Where Is the Internet of Health Things Data?

Evilasio Costa Junior[a], Rossana M. C. Andrade[b], Amanda D. P. Venceslau[c], Pedro Almir M. Oliveira[d], Ismayle S. Santos[e] and Breno S. Oliveira[f]
Group of Computer Networks, Software Engineering, Systems (Great), Federal University of Ceará (UFC), Ceará, Brazil

Keywords: Internet of Health Things, Databases, Systematic Multivocal Review.

Abstract: The advent of Internet of Things (IoT) and the smart objects popularization have boosted the data generation in many areas. Data have then become increasingly valuable as they can be used to "teach" machines to perform the most varied tasks. Health is among the areas that have benefited from such data, because there is, for example, a need for solutions that optimize the cost-benefit ratio of health systems. In this scenario, the Internet of Health Things (IoHT) uses smart sensors to collect patient data and intelligent algorithms to process this data for improving patient Quality of Life. However, researchers and practitioners have faced difficulties in finding and using public health care data sensor repositories. Therefore, we conducted a systematic multivocal review of IoHT databases to identify and characterize the existing datasets. We also bring as a contribution of this paper a set of guidelines about how new IoHT data repositories can be structured.

1 INTRODUCTION

The Internet of Things (IoT) was proposed over at least two decades (Ashton, 2009) and it was initially inspired by the glimpse of Mark Weiser related to ubiquitous computing together with the ideas to use sensors in order to enable computers to understand the world (Weiser, 1999). Since then, IoT has been adapted and strengthened from advances in many areas (Atzori et al., 2010), for example, miniaturization of sensors, expansion of data processing power, and improvements of machine learning algorithms.

These advances have enabled the Internet of Things use in many cross-section areas, achieving process enhancements and cost reductions. One area that has stood out in the use of this technology is healthcare (Islam et al., 2015). In the past, remote patient monitoring was complex and expensive. Nowadays, this kind of follow-up can be done using smartphone sensors (Meskó, 2014). As a consequence, a new research area has emerged: the Internet of Health Things (IoHT) (Rodrigues et al., 2018).

According to (Rodrigues et al., 2018), IoHT uses many kinds of sensors to collect patient data. Then, these data are transmitted to more robust nodes (*e.g.*, gateways), which can perform initial processing, or, if necessary, send the dataset to the cloud. Finally, the health data can be processed using Machine Learning techniques or analyzed by health professionals.

Given the advance in data storage and processing tools, datasets have become even more valuable, as they can be used to describe processes, optimize procedures, and for task automation using machine learning (Miloslavskaya and Tolstoy, 2016). However, despite the vast amount of available data, there are still challenges related to data silos and data lakes, standardization of devices, specialized IoHT platforms, quality assurance, data security, and privacy (Oliveira et al., 2022), in addition to the absence of public catalogs that facilitate access to such datasets (Selvaraj and Sundaravaradhan, 2020).

This paper focuses on investigating public catalogs of IoHT datasets and, for that, we performed a Multivocal Literature Review (MLR) to identify and characterize the existing datasets. We believe that the contributions of this work are as follows: (i) a set of datasets that can be used for other researchers to assess new proposals; (ii) a set of guidelines to organize the creation of new public datasets supporting the reuse by other researchers and (iii) Limitations and shortcomings of the literature regarding the

[a] https://orcid.org/0000-0002-0281-2964
[b] https://orcid.org/0000-0002-0186-2994
[c] https://orcid.org/0000-0003-4118-4224
[d] https://orcid.org/0000-0002-3067-3076
[e] https://orcid.org/0000-0001-5580-643X
[f] https://orcid.org/0000-0003-0079-8799

Costa Junior, E., Andrade, R., Venceslau, A., Oliveira, P., Santos, I. and Oliveira, B.
Where Is the Internet of Health Things Data?.
DOI: 10.5220/0011050300003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 39-49
ISBN: 978-989-758-569-2; ISSN: 2184-4992

datasets exposing challenges that may be interesting for future research (e.g., few descriptions regarding the pre-processing data, how to assess enough number of instances for the datasets, how to deal with the heterogeneity of data formats and the provenance of the collected data).

The paper outline is: Section 2 presents our study design; Section 3 discusses our results; Section 4 introduce a set of guidelines related to IoHT datasets; Section 5 points our some validity threats; and, finally, Sections 6 and 7 present the related work and our final considerations, respectively.

2 STUDY DESIGN

We performed a Multivocal Literature Review (MLR) about IoHT datasets. In this MLR study, we decided to search information both in the scientific literature (*e.g.*, articles, books, theses, and dissertations - white literature) and in the grey literature, that according (Garousi et al., 2019), includes preprints, e-prints, technical reports, lectures, datasets, audio-video media, and blogs.

Therefore, we based our Multivocal Literature Review on the methods proposed by (Brereton et al., 2007), (Kitchenham et al., 2009), and (Wohlin, 2014). For search in the grey literature, we also used the guidelines proposed in (Garousi et al., 2019). These are the most used methods for developing literature reviews in the software engineering area and have three activities: Planning, Execution (or conducting), and Presentation (or documentation). In the MLR planning, we define the research questions, the search strategy and generate the protocol that guides the execution. The latter contains the general objective of the review, the search strategy, the research questions, the papers' eligibility criteria, and the list of data that we would like to extract from the selected literature. In the conducting phase, we execute the search strategy and apply the eligibility criteria for selecting the papers. After this, we extract and synthesize the data. Finally, we generate the report in the presentation phase and discuss the results. This paper presents our report and contains both the results of the MLR and the discussion about them.

