Activity Recognition	LSTM
	A Semisupervised Recurrent Convolutional Attention Model for Human Activity Recognition	CNN, LSTM, SVM
	A Comparative Study on Missing Data Handling Using Machine Learning for	SVN, RF

	Human Activity Recognition	
	Group Activity Description and Recognition based on Trajectory Analysis and Neural Networks	SOM
	Hidden Markov Model-Based Fall Detection With Motion Sensor Orientation Calibration: A Case for Real-Life Home Monitoring	HMM
	Recognition of Real-life Activities with Smartphone Sensors using Deep Learning Approaches	CNN, LSTM
	Transition-Aware Housekeeping Task Monitoring Using Single Wrist-Worn Sensor	SVM, NB, LSTM
	Unsupervised Recognition of Multi-Resident Activities in Smart-Homes	HMM
Springer	The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition	PCA, SVM, HMM
	Human Action Prediction with 3D-CNN	CNN, LSTM
	Wrapper Filter Approach for Accelerometer-Based Human Activity Recognition	RF, k-NN, GB
	A revised framework of machine learning application for optimal activity recognition	SVM, MLP
	Feature learning for Human Activity Recognition using Convolutional Neural Networks	RF, CNN
	Multi modal human action recognition for video content matching	SVM
	A Framework for Semi-Supervised Adaptive Learning for Activity Recognition in Healthcare Applications	BN
	Efcacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments	LSTM, CNN
	Novel approaches to human activity recognition based on accelerometer data	CNN
	A resource conscious human action recognition framework using 26-layered deep convolutional neural network	CNN, SVM, k-NN
	Human-Sensing: Low Resolution Thermal Array Sensor Data Classification of Location-Based Postures	J48
	A fall detection method based on a joint motion map using double convolutional neural networks	CNN
Science Direct	Daily Human Activities Recognition Using Heterogeneous Sensors from Smartphones	MLP
	Cross-subject transfer learning in human activity recognition systems using generative adversarial networks	GAN
	Attention induced multi-head convolutional neural network for human activity recognition	CNN
	A smartphone sensors-based personalized human activity recognition system for sustainable smart cities	DRNN
	Efficiency investigation from shallow to deep neural network techniques in human activity recognition	ANN, CNN
	GCHAR: An efficient Group-based Context—aware human activity recognition on smartphone	k-NN, RF, j48
	Human activity recognition based on smartphone and wearable sensors using multiscale DCNN ensemble	CNN
	Human activity recognition with smartphone sensors using deep learning neural networks	ANN, SVM, MLP, J48, NB
	Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems	CNN
	Robust least squares twin support vector machine for human activity recognition	SVM
	Robust Human Activity Recognition using smartwatches and smartphones	RF, CNN, HMM, MLP, LSTM
	Online active learning for human activity recognition from sensory data streams	DT, SVM
	Multi-label classification based ensemble learning for human activity recognition in smart home	NB, DT, k-NN
	Modified deep residual network architecture deployed on serverless framework of IoT platform based on human activity recognition application	SVM, CNN, LSTM

ACKNOWLEDGMENT

We acknowledge the Department for the Economy (DfE) Ulster University for financial support and the School of Computing, Ulster University, Belfast Campus, UK.

REFERENCES

[1] S. Gupta, 'Deep learning based human activity recognition (HAR) using wearable sensor data', *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100046, Nov. 2021, doi: 10.1016/j.jjimei.2021.100046.

[2] A. Haleem, M. Javaid, M. Asim Qadri, R. Pratap Singh, and R. Suman, 'Artificial intelligence (AI) applications for marketing: A literature-based study', *Int. J. Intell. Netw.*, vol. 3, pp. 119–132, Jan. 2022, doi: 10.1016/j.ijin.2022.08.005.

[3] D. Sheth and M. Shah, 'Predicting stock market using machine learning: best and accurate way to know future stock prices', *Int. J. Syst. Assur. Eng. Manag.*, vol. 14, no. 1, pp. 1–18, Feb. 2023, doi: 10.1007/s13198-022-01811-1.

[4] Y.-T. Chang and N.-H. Fan, 'A novel approach to market segmentation selection using artificial intelligence techniques', *J. Supercomput.*, vol. 79, no. 2, pp. 1235–1262, Feb. 2023, doi: 10.1007/s11227-022-04666-2.

[5] S. P. Arnerić, V. D. Kern, and D. T. Stephenson, 'Regulatory-accepted drug development tools are needed to accelerate innovative CNS disease treatments', *Biochem. Pharmacol.*, vol. 151, pp. 291–306, May 2018, doi: 10.1016/j.bcp.2018.01.043.

[6] H. Liu, W. Zhang, B. Zou, J. Wang, Y. Deng, and L. Deng, 'DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy', *Nucleic Acids Res.*, vol. 48, no. D1, pp. D871–D881, Jan. 2020, doi: 10.1093/nar/gkz1007.

[7] M. Sun and J. Zhang, 'Research on the application of block chain big data platform in the construction of new smart city for low carbon emission and green environment', *Comput. Commun.*, vol. 149, pp. 332–342, Jan. 2020, doi: 10.1016/j.comcom.2019.10.031.

[8] N. S. Suhaimi, J. Mountstephens, and J. Teo, 'A Dataset for Emotion Recognition Using Virtual Reality and EEG (DER-VREEG): Emotional State Classification Using Low-Cost Wearable VR-EEG Headsets', *Big Data Cogn. Comput.*, vol. 6, no. 1, Art. no. 1, Mar. 2022, doi: 10.3390/bdcc6010016.

[9] C. Y. Wang, Q. Zhou, G. Fitzmaurice, and F. Anderson, 'VideoPoseVR: Authoring Virtual Reality Character Animations with Online Videos', *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. ISS, p. 575:448-575:467, Nov. 2022, doi: 10.1145/3567728.

[10] G. Tsueng *et al.*, 'Developing a standardized but extendable framework to increase the findability of infectious disease datasets', *Sci. Data*, vol. 10, no. 1, Art. no. 1, Feb. 2023, doi: 10.1038/s41597-023-01968-9.

