Pattern Recognition Letters 146 (2021) 46–54

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

KU-HAR: An open dataset for heterogeneous human activity recognition

Niloy Sikder[a], Abdullah-Al Nahid[b,*]

[a] Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh
[b] Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208, Bangladesh

ARTICLE INFO

Article history:
Received 30 June 2020
Revised 15 February 2021
Accepted 19 February 2021
Available online 17 March 2021

ABSTRACT

In Artificial Intelligence, Human Activity Recognition (HAR) refers to the capability of machines to identify various activities performed by the users. The knowledge acquired from these recognition systems is integrated into many applications where the associated device uses it to identify actions or gestures and performs predefined tasks in response. HAR requires a large quantity of meticulously collected action data of diverse nature to fuel its learning algorithms. This paper aims to introduce a new set of HAR data collected from 90 participants who are 18 to 34 years old. The constructed dataset is named KU-HAR, and it contains 1945 raw activity samples that belong to 18 different classes. We used built-in smartphone sensors (accelerometer and gyroscope) to collect these HAR samples. Apart from the original (raw) time-domain samples, the dataset contains 20,750 subsamples (extracted from them) provided separately, each containing 3 seconds of data of the corresponding activity. Some classification results have been provided as well to propound the quality of the proposed dataset. The acquired results show that Random Forest (RF), an ensemble learning algorithm, can classify the subsamples with almost 90% accuracy. This dataset will enable smartphones and other smart devices to identify new activities and help researchers to design more delicate models based on practical HAR data.

1. Introduction

At this moment, there are more than 3.5 billion smartphone users in the world; and the number is expected to reach four billion within a couple of years [1]. According to the latest statistics, 77% of US citizens own at least one smartphone, and the number is as high as 95% in South Korea, which is the leading country in smartphone ownership [2]. No other device in the history of technology has gained such popularity among people of all age groups in so little time. Modern mobile phones are capacious, powerful, multipurpose, equipped with various sensors, and even though they perform complex operations underneath, they are relatively easy to use. Today's mobile phones are "smart" because they can automate stuff, sense their surroundings, and respond accordingly. They can also learn from their users' behavior and use that knowledge to make educated guesses or suggestions. Many of these traits are possible because of the recent developments in the field of Artificial Intelligence (AI). Today, most smartphone applications (apps) are dynamic and have several intelligent algorithms working behind their seemingly modest user interface. The learning process of smartphones is heavily dependent on data collected from user inputs and a wide range of built-in sensors. Unless restricted or directed otherwise, a modern smartphone continuously collects auditory, visual, and sensory data in the form of numerals, signals, images, and videos. These data are then used to understand the key properties of the surrounding environment. But before that, the gathered data require some level of pre-processing since real data tend to be noisy. Fortunately, a wide range of Digital Signal Processing (DSP) and Digital Image Processing (DIP) techniques are available to carry out these pre-processing operations. Then comes the stage where powerful Machine Learning (ML) algorithms are put into action to learn from the collected data. This learning process enables smartphones to understand the users' needs more explicitly and become a more useful companion.

Human Activity Recognition (HAR), also known as Human Action Recognition, is a perfect example of modern smartphones' adaptive capacity. Apart from smartphones, any digital device such as tablets, smartwatches, or MP3 players with specific sensors installed can detect human movements and gestures and perform a set of tasks in response. Activity signal-based HAR usually takes two specific sensors' readings into account, namely accelerometer

* Corresponding author.
E-mail address: nahid.ece.ku@ku.ac.bd (A.-A. Nahid).