2.1 Planning

The first stage of planning consists of defining the objective of the literature review and specifying the Research Questions (RQ). This MLR aims to present a systematic multivocal review on the Internet of Health Things datasets, highlighting problems, technologies and limitations. Following RQs guided our study:

- **RQ1:** What are the existing IoHT public datasets?
- **RQ2:** What are the limitations of the existing Internet of Health Things datasets?
- **RQ3:** What technologies are relevant in creating and querying this kind of data sources?

We analyzed and discussed the answers to these questions in Section 3.

The **search strategy** of this MLR consists of two phases. In the first phase, we applied a seach string to find papers in scientific studies databases for white literature search and public repositories and internet search engines for grey literature search. In the second phase, we performed a manual procedure, known as snowballing forward (Wohlin, 2014), to analyze the citations of the articles previously selected in the first phase. Snowballing complements the search procedure in the public scientific datasets, making the white literature search coverage more comprehensive.

We chose Scopus, Web of Science, and Compendex for the white literature search. In addition, according to (Archambault et al., 2009), and (Aghaei Chadegani et al., 2013), which are relevant search datasets for Computer Science, aggregating works of several other relevant datasets for the area of Computing and related. Our search for grey literature was done using the Google Search Engine[1], Archive[2] and GitHub[3], which contain files of various formats and system source codes, as well as scientific articles not yet published or in the conception process.

Table 1: Identified elements of the PICo approach.

Aspect	Identified Element
Population	Academic Papers and Grey Literature
Interest	Public Databases, Public Datasets or Catalogs
Context	Internet of Things and Health

To built our **query string**, PICo approach was adopted (Pai et al., 2004). This method separates the question into three aspects: Population, Interest, and Context (PICo). The Population represents the kind of studies we would like to address in the research. The Interest corresponds to the research objective. Finally, the Context corresponds to the information we would like to find in our population studies. Table 1 shows the elements identified for each component of the PICo.

[1]Google website: https://www.google.com
[2]Archive website: https://archive.org
[3]GitHub website: https://github.com

Table 2: Final Query String.

((“Public Database” OR “Public Dataset” OR “Public Datasource” OR “Public Catalog” OR “Open Database” OR “Open Dataset” OR “Open Datasource” OR “Open Catalog”) AND (IoT OR “Internet of Things” OR “System of System” OR “Ubiquitous Computing” OR Sensors) AND (Health OR eHealth OR Telemedicine OR Wellbeing OR Wellness))

We evaluated many strings until we obtained the final version presented in Table 2. This search string was used for both white and grey literature searches.

For the selection of the most relevant studies, it is necessary to define inclusion and exclusion criteria (called **eligibility criteria**) that can be replicated by other researchers (Kitchenham et al., 2009).

The inclusion criteria used in this research are: (I1) Contains or presents addressing for datasets with health data; (I2) Only Datasets with free use licenses; and (I3) Only Datasets that contain sensor data. Moreover, we defined the following exclusion criteria for this MLR: (E1) Non-English papers; (E2) Papers with less than five pages (short paper); (E3) Video Datasets; (E4) The dataset does not contain sensor data characterization; (E5) The article or document does not contain a link to the base or base reference; (E6) The dataset does not contain characteristics of the individuals used in the experiments; and (E7) The dataset does not contain information on how and which experiments were performed.

In this MLR, the exclusion criteria operate in sequential order similar to an Access Control List (ACL) as in (Sandhu and Samarati, 1994). Thus, when we found a match on the list, we performed the exclusion action and did not check any other criterion.

To complete the planning phase, we defined the data extracted from the datasets found in this MLR and generated a data extraction form. The form containing the information to be extracted from each paper can be seen at the link https://bit.ly/3q5D5qD.

2.2 Conducting

In this phase, we executed a search with the query string in databases of academic papers and with the search filters referring to the exclusion criteria E1 and E2, which we applied directly in the search engines of the databases. Consequently, we found thirty-nine (39) papers and four hundred forty-four (444) repositories related to grey literature. We exclude twenty-four (24) papers and five (5) repositories by applying the exclusion and inclusion criteria based on reading the articles' title and abstract and the web repositories' title. Then, we performed the transversal reading of the fifteen (15) papers, and the analysis of the description and content of the four hundred thirty-nine (439) repositories remained. According to the eligibility criteria, we exclude nine (9) papers and four hundred thirty-three repositories (433). Hence, we selected six (6) papers and six (6) repositories containing datasets of sensors for use in health care and monitoring health applications.

There were many grey literature repositories excluded after analyzing their description and content, as we identified that most of these repositories contained applications and small datasets to be used as an example of the use of these applications. Also, there was no description of the data records in these datasets, making the use of them unfeasible.

Then, we applied the snowballing forward technique, identifying article citations in Google Scholar, as suggested by (Wohlin, 2014). Hence, we analyzed the title, abstract and executed the transversal reading of fifty-five (55) papers found. According to the eligibility rules, forty-nine (49) articles were excluded, leaving six (6) articles at the end. At the end, we obtained twelve studies (12) of white literature and six (6) repositories from grey literature.

After searching and selecting papers and grey literature repositories, we identified the datasets presented in the articles and grey literature repositories. Finally, we extracted the data from the datasets using the extraction form created in the planning phase. In all, we found forty-four (44) different datasets.

It is worth noting that some selected articles had more than one dataset. There are also datasets used in more than one article or present in more than one repository. Finally, some repositories presented more than one dataset that met the eligibility criteria, such as the Kaggle[4] and Physionet[5] repositories.