[11] R. Singh, A. Sonawane, and R. Srivastava, 'Recent evolution of modern datasets for human activity recognition: a deep survey', *Multimed. Syst.*, vol. 26, no. 2, pp. 83–106, Apr. 2020, doi: 10.1007/s00530-019-00635-7.

[12] A. D. Antar, M. Ahmed, and M. A. R. Ahad, 'Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: a review', in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, 2019, pp. 134–139.

[13] H. A. Piwowar and W. W. Chapman, 'Public sharing of research datasets: A pilot study of associations', *ASIST–ISSI Metr. Pre-Conf. Semin. Glob. Alliance*, vol. 4, no. 2, pp. 148–156, Apr. 2010, doi: 10.1016/j.joi.2009.11.010.

[14] A. Ambhaikar, 'A Survey on Health Care and Expert System', *Math. Stat. Eng. Appl.*, vol. 72, no. 1, Art. no. 1, Jan. 2023.

[15] Y. Li, G. Yang, Z. Su, S. Li, and Y. Wang, 'Human activity recognition based on multienvironment sensor data', *Inf. Fusion*, vol. 91, pp. 47–63, Mar. 2023, doi: 10.1016/j.inffus.2022.10.015.

[16] C. Nugent *et al.*, 'An initiative for the creation of open datasets within pervasive healthcare', ICST, the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2016, pp. 318–321. Accessed: Nov. 20, 2021. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-64771

[17] Z. Zhang, W. Wang, A. An, Y. Qin, and F. Yang, 'A human activity recognition method using wearable sensors based on convtransformer model', *Evol. Syst.*, Jan. 2023, doi: 10.1007/s12530-022-09480-y.

[18] N. Davies and S. Clinch, 'Pervasive Data Science', *IEEE Pervasive Comput.*, vol. 16, no. 3, pp. 50–58, 2017, doi: 10.1109/MPRV.2017.2940956.

[19] A. Bexheti, M. Langheinrich, and S. Clinch, 'Secure Personal Memory-Sharing with Co-located People and Places', in *Proceedings of the 6th International Conference on the Internet of Things*, in IoT'16. New York, NY, USA: Association for Computing Machinery, Nov. 2016, pp. 73–81. doi: 10.1145/2991561.2991577.

[20] P. Kumar and S. Suresh, 'Deep-HAR: an ensemble deep learning model for recognizing the simple, complex, and heterogeneous human activities', *Multimed. Tools Appl.*, Feb. 2023, doi: 10.1007/s11042-023-14492-0.

[21] Y. Chen, Y. Gu, X. Jiang, and J. Wang, 'OCEAN: a new opportunistic computing model for wearable activity recognition', in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, in UbiComp '16. New York, NY, USA: Association for Computing Machinery, Sep. 2016, pp. 33–36. doi: 10.1145/2968219.2971453.

[22] C.-Y. Huang *et al.*, 'Flexible Pressure Sensor with an Excellent Linear Response in a Broad Detection Range for Human Motion Monitoring', *ACS Appl. Mater. Interfaces*, vol. 15, no. 2, pp. 3476–3485, Jan. 2023, doi: 10.1021/acsami.2c19465.

[23] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, 'Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities', *Sensors*, vol. 23, no. 4, Art. no. 4, Jan. 2023, doi: 10.3390/s23042182.

[24] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, 'Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges', *Expert Syst. Appl.*, vol. 105, pp. 233–261, Sep. 2018, doi: 10.1016/j.eswa.2018.03.056.

[25] G. Alam, I. McChesney, P. Nicholl, and J. Rafferty, 'An Approach to Extract and Compare Metadata of Human Activity Recognition (HAR) Data Sets', in *Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2022)*, J. Bravo, S. Ochoa, and J. Favela, Eds., in Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2023, pp. 717–728. doi: 10.1007/978-3-031-21333-5_71.

[26] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, and A. Kumar, 'A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets', *Appl. Artif. Intell.*, vol. 36, no. 1, p. 2093705, Dec. 2022, doi: 10.1080/08839514.2022.2093705.

[27] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, 'A survey of video datasets for human action and activity recognition', *Comput. Vis. Image Underst.*, vol. 117, no. 6, pp. 633–659, Jun. 2013, doi: 10.1016/j.cviu.2013.01.013.

[28] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, 'Vision-based human activity recognition: a survey', *Multimed. Tools Appl.*, vol. 79, no. 41, pp. 30509–30555, Nov. 2020, doi: 10.1007/s11042-020-09004-3.

[29] S. Vishwakarma and A. Agrawal, 'A survey on activity recognition and behavior understanding in video surveillance', *Vis. Comput.*, vol. 29, no. 10, pp. 983–1009, Oct. 2013, doi: 10.1007/s00371-012-0752-6.

[30] L. Onofri, P. Soda, M. Pechenizkiy, and G. Iannello, 'A survey on using domain and contextual knowledge for human activity recognition in video streams', *Expert Syst. Appl.*, vol. 63, pp. 97–111, Nov. 2016, doi: 10.1016/j.eswa.2016.06.011.

[31] M. H. Arshad, M. Bilal, and A. Gani, 'Human Activity Recognition: Review, Taxonomy and Open Challenges', *Sensors*, vol. 22, no. 17, Art. no. 17, Jan. 2022, doi: 10.3390/s22176463.

[32] A. Lentzas and D. Vrakas, 'Non-intrusive human activity recognition and abnormal behavior detection on elderly people: a review', *Artif. Intell. Rev.*, vol. 53, no. 3, pp. 1975–2021, Mar. 2020, doi: 10.1007/s10462-019-09724-5.

[33] L. Minh Dang, K. Min, H. Wang, Md. Jalil Piran, C. Hee Lee, and H. Moon, 'Sensor-based and vision-based human activity recognition: A comprehensive survey', *Pattern Recognit.*, vol. 108, p. 107561, Dec. 2020, doi: 10.1016/j.patcog.2020.107561.

[34] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, 'Trends in human activity recognition using smartphones', *J. Reliab. Intell. Environ.*, vol. 7, no. 3, pp. 189–213, Sep. 2021, doi: 10.1007/s40860-021-00147-0.

[35] S. Zolfaghari, M. R. Keyvanpour, and R. Zall, 'Analytical Review on Ontological Human Activity Recognition Approaches', *Int. J. E-Bus. Res. IJEBR*, vol. 13, no. 2, pp. 58–78, Apr. 2017, doi: 10.4018/IJEBR.2017040104.