https://doi.org/10.1016/j.patrec.2021.02.024
0167-8655/

and gyroscope. An accelerometer measures the linear acceleration of its movements along the three axes (namely, X, Y, and Z). A gyroscope measures the angular rotational velocity along those axes. These properties are recorded in terms of their rate of change. If the smartphone is attached to or being carried by a person, these readings can be used to describe his/her movements — this idea forms the basics of activity recognition using smartphone sensors. However, readings from other sensors such as magnetometer and Global Positioning System (GPS) can also be incorporated for extra precision. Earliest works in HAR date back to the early 1980s [3]. However, because of the recent advancements in the smartphone, chip, and sensor designing technologies as well as in AI research, numerous breakthroughs in HAR research have occurred within the last two decades. As the demand for new functionalities of smartphones is higher than ever before, HAR is being researched extensively to make the devices smarter and more adaptive. However, the usability of this technology is not confined to mobile apps development. HAR is widely used in surveillance systems, automatic health-care monitoring systems, prosthetics, haptics, robotics, and many other sectors where intelligent machines are required to detect human motions. That been said, the principal goal remains the same — to correctly identify human actions or gestures based on single or multiple sensor outputs.

The activities in HAR research can be broadly divided into two categories: data collection and recognition model development. Researches that belong to the first category focus on gathering mass data of diverse activities from multiple subjects using external or built-in smartphone sensors. Upon performing some preprocessing and noise-removal operations, they provide other researchers the acquired data in a useable form. Studies that belong to the second category utilize those data and design recognition models using various DSP, DIP, and ML algorithms. This paper describes the outcomes of a research project that falls into the first category. We aim to introduce the KU-HAR dataset, which is available online and free to download, use, and modify (please see the Supplementary Material section to access the dataset). But before describing the proposed dataset, we want to present a chronological view of some popular and widely-used HAR datasets, which helped countless researchers to develop numerous activity-recognition models in the past decade.

In 2009, Yang et al. constructed a public-domain HAR dataset in the University of California Berkeley, California, based on the data collected from 20 participants [4]. Their dataset is called the Wearable Action Recognition Database (WARD), and it contains 13 action categories. The activity data were collected using a tri-axial accelerometer and a di-axial gyroscope. In 2010, Kawaguchi et al. started the HASC Challenge, and its goal was to amass large-scale accelerometer data of six human movements [5]. The HASC Challenge was not a competition but a technological challenge that gave 20 teams of volunteers a challenge to find at least five participants (each) and collect five samples of activity data (of each class) from them. At the time of publication in 2011, they accumulated 6700 activity samples from 540 different subjects.

Multiple HAR datasets were published in 2012, some of which are still considered benchmarks and are widely used to test and validate HAR models. The first one on this list is called the UCI HAR dataset and was constructed by Anguita et al. [6]. The dataset contains more than ten thousand activity samples of six different classes collected using built-in smartphone sensors from 30 participants. It is hosted by the University of California Irvine (UCI) ML repository [7]. The authors extended the dataset by adding six new types of activity signals in 2015 [8]. The added signals contain information on six different postural transitions, which refer to transitory movements or the change of body state from one posture to another. The extended dataset is also available in the UCI ML repository [9]. Published in 2012, Wireless Sensor Data Mining (WISDM) dataset presented by Kwapisz et al. is another popular HAR dataset. It contains 5424 samples collected from 29 different individuals [10]. This dataset includes only the outputs of accelerometers for six different types of activities. The Opportunity challenge developed by Chavarriaga et al. is another large-scale benchmark HAR dataset [11]. The specialty of this dataset is that the researchers included magnetometers' readings as well. And lastly, developed by Zhang and Sawchuk at the University of Southern California (USC), Los Angeles, USC-HAD is another HAR dataset that came out in 2012 [12]. Compared to the previously mentioned datasets, this one is relatively small as it contains only 840 samples collected from 14 contributors. However, the dataset has information on 12 activities, making it more diverse than the others.