Lastly, we arranged the extracted data in a spreadsheet and synthesized them. Then, we used the Tableau tool[6] for quantitative data analysis, and we performed the content analysis for subjective and qualitative interpretation of the extracted data.

3 RESULTS AND DISCUSSION

As previously described, in this investigation, we started the analysis with 483 items (among scientific articles and data repositories found in the grey literature). This number was refined until we had only

[4]Kaggle website: https://www.kaggle .com

[5]Physionet website: https://physionet.org

[6]Tableau website: https://www.tableau.com.

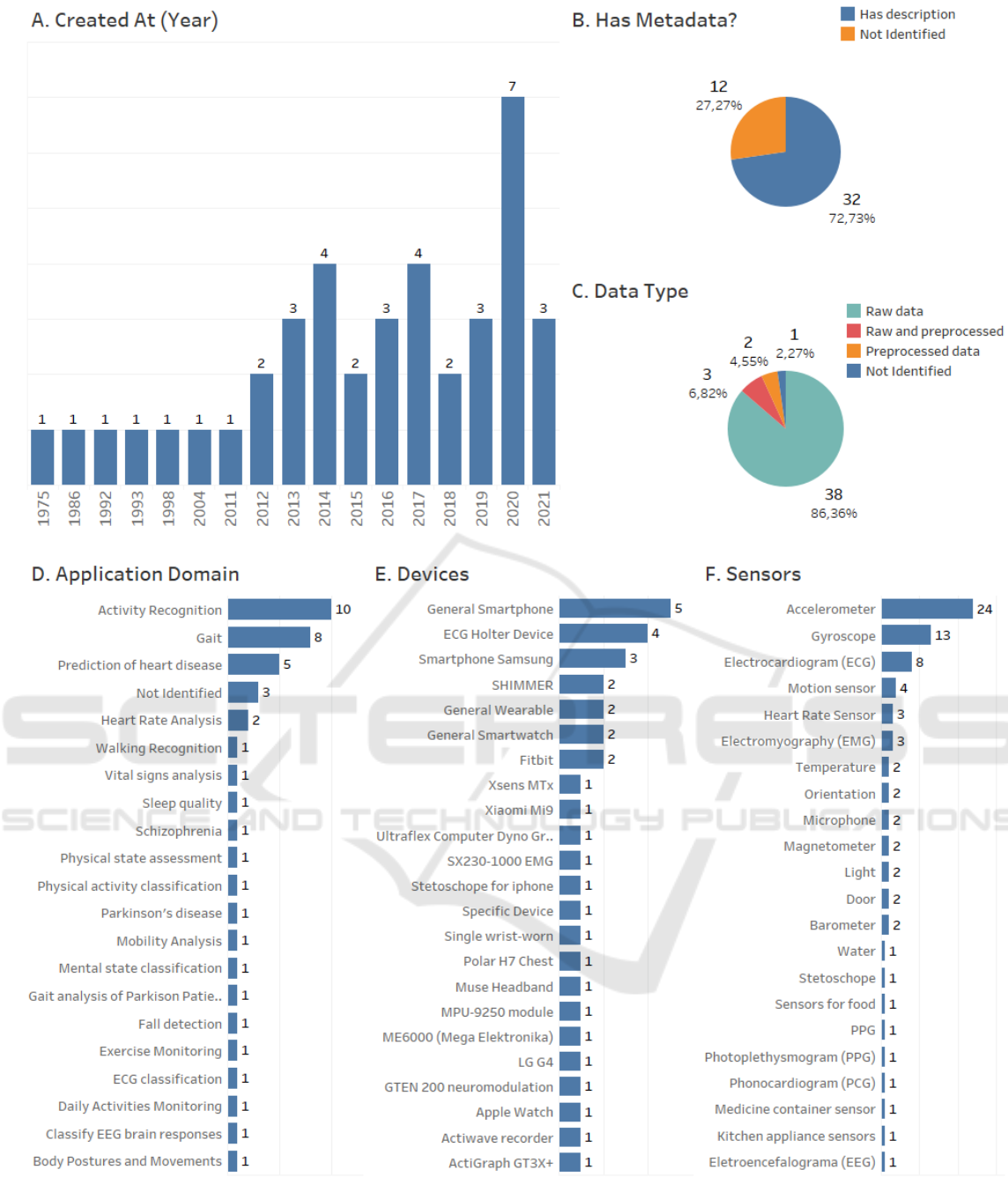

Figure 1: Dashboard summarizing the characteristics of the IoHT datasets.

those items suitable to answer our research questions. In this case, forty-four (44) data repositories were selected. It is noteworthy that the 483 items initially collected do not directly relate to the final number, because some articles may have links to one or more repositories. Moreover, other papers may not describe which repositories were used. Also, in the grey literature, we find several links to empty data warehouses.

Figure 1 exposes the characteristics of the datasets found considering the year of creation (A), whether they have described metadata (B), data type (C), application domain (D), devices (E), and sensors (F) used in the data collection.

It is noteworthy that although we selected a few repositories of grey literature at the end of this review, three of these repositories (Kaggle, Physionet,

and UCI[7]) present a large number of datasets, many of which have data obtained using IoT sensors for health applications. We argue that the construction of these repositories indicates the growing interest of the scientific community in sharing and providing subsidies for studies of new healthcare solutions. However, many of the datasets present in these repositories lack descriptions of their data and how to use them. All the details about the data repositories found in this study are available through the link bit.ly/3oXMSgh.