[36] I. H. Sarker, 'Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective', *SN Comput. Sci.*, vol. 2, no. 5, p. 377, Jul. 2021, doi: 10.1007/s42979-021-00765-8.

[37] Y. Wang *et al.*, 'A Novel Deep Multi-Feature Extraction Framework Based on Attention Mechanism Using Wearable Sensor Data for Human Activity Recognition', *IEEE Sens. J.*, pp. 1–1, 2023, doi: 10.1109/JSEN.2023.3242603.

[38] D. H. Zwitter Bruno S. Frey,Gerd Gigerenzer,Ernst Hafen,Michael Hagner,Yvonne Hofstetter,Jeroen van den Hoven,Roberto V. Zicari,Andrej, 'Will Democracy Survive Big Data and Artificial Intelligence?', *Scientific American*, Nov. 26, 2021. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/ (accessed Nov. 26, 2021).

[39] 'Open Government Data - OECD'. https://www.oecd.org/gov/digital-government/open-government-data.htm (accessed Nov. 26, 2021).

[40] 'What is Open Data?' https://opendatahandbook.org/guide/en/what-is-open-data/ (accessed Nov. 26, 2021).

[41] P. Andanda, 'Towards a Paradigm Shift in Governing Data Access and Related Intellectual Property Rights in Big Data and Health-Related Research', *IIC - Int. Rev. Intellect. Prop. Compet. Law*, vol. 50, no. 9, pp. 1052–1081, Nov. 2019, doi: 10.1007/s40319-019-00873-2.

[42] X. Zhu, C. Thomas, J. C. Moore, and S. Allen, 'Open Government Data Licensing: An Analysis of the U.S. State Open Government Data Portals', in *Diversity, Divergence, Dialogue*, K. Toeppe, H. Yan, and S. K. W. Chu, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 260–273. doi: 10.1007/978-3-030-71305-8_21.

[43] C. Pernet, C. Svarer, R. Blair, J. D. Van Horn, and R. A. Poldrack, 'On the Long-term Archiving of Research Data', *Neuroinformatics*, Feb. 2023, doi: 10.1007/s12021-023-09621-x.

[44] 'These Are The Best Free Open Data Sources Anyone Can Use', *freeCodeCamp.org*, Jan. 10, 2019. https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/ (accessed Nov. 19, 2021).

[45] T. Davidson, E. Wall, and J. Mace, 'A Qualitative Interview Study of Distributed Tracing Visualisation: A Characterisation of Challenges and Opportunities', *IEEE Trans. Vis. Comput. Graph.*, pp. 1–12, 2023, doi: 10.1109/TVCG.2023.3241596.

[46] D. Cook, K. D. Feuz, and N. C. Krishnan, 'Transfer learning for activity recognition: a survey', *Knowl. Inf. Syst.*, vol. 36, no. 3, pp. 537–556, Sep. 2013, doi: 10.1007/s10115-013-0665-3.

[47] P. Vepakomma, D. De, S. K. Das, and S. Bhansali, 'A-Wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities', in *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, Jun. 2015, pp. 1–6. doi: 10.1109/BSN.2015.7299406.

[48] Y. Kim and Y. Li, 'Human Activity Classification With Transmission and Reflection Coefficients of On-Body Antennas Through Deep Convolutional Neural Networks', *IEEE Trans. Antennas Propag.*, vol. 65, no. 5, pp. 2764–2768, May 2017, doi: 10.1109/TAP.2017.2677918.

[49] N. Y. Hammerla, S. Halloran, and T. Ploetz, 'Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables', *ArXiv160408880 Cs Stat*, Apr. 2016, Accessed: Nov.

20, 2021. [Online]. Available: http://arxiv.org/abs/1604.08880

[50] J. Qin, L. Liu, Z. Zhang, Y. Wang, and L. Shao, 'Compressive Sequential Learning for Action Similarity Labeling', *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 756–769, Feb. 2016, doi: 10.1109/TIP.2015.2508600.

[51] Y.-H. Kim *et al.*, 'MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech', in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, in CHI '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–21. doi: 10.1145/3491102.3517457.

[52] M. T. Leving, H. L. D. Horemans, R. J. K. Vegter, S. de Groot, J. B. J. Bussmann, and L. H. V. van der Woude, 'Validity of consumer-grade activity monitor to identify manual wheelchair propulsion in standardized activities of daily living', *PLOS ONE*, vol. 13, no. 4, p. e0194864, Apr. 2018, doi: 10.1371/journal.pone.0194864.

[53] X. Yu, J. Jang, and S. Xiong, 'A Large-Scale Open Motion Dataset (KFall) and Benchmark Algorithms for Detecting Pre-impact Fall of the Elderly Using Wearable Inertial Sensors', *Front. Aging Neurosci.*, vol. 13, p. 692865, Jul. 2021, doi: 10.3389/fnagi.2021.692865.

[54] J. C. E. Guerrero, E. M. España, M. M. Añasco, and J. E. P. Lopera, 'Dataset for human fall recognition in an uncontrolled environment', *Data Brief*, vol. 45, p. 108610, Sep. 2022, doi: 10.1016/j.dib.2022.108610.

[55] A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, 'SisFall: A Fall and Movement Dataset', *Sensors*, vol. 17, no. 1, Art. no. 1, Jan. 2017, doi: 10.3390/s17010198.

[56] R. Singh, A. Sonawane, and R. Srivastava, 'Recent evolution of modern datasets for human activity recognition: a deep survey', *Multimed. Syst.*, vol. 26, no. 2, pp. 83–106, Apr. 2020, doi: 10.1007/s00530-019-00635-7.

[57] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, 'ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding', 2015, pp. 961–970. Accessed: Nov. 20, 2021. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.html

[58] H. Zheng, D. Liu, and Y. Liu, 'Design and research on automatic recognition system of sports dance movement based on computer vision and parallel computing', *Microprocess. Microsyst.*, vol. 80, p. 103648, Feb. 2021, doi: 10.1016/j.micpro.2020.103648.

[59] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, 'From action to activity: Sensor-based activity recognition', *Neurocomputing*, vol. 181, pp. 108–115, Mar. 2016, doi: 10.1016/j.neucom.2015.08.096.