Among the most recent HAR datasets, UMAFall is one of the notable ones. The dataset was published in 2017 by Casilari et al. and is used in various kinds of fall detection [13]. Similar to the WISDM dataset, it incorporates three different sensor outputs of 11 activities gathered from 17 participants. In 2018, Saha et al. formulated an open-source HAR dataset called the University of Dhaka Mobility Dataset (DU-MD) [14,15]. This sizable dataset contains 5000 activity samples rounded up from 50 subjects who performed ten different activities under observation. In 2020, Garcia-Gonzalez et al. put together another dataset incorporating four different sensors [16]. The particularity of this work is its independence of subjects and smartphone orientation. Each of these datasets is slightly different from the others regarding the activities involved, the surrounding environment, or the data collection method. In truth, human activities are complex phenomena. They vary from person to person and are hard to categorize in a finite number of classes, which is why more data of diverse activities collected from many individuals are required to advance HAR research. And that was our motivation behind undertaking this data collection project which led to the formulation and publication of the KU-HAR dataset. We wanted to collect action data of new and diverse human activities, and add more samples to the existing ones while accumulating samples for this dataset. This dataset will allow ML-based HAR models to learn from these samples and be more robust and accurate while identifying the corresponding activities.

The rest of the paper is organized in the following order. Section 2 elaborately describes our HAR data collection procedure with the necessary facts and figures. Section 3 provides insights into the uploaded KU-HAR dataset and explains how to use it to train ML models. Section 4 presents some classification results by employing a simple classification framework. Finally, Section 5 mentions the takeaways of the paper and outlines the authors' further plans regarding this dataset.

2. Data collection methodology

The activity data of the KU-HAR dataset were collected in a practical environment at Khulna University, Khulna-9208, Bangladesh. The data were collected from 90 participants (75 male and 15 female). The participants belong to a range of ages and weight classes. More statistical information on the participants is provided in Table 1. Prior to participation, each volunteer was briefed on the reason and the procedure of the data collection as well as how and for what purpose the obtained data will be used. Then they were handed a consent form containing all the rules, regulations, and potential risk factors (please see the Supplementary Material section to read the consent form). They were also provided with a withdrawal form outlining the procedure of withdrawal of their personal activity data if they wish to do so afterward (please see the Supplementary Material section). Each participant had the right to deny to perform any activity on the list

Table 1
Details on the participants who provided HAR data.

No of total participants:	90
No of male participants:	75
No of female participants:	15
Age range of the participants (yr):	18 – 34
Average age of the participants (yr):	21.7
Weight range of the participants (kg):	42.2 – 100.1
Average weight of the participants (kg):	63.2
Participants with prior heart conditions:	2

or stop the procedure involving him/her altogether. They were also instructed to immediately report any discomfort or injury during or after the session. Upon acknowledging all the rules and their rights, they put their names, student IDs (if applicable), and signatures with dates at the end of the consent form. Only after completing these formalities, a participant was enlisted in the project database. A few individuals were selected to assist the data collection process, who were briefed separately about their duties during the whole procedure. Although no such cases occurred, a first-aid box was always kept within reach, and an assistant was instructed on how to quickly respond in the event of a severe injury. The safety and willingness of the participants were given the top priority throughout the whole process.

The age and the weight of each participant, along with any known heart conditions, were recorded at the beginning of the procedure. He/she was given a unique Participant ID to keep track of the samples collected from his/her activities. Then, an assistant fastened a bag (also known as a fanny pack) around his/her waist containing a smartphone equipped with an accelerometer and a gyroscope. The smartphone was connected to a nearby laptop via a wireless network. The phone records the outputs of the sensors as (six different streams of) time-series data (tri-axial readings of two sensors) and transmits them to the laptop. We used five android smartphones in this data collection project. Models of these smartphones are: Samsung Galaxy J7 (2017), XiaomiRedmi Note 4, Realme 3 Pro, Realme 5i, and Realme C3. Fig. 1(a) shows a participant standing still wearing a waist bag while providing activity data. The waist bag is enhanced in Fig. 1(b). Fig. 1(c) shows the screen of the smartphone inside the bag while recording and transmitting activity data.