Regarding RQ1, the rationale was to find IoHT datasets in order to discuss how they are organized. We found forty-four (44) data repositories. Most of these data repositories were created in recent years, but we found some even before the advent of the Internet of Health Things. This situation occurs because the data were collected by long-standing devices such as ECG (electrocardiogram) sensors. In four data repositories, it was impossible to identify the creation year. Most repositories provide raw data (86%) and have a meta-data description (72%).

Concerning the application domain, the three domains of most significant interest were activity recognition (10 repositories), gait analysis (8 repositories), and prediction of heart disease (5 repositories). This aspect (application domain) is directly related to the devices and sensors used in most data collections. Usually, it is used smartphones or wearables to collect data from accelerometers (24), gyroscopes (13), and electrocardiogram (8) sensors.

Most datasets found use low-cost IoT devices, smartphones, or wearables to collect data. These devices collects data in different environments, not restricting the participants of the experiments that had their data collected to compose the bases to specific environments. In addition, many of the devices are low-cost and collect data from different sensors simultaneously, thus allowing a correlation to be made between the different types of data with the health status of the participants in the experiments.

Moreover, much of the information in the found datasets was obtained using sensors of a more generalist nature, which are not directly aimed at collecting health data, such as accelerometers, gyroscopes, and environmental sensors, such as smoke and lighting sensors. For this, it needs to have a suitable categorization to identify which health issues, or health profiles, are characterized by the data from these sensors. In this sense, it would be interesting for future work to use semantics that allow a clear understanding of how the data can be used and how it is possible to correlate these data to health status.

- **RQ1:** What are the public existing IoHT datasets?

  Summarized answer: it was identified 483 studies in scientific and grey literature from which we have selected 44 data repositories that have raw or pre-processed data from IoT sensors to characterize information related to health monitoring.

  Figure 1 presents the main characteristics of these datasets. In addition, we can also highlight some application domains found. Namely, prediction of heart disease, Gait Recognition, Fall detection, Activity Recognition, Parkinson's disease, Classification of Body Postures and Movement, Schizophrenia, Mental state classification, ECG classification, Sleep and Exercise Monitoring.

  Most datasets present metadata from the set rather than the data, with no provenance description. Some datasets highlight pre-processing data but do not indicate which techniques were used.

Although it is possible to find sensor data for multiple healthcare application domains, we have seen that there are still many limitations that make it challenging to use this data broadly. Among the main limitations identified in this study, we highlight the lack of standard regarding the number of instances, the high heterogeneity in data storage formats, the absence of Application Program Interfaces (APIs) or query tools for on-demand access to data repositories, and, finally, the lack of details about the data collection context (device specification, frequency, accuracy, environment and subjects characteristics).

Regarding data storage formats, we found many different types (*e.g.*, CSV, TXT, DAT, JSON). Unfortunately, the internal organization of these datasets does not follow a standard either. Thus, this makes data processing and integration difficult. Another challenge related to accessing data repositories is the absence of APIs or query tools. Usually, most data repositories have only the download option, which can be negative in the case of large datasets.

Concerning data repository metadata, 32 repositories (72%) have description. However, such descriptions still lack details about the context of the collection. For example, it is essential to know the specification of devices, collection frequency, and accuracy to ensure the correct use of the repository. The dataset, namely "User Identification From Walking Activity"[8] from UCI presents a suitable detail of the collection procedure, participants, and storage structure. However, repository do not show the characteristics (such as smartphone hardware detail, sensor precision, data collection frequency) of the sensors used.

In addition to the lack of standards regarding the

[7]UCI website: https://archive.ics.uci.edu.

[8]Daily and Sports: archive.ics.uci.edu/ml/datasets/User Identification From Walking Activity

number of instances, heterogeneity in formats, and the absence of APIs, data quality can be another limiting factor for the use of datasets. For example, we did not identify any standard regarding the sensors and frequencies for data collection. Furthermore, we did not find any reference to measure the quality of data available in the repositories found.

Considering this context, we reinforce that semantics can help improve existing datasets and build datasets in the future. Thus, studies addressing the construction and use of semantics in datasets of IoT sensors for healthcare are promising.

Another point to be highlighted is the profile of the participants used in the experiments or case studies where the data that make up the datasets we found were collected. We identified the number of participants in just over 81% of the datasets (36 datasets). Still, not all of these datasets presented a profile for the participants of the experiments or case studies. In most cases, the only characteristics of the participants in these studies are the identification of sex and age. A possible reason for this is the need to anonymize the data.

Furthermore, depending on the use of the data in the original study, there is no need for a more detailed characterization of the participants.

However, other characteristics do not make data anonymization impossible, such as height, weight, or even the position in which the sensors, when wearables, were located during collection. In this sense, a challenge to be addressed in future work is related to what types of user profile information that do not affect the privacy and anonymization of data should be interesting for different types of application domains focused on health. In addition, this kind of information can support the reuse of data repositories in further in-deep investigations.

- **RQ2:** What are the limitations of the existing Internet of Health Things datasets?

  Summarized answer: each study uses its dataset obtained under different conditions. One of these conditions concerns the number of samples or instances. As a result, the datasets found have varied instances, and almost half of the datasets do not provide the number of instances available.

  This variability in dataset characteristics can reflect on the performance of the algorithms, generating different results for the performances declared in studies of the same concentration area, such as, for example, gait recognition.

  Furthermore, another limitation is the heterogeneity of available formats such as CSV, JSON, ZIP, DAT, and TXT, which requires that applications or systems that want to use different datasets to implement wrappers to acquire the data. Application Program Interfaces, tools, or query languages are unavailable for data access.

  Another limiting factor is the lack of provenance of the collected data. The metadata provided are descriptions of the dataset and not about each detected data, making tracking and use by analytics and recognition applications difficult.