[60] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Qiu Chen, 'Temporal context network for activity localization in videos', in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5793–5802.

[61] D. Wu, N. Sharma, and M. Blumenstein, 'Recent advances in video-based human action recognition using deep learning: A review', in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 2865–2872. doi: 10.1109/IJCNN.2017.7966210.

[62] T. Hassner, 'A critical review of action recognition benchmarks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 245–250.

[63] N. Ikizler-Cinbis, R. Gokberk Cinbis, and S. Sclaroff, 'Learning actions from the Web', in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 995–1002. doi: 10.1109/ICCV.2009.5459368.

[64] A. Torralba and A. A. Efros, 'Unbiased look at dataset bias', in *CVPR 2011*, Jun. 2011, pp. 1521–1528. doi: 10.1109/CVPR.2011.5995347.

[65] 'UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set'. http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones (accessed Nov. 20, 2021).

[66] 'Harvard Dataverse'. https://dataverse.harvard.edu/ (accessed Nov. 20, 2021).

[67] 'Dataset Search'. https://datasetsearch.research.google.com/ (accessed Nov. 20, 2021).

[68] 'IEEE DataPort', *IEEE DataPort*. https://ieee-dataport.org/ (accessed Nov. 20, 2021).

[69] 'Zenodo - Research. Shared.' https://zenodo.org/ (accessed Nov. 20, 2021).

[70] 'figshare - credit for all your research'. https://figshare.com/ (accessed Nov. 26, 2021).

[71] 'hollywood data on data.world | 10 datasets available', *data.world*. https://data.world/datasets/hollywood (accessed Nov. 26, 2021).

[72] 'Action Similarity Labeling Challenge'. https://talhassner.github.io/home/projects/ASLAN/ASLAN-main.html (accessed Nov. 27, 2021).

[73] 'Papers with Code - YouCook Dataset'. https://paperswithcode.com/dataset/youcook (accessed Nov. 27, 2021).

[74] 'Welcome to CASAS'. http://casas.wsu.edu/datasets/ (accessed Nov. 27, 2021).

[75] 'UCI Machine Learning Repository: OPPORTUNITY Activity Recognition Data Set'. https://archive.ics.uci.edu/ml/datasets/opportunity+activity+recognition (accessed Nov. 27, 2021).

[76] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, 'From individual to group-level emotion recognition: EmotiW 5.0', in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, in ICMI '17. New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 524–528. doi: 10.1145/3136755.3143004.

[77] 'UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set'. http://archive.ics.uci.edu/ml/datasets/WISDM+Smartphone+and+Smartwatch+Activity+and+Biometrics+Dataset+ (accessed Aug. 05, 2022).

[78] 'UCI Machine Learning Repository: PAMAP2 Physical Activity Monitoring Data Set'. https://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring (accessed Aug. 05, 2022).

[79] 'Sussex-Huawei Locomotion Dataset'. http://www.shl-dataset.org/ (accessed Aug. 05, 2022).

[80] F. Mastrogiovanni, 'fulviomas/WHARF'. Dec. 11, 2014. Accessed: Aug. 05, 2022. [Online]. Available: https://github.com/fulviomas/WHARF

[81] M. Soliman, T. Fatnassi, I. Elgammal, and R. Figueiredo, 'Exploring the Major Trends and Emerging Themes of Artificial Intelligence in the Scientific Leading Journals amidst the COVID-19 Era', *Big Data Cogn. Comput.*, vol. 7, no. 1, Art. no. 1, Mar. 2023, doi: 10.3390/bdcc7010012.

[82] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, 'Survey on sentiment analysis: evolution of research methods and topics', *Artif. Intell. Rev.*, Jan. 2023, doi: 10.1007/s10462-022-10386-z.

[83] J. Osterrieder, 'A Primer on Artificial Intelligence and Machine Learning for the Financial Services Industry'. Rochester, NY, Feb. 05, 2023. doi: 10.2139/ssrn.4349078.

[84] M. Chhabra, K. K. Ravulakollu, M. Kumar, A. Sharma, and A. Nayyar, 'Improving automated latent fingerprint detection and segmentation using deep convolutional neural network', *Neural Comput. Appl.*, vol. 35, no. 9, pp. 6471–6497, Mar. 2023, doi: 10.1007/s00521-022-07894-y.

[85] D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, and M. R. Luaces, 'New machine learning approaches for real-life human activity recognition using smartphone sensor-based data', *Knowl.-Based Syst.*, vol. 262, p. 110260, Feb. 2023, doi: 10.1016/j.knosys.2023.110260.

[86] L. Bai, H. Li, W. Gao, J. Xie, and H. Wang, 'A joint multiobjective optimization of feature selection and classifier design for high-dimensional data classification', *Inf. Sci.*, vol. 626, pp. 457–473, May 2023, doi: 10.1016/j.ins.2023.01.069.

[87] S. Buyrukoğlu and S. Savaş, 'Stacked-Based Ensemble Machine Learning Model for Positioning Footballer', *Arab. J. Sci. Eng.*, vol. 48, no. 2, pp. 1371–1383, Feb. 2023, doi: 10.1007/s13369-022-06857-8.

[88] H. H. Ali, H. M. Moftah, and A. A. A. Youssif, 'Depth-based human activity recognition: A comparative perspective study on feature extraction', *Future Comput. Inform. J.*, vol. 3, no. 1, pp. 51–67, Jun. 2018, doi: 10.1016/j.fcij.2017.11.002.

[89] A. Ray, M. H. Kolekar, R. Balasubramanian, and A. Hafiane, 'Transfer Learning Enhanced Vision-based Human Activity Recognition: A Decade-long Analysis', *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 1, p. 100142, Apr. 2023, doi: 10.1016/j.jjimei.2022.100142.

[90] K. Henricksen, J. Indulska, and A. Rakotonirainy, 'Modeling Context Information in Pervasive Computing Systems', in *Pervasive Computing*, F. Mattern and M. Naghshineh, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2002, pp. 167–180. doi: 10.1007/3-540-45866-2_14.