Fig. 2 illustrates a schematic diagram of the data collection process involving a smartphone, a laptop, and a permanent data repository. It also shows the smartphone's orientation (left side down, the screen was facing the same direction as the participant) while recording the HAR data. The laptop interface helped us monitor the acquired data in real-time and detect any disruption such as interruption in data collection or a break down in the wireless connectivity between smartphones and laptops. Upon collection, the data were stored in a secure database for further processing. To speed things up, several groups of assistants worked simultaneously and collected data from multiple participants. However, the entire procedure was closely monitored at all times. A hand-written sheet was strictly maintained containing information on the samples gathered by each group and from each participant.

The KU-HAR dataset contains samples of 18 different activity classes. These activities are listed and described in Table 2. Broadly, the classes can be divided into two groups—indoor activities and outdoor activities. The first 11 activities' data (Class ID 0–10) were collected indoors since they do not require a large space to perform. The participants performed these activities inside a classroom within an area around 180 square feet. The classroom is on the ground floor of the building. The volunteers monitored the collected data sitting on the seats of the classroom in front of the participants. The next four activities (Class ID 11–14) require an open area, and hence, were arranged outdoors. The corridor in front of the classroom was used to perform these activities and record their samples. The staircase used for collecting the next two classes' data (*Stair-up* and *Stair-down*) is also situated in front of the classroom and has nine flights of stairs reaching from the ground floor to the 3rd floor (three in between each floor). And finally, the *Table-tennis* data were collected in the common room (ground floor) of Khan Jahan Ali Hall, Khulna University. Table 2 also contains the duration of a typical sample and the number of total collected samples of each class. However, some activities are slightly cumbersome and require a certain level of physical fitness to perform. For those activities, the duration or repetition of each sample was kept at a reasonable range. Constant communication was maintained with

Table 2
Description of the activity classes in the KU-HAR dataset.

Class name	Class ID	Performed activity	Duration or repetitions per sample[#]	Number of collected samples	Number of extracted subsamples
Stand	0	Standing still on the floor	1 min	91	1886
Sit	1	Sitting still on a chair	1 min	90	1874
Talk-sit	2	Talking with hand movements while sitting on a chair	1 min	86	1797
Talk-stand	3	Talking with hand movements while standing up or sometimes walking around within a small area	1 min	88	1866
Stand-sit	4	Repeatedly standing up and sitting down*	5 times	339	2178
Lay	5	Laying still on a plain surface (a table)	1 min	87	1813
Lay-stand	6	Repeatedly standing up and laying down*	5 times	148	1762
Pick	7	Picking up an object from the floor by bending down	10 times	105	1333
Jump	8	Jumping repeatedly on a spot	10 times	130	666
Push-up	9	Performing full push-ups with a wide-hand position	5 times	111	480
Sit-up	10	Performing sit-ups with straight legs on a plain surface	5 times	121	1005
Walk	11	Walking 20 meters at a normal pace	≈ 12 s	188	882
Walk-backward	12	Walking backwards for 20 meters at a normal pace	≈ 20 s	50	317
Walk-circle	13	Walking at a normal pace along a circular path	≈ 20 s	35	259
Run	14	Running 20 meters at a high speed	≈ 7 s	146	595
Stair-up	15	Ascending on a set of stairs at a normal pace	≈ 1 min	53	798
Stair-down	16	Descending from a set of stairs at a normal pace	≈ 50 s	57	781
Table-tennis	17	Playing table tennis	1 min	20	458
			Total	1945	20,750

* Postural transition activity
\# Subjected to change based on the ability and willingness of the participant