Finally, in RQ3, we investigated the relevant technologies for these IoHT datasets. Again, we found many different items, but the most common were smartphones and wearables with accelerometers, indicating that there is still room for developing and using new IoHT devices. For this, barriers such as the difficulty of hardware miniaturization, energy supply, and user engagement must be overcome.

The extensive use of wearables and smartphones is related to the low cost, improvements in the quality of the sensors and the fact that they do not limit the participant's mobility, unlike fixed smart objects in the environment, or require manipulation by experts. Furthermore, there is also the possibility of collecting multiple data simultaneously by these mobile devices. Therefore, we assume that the number of collections using these sensors tends to grow.

- **RQ3:** What technologies are relevant in creating and querying this kind of data sources?

  Summarized answer: ECG Holter Device, Fitbit, LG G4, Polar H7 Chest Sensor with Elite HRV, Samsung Galaxy, SHIMMER Sensor, Xiaomi Mi9, and Xsens MTx are some of the devices found in the review that gain prominence as cutting edge technologies in the data acquisition stage and creation of the datasets. We can also highlight the sensors used in existing datasets in healthcare IoT applications. They are: Accelerometer, Electromyography (EMG), Gyroscope, Magnetometer, Motion, Water, Door, Light, Temperature, and others.

4 GUIDELINES FOR IoHT DATASETS

Based on the results found in our Multivocal Literature Review, we present in this section some guidelines that we identified for building sensor datasets for use on Internet of Health Things applications. We argue that these guidelines can enhance IoHT data repositories' quality, promote their use in different studies and/or applications, and reduce the data silos issue. Figure 2 links these guidelines together to reinforce their relevance in the IoHT data sharing process.

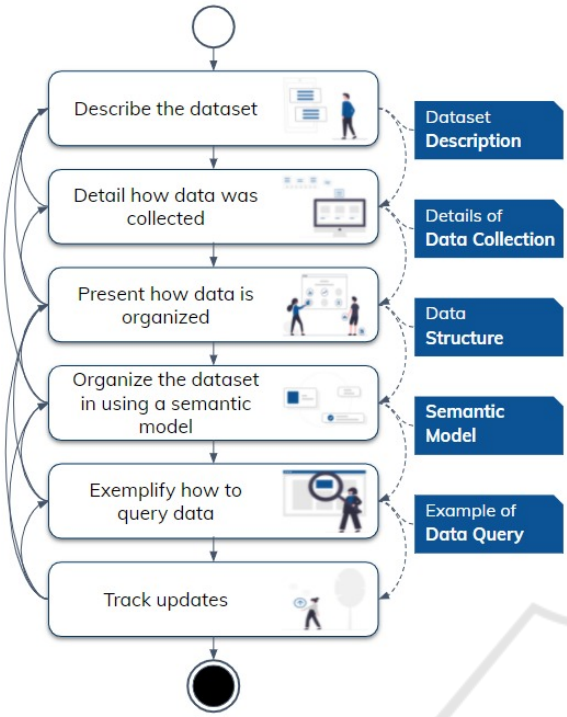

Figure 2: Guidelines for IoHT datasets.

4.1 Describe the Dataset

After analyzing the 44 datasets founded, we noticed that not all have a suitable description. For instance, some of them do not have the information of when the dataset was built, the data context, or even how to cite the dataset.

Therefore, to support the better usage of IoHT datasets, it is essential to have a minimal set of information about the dataset. We suggest the following general information:

- dataset title,
- creation and last update,
- dataset main goal, and
- how to cite the dataset.

The dataset title allows identifying the dataset. In addition, the creation date and the last update date allow identifying a time frame of the dataset, which can help understand the nature of some data present in the dataset. Furthermore, understanding the purpose of the dataset creation and the purpose of the data presented in it is essential to understand what the dataset's data characterize and, therefore, for what purpose they should be used. Finally, it is essential to specify how the dataset should be referenced in scientific publications of studies that use it.

Other information such as papers that have already used this dataset or related datasets could also be interesting to the researchers.

An example of a dataset that maintains these types of information for reference is the *MIT-BIH Arrhythmia dataset* (Moody and Mark, 2001), which was created in 1975 and updated until 2018 and has as its primary objective the Prediction of heart disease. Also, the *Modulation of Plantar Pressure and Muscle During Gait dataset* (Moriguchi et al., 2018) is another good example. This dataset was created in 2018 and, as its title claims, aims to analyze the Modulation of Plantar Pressure using gait data.

4.2 Detail How Data Was Collected

Another essential piece of information to enable data reuse is how the data was collected. This information is needed since the researcher or practitioner that will use the dataset (*i.e.*, dataset consumer) should know how the data was collected, which sensors were used, whether the sensors were calibrated or not. Furthermore, such information is helpful to support the data analysis and discussion of the conclusions obtained from the data. Hence, regarding the raw data, we argue that the datasets should describe at least the following items:

- how the data was collected,
- when the data was collected,
- detail which sensors and devices were used,
- discuss the quality of the sensors/devices, and
- describe the participants' profile.

Presenting the way the data was collected, the profile of the participants, which sensors and devices were used in the collection, and identifying some information about the sensor's quality, such as frequency used, allows the experiments carried out for the data collection to be replicated. Moreover, this information set allows dataset consumers to identify if they can use the data present in the dataset in their work. Furthermore, knowing when the data were collected can support studies that need temporal information for some of their goals.

The UMAFall (Santoyo-Ramón et al., 2018; Casilari et al., 2017) and the HuGaDB (Chereshnev and Kertész-Farkas, 2017) are examples of datasets that present information as proposed in this guideline.