[91] B. Šumak, S. Brdnik, and M. Pušnik, 'Sensors and Artificial Intelligence Methods and Algorithms for Human–Computer Intelligent Interaction: A Systematic Mapping Study', *Sensors*, vol. 22, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/s22010020.

[92] Y. Wang, S. Cang, and H. Yu, 'A survey on wearable sensor modality centred human activity recognition in health care', *Expert Syst. Appl.*, vol. 137, pp. 167–190, Dec. 2019, doi: 10.1016/j.eswa.2019.04.057.

[93] S. K. Challa, A. Kumar, and V. B. Semwal, 'A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data', *Vis. Comput.*, vol. 38, no. 12, pp. 4095–4109, Dec. 2022, doi: 10.1007/s00371-021-02283-3.

[94] Q. Shen, H. Feng, R. Song, D. Song, and H. Xu, 'Federated Meta-Learning with Attention for Diversity-Aware Human Activity Recognition', *Sensors*, vol. 23, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/s23031083.

[95] J. C. Stamper *et al.*, 'Managing the Educational Dataset Lifecycle with DataShop', in *Artificial Intelligence in Education*, G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 557–559. doi: 10.1007/978-3-642-21869-9_100.

[96] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, 'Data and its (dis)contents: A survey of dataset development and use in machine learning research', *Patterns*, vol. 2, no. 11, p. 100336, Nov. 2021, doi: 10.1016/j.patter.2021.100336.

[97] K. KELLEY, B. CLARK, V. BROWN, and J. SITZIA, 'Good practice in the conduct and reporting of survey research', *Int. J. Qual. Health Care*, vol. 15, no. 3, pp. 261–266, May 2003, doi: 10.1093/intqhc/mzg031.

[98] 'Open-Ended Questions: How to Code & Analyze for Insights [2018]', *Thematic*, Jun. 08, 2018. https://getthematic.com/insights/code-open-ended-questions-in-surveys-to-get-deep-insights/ (accessed Aug. 09, 2022).

[99] A. Castleberry and A. Nolen, 'Thematic analysis of qualitative research data: Is it as easy as it sounds?', *Curr. Pharm. Teach. Learn.*, vol. 10, no. 6, pp. 807–815, Jun. 2018, doi: 10.1016/j.cptl.2018.03.019.

[100] T. Singh and D. K. Vishwakarma, 'Video benchmarks of human action datasets: a review', *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 1107–1154, Aug. 2019, doi: 10.1007/s10462-018-9651-1.

[101] F. Cruciani, I. Cleland, C. Nugent, P. McCullagh, K. Synnes, and J. Hallberg, 'Automatic Annotation for Human Activity Recognition in Free Living Using a

Smartphone', *Sensors*, vol. 18, no. 7, Art. no. 7, Jul. 2018, doi: 10.3390/s18072203.

[102] F. Alharbi, L. Ouarbya, and J. A. Ward, 'Comparing Sampling Strategies for Tackling Imbalanced Data in Human Activity Recognition', *Sensors*, vol. 22, no. 4, p. 1373, 2022.

[103] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, 'Emotion recognition in the wild from videos using images', in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, in ICMI '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 433–436. doi: 10.1145/2993148.2997627.

[104] L. Surace, M. Patacchiola, E. Battini Sönmez, W. Spataro, and A. Cangelosi, 'Emotion recognition in the wild using deep neural networks and Bayesian classifiers', in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, in ICMI '17. New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 593–597. doi: 10.1145/3136755.3143015.

[105] S. Brajesh and I. Ray, 'Ensemble approach for sensor-based human activity recognition', in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, in UbiComp-ISWC '20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 296–300. doi: 10.1145/3410530.3414352.

[106] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, 'Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild', in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, in ICMI '17. New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 529–535. doi: 10.1145/3136755.3143005.

[107] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, 'HoloNet: towards robust emotion recognition in the wild', in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, in ICMI '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 472–478. doi: 10.1145/2993148.2997639.

[108] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, 'Deep Triplet Networks with Attention for Sensor-based Human Activity Recognition', in *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Mar. 2021, pp. 1–10. doi: 10.1109/PERCOM50583.2021.9439116.

[109] R. Mohamed, T. Perumal, Md. N. Sulaiman, N. Mustapha, and M. N. Razali, 'Conflict resolution using enhanced label combination method for complex activity recognition in smart home environment', in *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, Oct. 2017, pp. 1–3. doi: 10.1109/GCCE.2017.8229477.

[110] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, 'A semisupervised recurrent convolutional attention model for human activity recognition', *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1747–1756, 2019.

[111] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, J. Garcia-Rodriguez, M. Cazorla, and M. T. Signes-Pont, 'Group activity description and recognition based on trajectory analysis and neural networks', in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2016, pp. 1585–1592.

[112] H. Bi, M. Perello-Nieto, R. Santos-Rodriguez, and P. Flach, 'Human activity recognition based on dynamic active learning', *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 922–934, 2020.

[113] K.-C. Liu, C.-Y. Hsieh, and C.-T. Chan, 'Transition-Aware Housekeeping Task Monitoring Using Single Wrist-Worn Sensor', *IEEE Sens. J.*, vol. 18, no. 21, pp. 8950–8962, Nov. 2018, doi: 10.1109/JSEN.2018.2868278.

[114] D. Riboni and F. Murru, 'Unsupervised Recognition of Multi-Resident Activities in Smart-Homes', *IEEE Access*, vol. 8, pp. 201985–201994, 2020, doi: 10.1109/ACCESS.2020.3036226.

[115] S. Jha, M. Schiemer, F. Zambonelli, and J. Ye, 'Continual learning in sensor-based human activity recognition: An empirical benchmark analysis', *Inf. Sci.*, vol. 575, pp. 1–21, Oct. 2021, doi: 10.1016/j.ins.2021.04.062.

[116] Z. N. Khan and J. Ahmad, 'Attention induced multi-head convolutional neural network for human activity recognition', *Appl. Soft Comput.*, vol. 110, p. 107671, Oct. 2021, doi: 10.1016/j.asoc.2021.107671.

[117] J. Sena, J. Barreto, C. Caetano, G. Cramer, and W. R. Schwartz, 'Human activity recognition based on smartphone and wearable sensors using multiscale DCNN ensemble', *Neurocomputing*, vol. 444, pp. 226–243, Jul. 2021, doi: 10.1016/j.neucom.2020.04.151.

[118] C. A. Ronao and S.-B. Cho, 'Human activity recognition with smartphone sensors using deep learning neural networks', *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016, doi: 10.1016/j.eswa.2016.04.032.

[119] M. Jethanandani, A. Sharma, T. Perumal, and J.-R. Chang, 'Multi-label classification based ensemble learning for human activity recognition in smart home', *Internet Things*, vol. 12, p. 100324, Dec. 2020, doi: 10.1016/j.iot.2020.100324.

[120] B. M. Abidine, L. Fergani, B. Fergani, and M. Oussalah, 'The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition', *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 119–138, Feb. 2018, doi: 10.1007/s10044-016-0570-y.

[121] L. Al-Frady and A. Al-Taei, 'Wrapper Filter Approach for Accelerometer-Based Human Activity Recognition', *Pattern Recognit. Image Anal.*, vol. 30, no. 4, pp. 757–764, Oct. 2020, doi: 10.1134/S1054661820040033.

[122] P. Gupta, R. McClatchey, and P. Caleb-Solly, 'Tracking changes in user activity from unlabelled

smart home sensor data using unsupervised learning methods', *Neural Comput. Appl.*, vol. 32, no. 16, pp. 12351–12362, Aug. 2020, doi: 10.1007/s00521-020-04737-6.

[123] R. A. Hamad, M. Kimura, and J. Lundström, 'Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments', *SN Comput. Sci.*, vol. 1, no. 4, p. 204, Jun. 2020, doi: 10.1007/s42979-020-00211-1.

[124] A. Jordao, L. A. B. Torres, and W. R. Schwartz, 'Novel approaches to human activity recognition based on accelerometer data', *Signal Image Video Process.*, vol. 12, no. 7, pp. 1387–1394, Oct. 2018, doi: 10.1007/s11760-018-1293-x.

[125] B. Pontes, M. Cunha, R. Pinho, and H. Fuks, 'Human-Sensing: Low Resolution Thermal Array Sensor Data Classification of Location-Based Postures', in *Distributed, Ambient and Pervasive Interactions*, N. Streitz and P. Markopoulos, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 444–457. doi: 10.1007/978-3-319-58697-7_33.

[126] L. Yao, W. Yang, and W. Huang, 'A fall detection method based on a joint motion map using double convolutional neural networks', *Multimed. Tools Appl.*, vol. 81, no. 4, pp. 4551–4568, Feb. 2022, doi: 10.1007/s11042-020-09181-1.

[127] H. Ye *et al.*, 'Face Recognition via Active Annotation and Learning', in *Proceedings of the 24th ACM international conference on Multimedia*, in MM '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 1058–1062. doi: 10.1145/2964284.2984059.

[128] S. Pini, O. B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet, 'Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild', in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, in ICMI '17. New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 536–543. doi: 10.1145/3136755.3143006.

[129] S. S. Alia *et al.*, 'Summary of the 2nd nurse care activity recognition challenge using lab and field data', in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, in UbiComp-ISWC '20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 378–383. doi: 10.1145/3410530.3414611.

[130] E. Soleimani and E. Nazerfard, 'Cross-subject transfer learning in human activity recognition systems using generative adversarial networks', *Neurocomputing*, vol. 426, pp. 26–34, Feb. 2021, doi: 10.1016/j.neucom.2020.10.056.

[131] A. R. Javed, R. Faheem, M. Asim, T. Baker, and M. O. Beg, 'A smartphone sensors-based personalized human activity recognition system for sustainable smart cities', *Sustain. Cities Soc.*, vol. 71, p. 102970, Aug. 2021, doi: 10.1016/j.scs.2021.102970.

[132] L. Cao, Y. Wang, B. Zhang, Q. Jin, and A. V. Vasilakos, 'GCHAR: An efficient Group-based Context—aware human activity recognition on smartphone', *J. Parallel Distrib. Comput.*, vol. 118, pp. 67–80, Aug. 2018, doi: 10.1016/j.jpdc.2017.05.007.

[133] R. Khemchandani and S. Sharma, 'Robust least squares twin support vector machine for human activity recognition', *Appl. Soft Comput.*, vol. 47, pp. 33–46, Oct. 2016, doi: 10.1016/j.asoc.2016.05.025.

[134] R. San-Segundo, H. Blunck, J. Moreno-Pimentel, A. Stisen, and M. Gil-Martín, 'Robust Human Activity Recognition using smartwatches and smartphones', *Eng. Appl. Artif. Intell.*, vol. 72, pp. 190–202, Jun. 2018, doi: 10.1016/j.engappai.2018.04.002.

[135] S. Mohamad, M. Sayed-Mouchaweh, and A. Bouchachia, 'Online active learning for human activity recognition from sensory data streams', *Neurocomputing*, vol. 390, pp. 341–358, May 2020, doi: 10.1016/j.neucom.2019.08.092.

[136] P. Gupta and P. Caleb-Solly, 'A Framework for Semi-Supervised Adaptive Learning for Activity Recognition in Healthcare Applications', in *Engineering Applications of Neural Networks*, E. Pimenidis and C. Jayne, Eds., in Communications in Computer and Information Science. Cham: Springer International Publishing, 2018, pp. 3–15. doi: 10.1007/978-3-319-98204-5_1.

[137] C. Lübbe, B. Friedrich, S. Fudickar, S. Hellmers, and A. Hein, 'Feature based random forest nurse care activity recognition using accelerometer data', in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, in UbiComp-ISWC '20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 408–413. doi: 10.1145/3410530.3414340.

[138] M. S. Siraj *et al.*, 'UPIC: user and position independent classical approach for locomotion and transportation modes recognition', in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, in UbiComp-ISWC '20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 340–345. doi: 10.1145/3410530.3414343.

[139] S. Mekruksavanich and A. Jitpattanakul, 'Recognition of Real-life Activities with Smartphone Sensors using Deep Learning Approaches', in *2021 IEEE 12th International Conference on Software Engineering and Service Science (ICSESS)*, Aug. 2021, pp. 243–246. doi: 10.1109/ICSESS52187.2021.9522231.