The UMAFall is a dataset that contains data used to characterize activity daily living and falls. For data collection, accelerometers present in cellular devices (LG G4 and SAMSUNG S5) and accelerometer, gyroscope, and magnetometer present in an MPU-9250

module were used. These data were collected between 2016 and 2017 in experiments performed with 19 men and women aged between 19 and 67 years. This type of dataset can be used for fall classification and detection studies as (Saha et al., 2018; Junior et al., 2021).

HuGaDB is a dataset used to characterize gait patterns. Data collection was performed using specific devices containing accelerometer, gyroscope, and electromyogram sensors. These data were collected in experiments performed with 18 men and women aged 18 and 35. This dataset can be used for studies about gait patterns as (Qiu et al., 2018; Sun et al., 2020)

4.3 Present How Data Is Organized

During this study, we also observed different ways of data organization within the dataset. For example, the dataset named GP Data Analysis and ML[9] that was found in our review has a CSV file with accelerometer data. However, there is no description of the relationship of these data with the problem (in this case, gait analysis). On the other hand, the dataset named Modulation of Plantar Pressure and Muscle During Gait[10] (also found in our review) has a detailed description of the data collection and data files structure. We highlight that this organization affects the understanding and usage of the dataset.

In this scenario, we identified that it is essential to provide a clear data organization to leverage the data reuse by others researchers. Thus, we propose the two specific points: i) to detail how the data is organized in the dataset, and ii) to discuss relationships among data and health.

The latter is needed, for instance, since a set of streaming accelerometer data may be related to a specific type of movement.

4.4 Organize the Information Present in the Dataset using Semantics

As a result of our review, we observed the lack of data semantics and semantic technologies, such as ontologies, representing concepts semantically. As a result, different devices capture and make available data, often characterized by similar concepts.

Through standard vocabularies, it is possible to represent concepts obtained from heterogeneous sources and allow the interoperability of systems and platforms. The authors (Malik and Malik, 2020), for example, reinforce that the use of semantic web technologies in IoT is an emerging technology that can be used to address concerns in the healthcare domain, such as data interoperability.

Furthermore, ontologies provide semantics representation about the dataset construction process, describing, for example, algorithms used in noisy data cleaning and uncertainty handling (Elsaleh et al., 2020). Considering this context, ontology catalogs for IoT, such as the LOV4IoT[11] can be an opportunity for reuse and modeling for new and existing datasets (Venceslau et al., 2019).

4.5 Exemplify How to Query Data

Most of the datasets found provide data for download, and in a few cases, an API or own script is provided. However, some data provided in a columnar format often does not define its usefulness and purpose in the application scenario.

We have faced a scenario of little or no semantic representation of concepts and their relationships with other data. It would be interesting to present examples of how to query the data, facilitate the users' understanding of how to use the dataset, and use semantic technologies, particularly queries and their results. Furthermore, the download option can not be suitable for repositories with extensive datasets. The ideal would be to allow a data stream through APIs. In the literature, it is possible to find works (Mohammed and Fiaidhi, 2021) that seek to tackle the challenges of structuring and facilitating access to a patient's heterogeneous data record using knowledge graphs with the Neo4J tool[12].

To conclude, we can highlight as good examples the repositories hosted on the Kaggle, as it is possible to create code notebooks to access, process, and analyze such information within the Kaggle platform. Thus, there is no need to consume local disk space.

4.6 Track Updates

Finally, our last guideline is related to update tracking. Usually, the data repository is often updated from time to time. Therefore, it is essential to identify what has been added, updated, or removed to ensure that the data repository has maintained its consistency. In addition, temporal information is, in many cases, highly relevant data for studies, which is why it is also essential to identify the changes that occur in the dataset, portraying which data were affected and the possible addition of concepts.

[9] github/abdallahkhairy/GP-Data_Analysis_and_ML

[10] https://physionet.org/content/plantar/1.0.0

[11] LOV4IoT website: http://lov4iot.appspot.com.

[12] Neo4J website: https://neo4j.com.

As a result, it was possible to observe that the datasets propose the update dates but do not portray in representation models or files which concepts and the number of samples were affected. This aspect can cause divergences in the treatment of new data by the applications and make it difficult to compare studies since each study uses the data set obtained under different conditions. Thus, proposals that aim to annotate the data as it is acquired and processed can guarantee this information about its origin, facilitating the detection of changes (Elsaleh et al., 2020).

5 VALIDITY THREATS

As an empirical study, this work contains some threats to validity and, according to (Kitchenham et al., 2009), it is fundamental to identify and mitigate them. This section, thus, presents and discusses our work validity threats and how they were mitigated.

Even considering a review protocol and looking at academic and grey literature, we cannot guarantee that all sensor data and health records datasets were identified. The reasons for that are the following: i) it is common to find papers that do not present what datasets were considered for the study; ii) the paper that considers a specific data repository can not be indexed by the search sources selected in this systematic review; and, iii) the dataset and the papers using it are not achieved by the string search applied.

To mitigate these threats, we performed a snowballing process beyond the systematic search in datasets to amplify the search. Also, regarding the search string, it was defined based on several keywords related to sensors, IoT, health, and datasets.

Moreover, it is essential to highlight that we also searched four widely used databases (Scopus, Web of Science, Compendex, and PubMed) and other three sources for grey literature (Google, GitHub, and Archive). These data sources were selected based on their representativeness. We argue that Scopus, Web of Science, Compendex, and PubMed contain the most relevant studies for the IoHT area. Google, GitHub, and Achieve, in turn, contain most relevant grey literature regarding many different subjects.