[140] J. Suto and S. Oniga, 'Efficiency investigation from shallow to deep neural network techniques in human activity recognition', *Cogn. Syst. Res.*, vol. 54, pp. 37–49, May 2019, doi: 10.1016/j.cogsys.2018.11.009.

[141] R. Alfaifi and A. M. Artoli, 'Human Action Prediction with 3D-CNN', *SN Comput. Sci.*, vol. 1, no. 5, p. 286, Aug. 2020, doi: 10.1007/s42979-020-00293-x.

[142] M. Bilal, F. K. Shaikh, M. Arif, and M. F. Wyne, 'A revised framework of machine learning application for optimal activity recognition', *Clust. Comput.*, vol. 22, no. 3, pp. 7257–7273, May 2019, doi: 10.1007/s10586-017-1212-x.

[143] F. Cruciani *et al.*, 'Feature learning for Human Activity Recognition using Convolutional Neural Networks', *CCF Trans. Pervasive Comput. Interact.*, vol. 2, no. 1, pp. 18–32, Mar. 2020, doi: 10.1007/s42486-020-00026-2.

[144] J. Guo, H. Bai, Z. Tang, P. Xu, D. Gan, and B. Liu, 'Multi modal human action recognition for video content matching', *Multimed. Tools Appl.*, vol. 79, no. 45, pp. 34665–34683, Dec. 2020, doi: 10.1007/s11042-020-08998-0.

[145] M. A. Khan, Y.-D. Zhang, S. A. Khan, M. Attique, A. Rehman, and S. Seo, 'A resource conscious human action recognition framework using 26-layered deep convolutional neural network', *Multimed. Tools Appl.*, vol. 80, no. 28, pp. 35827–35849, Nov. 2021, doi: 10.1007/s11042-020-09408-1.

[146] T. Wu *et al.*, 'CARMUS: Towards a General Framework for Continuous Activity Recognition with Missing Values on Smartphones', in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, Jul. 2018, pp. 850–859. doi: 10.1109/COMPSAC.2018.00148.

[147] T. Hossain and S. Inoue, 'A Comparative study on missing data handling using machine learning for human activity recognition', in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, 2019, pp. 124–129.

[148] S. Yu, H. Chen, and R. A. Brown, 'Hidden Markov model-based fall detection with motion sensor orientation calibration: A case for real-life home monitoring', *IEEE J. Biomed. Health Inform.*, vol. 22, no. 6, pp. 1847–1853, 2017.

[149] A. Keshavarzian, S. Sharifian, and S. Seyedin, 'Modified deep residual network architecture deployed on serverless framework of IoT platform based on human activity recognition application', *Future Gener. Comput. Syst.*, vol. 101, pp. 14–28, Dec. 2019, doi: 10.1016/j.future.2019.06.009.

[150] S. Hossain Arib, R. Akter, O. Shahid, and M. A. R. Ahad, 'Classical Machine Learning Approach for Human Activity Recognition Using Location Data', in *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, in UbiComp '21. New York, NY, USA: Association for Computing Machinery, Sep. 2021, pp. 340–345. doi: 10.1145/3460418.3479376.

[151] E. Kwon, H. Park, S. Byon, E.-S. Jung, and Y.-T. Lee, 'HaaS (Human Activity Analytics as a Service) Using Sensor Data of Smart Devices', in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, 2018, pp. 1500–1502.

[152] U. Ozbulak, B. Vandersmissen, A. Jalalvand, I. Couckuyt, A. Van Messem, and W. De Neve, 'Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems', *Comput. Vis. Image Underst.*, vol. 202, p. 103111, Jan. 2021, doi: 10.1016/j.cviu.2020.103111.

[153] M.-S. Dao, T.-A. Nguyen-Gia, and V.-C. Mai, 'Daily Human Activities Recognition Using Heterogeneous Sensors from Smartphones', *Procedia Comput. Sci.*, vol. 111, pp. 323–328, Jan. 2017, doi: 10.1016/j.procs.2017.06.030.

[154] N. M. Richards and J. H. King, 'Big data ethics', *Wake For. Rev*, vol. 49, p. 393, 2014.

[155] T. Simonite, 'Google's AI guru wants computers to think more like brains', *Wired*, 2018.

[156] S. Mohamed, M.-T. Png, and W. Isaac, 'Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence', *Philos. Technol.*, vol. 33, no. 4, pp. 659–684, 2020.

[157] P. Chahuara, A. Fleury, M. Vacher, and F. Portet, 'Méthodes SVM et MLN pour la reconnaissance automatique d'activités humaines dans les habitats perceptifs: tests et perspectives', in *RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*, 2012, pp. 978–2.

[158] N. Mollet and R. Chellali, 'Détection et interprétation des gestes de la main', in *2005 3rd International Conference on SETIT*, 2005.

[159] R. Merkley, 'Use and Fair Use: Statement on shared images in facial recognition AI', *Creat. Commons*, 2019.

[160] D. Sculley, J. Snoek, A. Wiltschko, and A. Rahimi, 'Winner's curse? On pace, progress, and empirical rigor', 2018.

[161] G. Bhat, N. Tran, H. Shill, and U. Y. Ogras, 'w-HAR: An activity recognition dataset and framework using low-power wearable devices', *Sensors*, vol. 20, no. 18, p. 5356, 2020.

[162] S.-J. VAN ELS, D. GRAUS, and E. BEAUXIS-AUSSALET, 'Improving Fairness Assessments with Synthetic Data: a Practical Use Case with a Recommender System for Human Resources'.

[163] B. Xin *et al.*, 'Federated synthetic data generation with differential privacy', *Neurocomputing*, vol. 468, pp. 1–10, Jan. 2022, doi: 10.1016/j.neucom.2021.10.027.

[164] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, 'Deep convolutional neural networks on multichannel time series for human activity recognition', in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

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

[166] Y. LeCun, F. J. Huang, and L. Bottou, 'Learning methods for generic object recognition with invariance to pose and lighting', in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Jun. 2004, p. II-104 Vol.2. doi: 10.1109/CVPR.2004.1315150.

[167] J. J.-C. Ying, B.-H. Lin, V. S. Tseng, and S.-Y. Hsieh, 'Transfer Learning on High Variety Domains for Activity Recognition', in *Proceedings of the ASE BigData & SocialInformatics 2015*, in ASE BD&SI '15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1–6. doi: 10.1145/2818869.2818890.

[168] H. M. S. Hossain, M. A. A. H. Khan, and N. Roy, 'Active learning enabled activity recognition', *Pervasive Mob. Comput.*, vol. 38, pp. 312–330, Jul. 2017, doi: 10.1016/j.pmcj.2016.08.017.

[169] V. Xafis and M. K. Labude, 'Openness in Big Data and Data Repositories', *Asian Bioeth. Rev.*, vol. 11, no. 3, pp. 255–273, Sep. 2019, doi: 10.1007/s41649-019-00097-z.

[170] 'Center for Data and Visualization Sciences | Duke University Libraries'. https://library.duke.edu/data (accessed Aug. 09, 2022).

[171] U. of Bristol, 'Managing research data'. http://www.bristol.ac.uk/staff/researchers/data/ (accessed Aug. 09, 2022).

[172] H. Wang, 'LibGuides: Machine Learning and AI: Home'. https://guides.library.cmu.edu/machine-learning/home (accessed Aug. 09, 2022).

[173] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, 'Data Cleaning: Overview and Emerging Challenges', in *Proceedings of the 2016 International Conference on Management of Data*, in SIGMOD '16. New York, NY, USA: Association for Computing Machinery, Jun. 2016, pp. 2201–2206. doi: 10.1145/2882903.2912574.

[174] S. Singhal and M. Jena, 'A study on WEKA tool for data preprocessing, classification and clustering', *Int. J. Innov. Technol. Explor. Eng. IJItee*, vol. 2, no. 6, pp. 250–253, 2013.

[175] S. A. Alasadi and W. S. Bhaya, 'Review of data preprocessing techniques in data mining', *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

[176] K. Jee and G.-H. Kim, 'Potentiality of big data in the medical sector: focus on how to reshape the healthcare system', *Healthc. Inform. Res.*, vol. 19, no. 2, pp. 79–85, 2013.

[177] M. Mancini, 'Exploiting big data for improving healthcare services', *J. E-Learn. Knowl. Soc.*, vol. 10, no. 2, 2014.

[178] E. Baro, S. Degoul, R. Beuscart, and E. Chazard, 'Toward a literature-driven definition of big data in healthcare', *BioMed Res. Int.*, vol. 2015, 2015.

[179] M. B. Howren, M. W. Vander Weg, and F. D. Wolinsky, 'Computerized cognitive training interventions to improve neuropsychological outcomes: evidence and future directions', *J. Comp. Eff. Res.*, vol. 3, no. 2, pp. 145–154, 2014.

[180] W. Raghupathi and V. Raghupathi, 'Big data analytics in healthcare: promise and potential', *Health Inf. Sci. Syst.*, vol. 2, no. 1, pp. 1–10, 2014.

[181] L. M. Fernandes, M. O'Connor, and V. Weaver, 'Big data, bigger outcomes', *J. AHIMA*, vol. 83, no. 10, pp. 38–43, 2012.

[182] J.-C. Hsieh, A.-H. Li, and C.-C. Yang, 'Mobile, cloud, and big data computing: contributions, challenges, and new directions in telecardiology', *Int. J. Environ. Res. Public. Health*, vol. 10, no. 11, pp. 6131–6153, 2013.

[183] A. Barzegar Khanghah, G. Fernie, and A. Roshan Fekr, 'Design and Validation of Vision-Based Exercise Biofeedback for Tele-Rehabilitation', *Sensors*, vol. 23, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/s23031206.

[184] J. Wu, S. G. Faux, I. Harris, C. J. Poulos, and T. Alexander, 'Record linkage is feasible with non-identifiable trauma and rehabilitation datasets', *Aust. N. Z. J. Public Health*, vol. 40, no. 3, pp. 245–249, 2016, doi: 10.1111/1753-6405.12510.

[185] F. Cruciani *et al.*, 'A Public Domain Dataset for Human Activity Recognition in Free-Living Conditions', in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2019, pp. 166–171. doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00071.

Gulzar Alam is currently pursuing a Ph.D. in Computing Science at Ulster University, School of Computing, United Kingdom. He has devoted his career to enhancing the field of computer science through his contributions to data curation and the quality of data sets for human activity recognition (HAR). Gulzar received his Master of Science in Software Engineering from the King Fahd University of Petroleum and Minerals (KFUPM) in Saudi Arabia, where he also served as an undergraduate teaching assistant. Following the completion of his master's degree, Gulzar worked on several cybersecurity and software engineering related research projects. He obtained valuable experience conducting research, collaborating with teams, and publishing academic papers and a patent during this time. His current research focuses on enhancing the quality of data sets in the realm of HAR. Gulzar has also collaborated with academic institutions on various funded projects to advance the current state of knowledge in computer science.

Ian McChesney is a Senior Lecturer in Computing Science. His first degree is in computer science and he has a PhD in software engineering. He is a Fellow of The British Computer Society, a Chartered Engineer and a Senior Fellow of the Higher Education Academy. Dr. McChesney has higher education experience in research, teaching and knowledge transfer with industry. His research interests in software engineering have focused on socio-technical themes such as software estimation, program comprehension and software team coordination, using methods such as field survey, empirical software engineering and eye tracking. In pervasive computing he has

worked on experimental human activity recognition and support tools for managing open data sets. His knowledge transfer experience covers areas such as software project management and requirements engineering.

Peter Nicholl received the B.Eng degree in Electronic Systems and the Ph.D. degree in feature encoding from the University of Ulster, Belfast, U.K. in 1991 and 1994, respectively. He is a Senior Lecturer in Computing Science within the School of Computing, Ulster University. He is a Senior Fellow of the Higher Education Academy. Dr Nicholl has higher education experience in research, teaching and knowledge transfer with industry. His research interests include intelligent transportation, computer vision, and deep learning.

Joseph Rafferty received the B.Eng. degree in computer science from Queen's University Belfast, and the M.Sc. degree in computing and the Ph.D. degree in computer science from Ulster University. He is currently a Lecturer with the School of Computing, Ulster University. His research interests include intention recognition, smart environments, agent-based systems, connected health, sensor technology, and planning and intelligent systems.