Some datasets are also grouped in large data repositories, such as Kaggle, Physionet and UCI. The latter, for instance, embraces at least 588 different datasets. In these cases, we applied our search string in order to filter the number of datasets that should be manually analyzed. Since we applied our search string, our keywords cannot reach some related dataset. However, we reinforce that we tested our keywords to define a suitable search string.

Lastly, most of the repositories retrieved do not follow a systematic presentation of their information. Hence, we manually extracted the data from each dataset. However, this was needed since there is a lack of a standard for presenting the data information. For instance, some dataset clearly states how the data was collected while others do not. In this case, to improve the confidence of the results, we reviewed the extracted data with the support of four researchers.

6 RELATED WORK

This section briefly reviews works related to ours, such as reviews, surveys, or presentations of different public datasets on health applications. In this sense, considering that this study is motivated by the need to present an overview of IoHT datasets, we also review papers that present public datasets using sensors.

In the work proposed by (Cohoon and Bhavnani, 2020), the authors address types of datasets produced from digital health technologies, analytical methods, and how they can better translate the interpretation of these findings into patient care. In this perspective, the authors report public datasets and their applications in artificial intelligence algorithms. For example, the PTB Diagnostic ECG Dataset is an open-access dataset with 549 ECGs (Electrocardiograms) from 290 patients. Applying a Convolutional Neural Network (CNN) to this dataset, it was possible to detect, for example, myocardial ischemia in patients (Strodthoff and Strodthoff, 2019). In another application, paired ECGs and echocardiograms from nearly 45,000 patients at the Mayo Clinic were used to train a CNN to identify a left ventricular ejection fraction of less than 35% of the ECG data alone (Attia et al., 2019). The study explores public health digital datasets within applications that use ECG data. However, the authors did not conduct a multi-vocal review on IoHT datasets.

The work proposed by (Shuja et al., 2021) presents a survey that provides a discussion of COVID-19 open-source datasets and efforts to promote extension, validation, and scientific collaboration. In addition, the authors compare scientific papers accompanied by open-source code and data for providing future research guidance, highlighting the challenges and opportunities for missing or limited datasets. The authors present the results through a taxonomy, identifying the main characteristics of open-source datasets in terms of their type, applications, and methods. Similar to our approach, the authors present investigations from the literature and use two repositories, GitHub and Kaggle, for datasets on

domains of health applications. However, our scenario is considered more comprehensive, since we apply for a multi-voice review in healthcare applications that use sensors in data acquisition. Therefore, our study encompasses, in addition to papers, other public data repositories available on the internet.

In (Igual et al., 2015), the authors discuss the fall detection rates presented by different studies and the difficulty in comparing different fall detection studies, since each study uses its dataset obtained under other conditions. Then, using different publicly available datasets, the authors propose an investigation to determine whether the datasets influence reported performances. As a result, the authors argue that the performances of fall detection techniques are affected, to a greater or lesser degree, by the specific datasets used to validate them. Furthermore, they conclude that dataset characteristics also influence performance, while the algorithms seem less sensitive to sampling frequency or acceleration interval. Our proposal also includes public datasets related to healthcare. Therefore, it is possible to notice that our proposal can be used for different applications to compare public datasets and their influence on the performances presented in the literature.

7 FINAL REMARKS

IoT brings advances in many domains, for example, Healthcare, which has been benefited from smart things that support health data collection. This technology can be used, for example, to monitor patients, to detect and prevent falls, and support better decisions. While developing IoHT solutions, researchers and engineers often create their dataset or try to use a public one. However, they face two problemas as follows. The former requires the knowledge of the sensor data and how to collect and store them. The latter, in turn, is not easy to find.

Thus, this paper presents the results of a Multivocal Literature Review aiming to identify and characterize the existing datasets with health data collected by sensors. As a result, we found 44 datasets that match this criterion and we classified them regarding metadata, devices, domain, and data types.

Furthermore, by exploring these datasets, we perceived lack of standards and the essential information to their use by other researchers and engineers. Hence, we also discuss practices that could be used by the datasets provided in order to increase their understanding and usage by the third party.

For future work, we intend to build new health datasets using the proposed guidelines. We will also analyze different collecting methods to extend our guidelines for how the data is collected. Moreover, proposing semantics for health datasets is another future direction. Lastly, we intend to detail the process of organizing our Fall detection database, presented in (Linhares et al., 2020), following the guidelines proposed in this paper.

8 CODE AND DATA AVAILABILITY

All data used in this investigation are available on the Internet to ensure the its reproducibility and allow future in-deep analysis.

- **Protocol**: https://bit.ly/3loAjcX
- **Raw Data** (databases): bit.ly/3oXMSgh
- **Enlarged images**: https://bit.ly/3I3QC8T

ACKNOWLEDGMENTS

The authors would like to thank CNPQ (Brazilian National Council for Scientific and Technological) for the Productivity Scholarship of Rossana Maria de Castro Andrade DT-1 (N^o 306362/2021-0).

REFERENCES

Aghaei Chadegani, A., Salehi, H., Yunus, M., Farhadi, H., Fooladi, M., Farhadi, M., and Ale Ebrahim, N. (2013). A comparison between two main academic literature collections: Web of science and scopus databases. *Asian Social Science*, 9(5):18–26.

Archambault, É., Campbell, D., Gingras, Y., and Larivière, V. (2009). Comparing bibliometric statistics obtained from the web of science and scopus. *Journal of the American society for information science and technology*, 60(7):1320–1326.

Ashton, K. (2009). That "internet of things", in the real world things matter than ideas. *RFID Journal*.

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., et al. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature medicine*, 25(1):70–74.

Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: A survey. *Computer networks*, 54(15):2787–2805.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review process within the software

engineering domain. *Journal of systems and software*, 80(4):571–583.

Casilari, E., Santoyo-Ramón, J. A., and Cano-García, J. M. (2017). Umafall: A multisensor dataset for the research on automatic fall detection. *Procedia Computer Science*, 110:32–39.

Chereshnev, R. and Kertész-Farkas, A. (2017). Hugadb: Human gait database for activity recognition from wearable inertial sensor networks. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 131–141. Springer.

Cohoon, T. J. and Bhavnani, S. P. (2020). Toward precision health: applying artificial intelligence analytics to digital health biometric datasets. *Personalized Medicine*, 17(4):307–316.

Elsaleh, T., Enshaeifar, S., Rezvani, R., Acton, S. T., Janeiko, V., and Bermudez-Edo, M. (2020). Iotstream: A lightweight ontology for internet of things data streams and its use with data analytics and event detection services. *Sensors*, 20(4):953.

Garousi, V., Felderer, M., and Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology*, 106:101–121.

Igual, R., Medrano, C., and Plaza, I. (2015). A comparison of public datasets for acceleration-based fall detection. *Medical engineering & physics*, 37(9):870–878.

Islam, S. R., Kwak, D., Kabir, M. H., Hossain, M., and Kwak, K.-S. (2015). The internet of things for health care: a comprehensive survey. *IEEE Access*, 3:678–708.

Junior, E. C., Andrade, R. M., Rocha, L. S., Taramasco, C., and Ferreira, L. (2021). Computational solutions for human falls classification. *IEEE Access*.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15.

Linhares, I., Andrade, R., Costa Junior, E., Oliveira, P. A., Oliveira, B., and Aguilar, P. (2020). Lessons learned from the development of mobile applications for fall detection. In *GLOBAL HEALTH 2020*, pages 18–25.

Malik, N. and Malik, S. K. (2020). Using iot and semantic web technologies for healthcare and medical sector. *Ontology-Based Information Retrieval for Healthcare Systems*, pages 91–115.

Meskó, B. (2014). *The guide to the future of medicine: technology and the human touch*. Webicina kft.

Miloslavskaya, N. and Tolstoy, A. (2016). Big data, fast data and data lake concepts. *Procedia Computer Science*, 88:300–305.

Mohammed, S. and Fiaidhi, J. (2021). The road map of building e-diagnostics services using neo4j graph connectivity and analytics for the internet of healthcare things (ioht). *International Information Institute (Tokyo)*, 24(2):93–106.

Moody, G. B. and Mark, R. G. (2001). The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50.

Moriguchi, M., Maeshige, N., Ueno, M., Yoshikawa, Y., Terashi, H., and Fujino, H. (2018). Modulation of plantar pressure and gastrocnemius activity during gait using electrical stimulation of the tibialis anterior in healthy adults. *Plos one*, 13(5):e0195309.

Oliveira, P. A. M., Andrade, R. M. C., Neto, P. S. N., and Oliveira, B. S. (2022). Internet of health things for quality of life: Open challenges based on a systematic literature mapping. In *15th International Conference on Health Informatics (HEALTHINF)*. INSTICC.

Pai, M., McCulloch, M., Gorman, J. D., Pai, N., Enanoria, W., Kennedy, G., Tharyan, P., and Colford Jr, J. M. (2004). Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National medical journal of India*, 17(2):86–95.

Qiu, S., Wang, Z., Zhao, H., Liu, L., Li, J., Jiang, Y., and Fortino, G. (2018). Body sensor network based robust gait analysis: Toward clinical and at home use. *IEEE Sensors Journal*.

Rodrigues, J. J., Segundo, D. B. D. R., Junqueira, H. A., Sabino, M. H., Prince, R. M., Al-Muhtadi, J., and De Albuquerque, V. H. C. (2018). Enabling technologies for the internet of health things. *IEEE Access*, 6:13129–13141.

Saha, S. S., Rahman, S., Rasna, M. J., Zahid, T. B., Islam, A. M., and Ahad, M. A. R. (2018). Feature extraction, performance analysis and system design using the du mobility dataset. *IEEE Access*, 6:44776–44786.

Sandhu, R. S. and Samarati, P. (1994). Access control: principle and practice. *IEEE communications magazine*, 32(9):40–48.

Santoyo-Ramón, J. A., Casilari, E., and Cano-García, J. M. (2018). Analysis of a smartphone-based architecture with multiple mobility sensors for fall detection with supervised learning. *Sensors*, 18(4):1155.

Selvaraj, S. and Sundaravaradhan, S. (2020). Challenges and opportunities in iot healthcare systems: a systematic review. *SN Applied Sciences*, 2(1):139.

Shuja, J., Alanazi, E., Alasmary, W., and Alashaikh, A. (2021). Covid-19 open source data sets: a comprehensive survey. *Applied Intelligence*, 51(3):1296–1325.

Strodthoff, N. and Strodthoff, C. (2019). Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiological measurement*, 40(1):015001.

Sun, F., Zang, W., Gravina, R., Fortino, G., and Li, Y. (2020). Gait-based identification for elderly users in wearable healthcare systems. *Information Fusion*, 53:134–144.

Venceslau, A., Andrade, R., Vidal, V., Nogueira, T., and Pequeno, V. (2019). Iot semantic interoperability: a systematic mapping study. In *ICEIS*, pages 535–544.

Weiser, M. (1999). The computer for the 21st century. *ACM SIGMOBILE mobile computing and communications review*, 3(3):3–11.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, page 38. ACM.