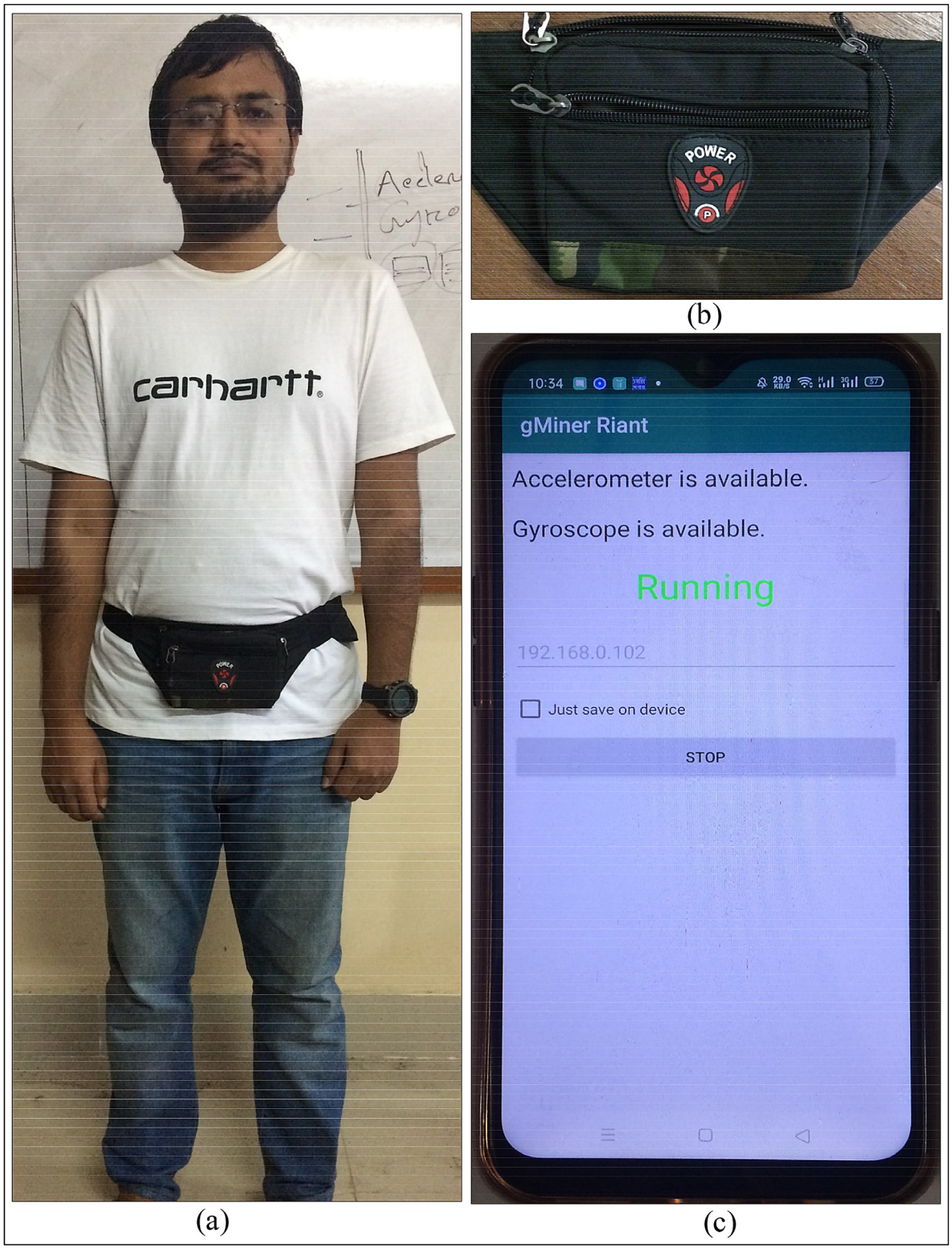


Fig. 1. (a) A participant providing HAR data, (b) a waist bag used to put the smartphone on him, and (c) the screen of the smartphone while recording.

the participants, and their willingness and capability were given the utmost importance. Activity signals of particular classes do not follow any pattern and hence are aperiodic in nature. However, class samples that involve repetition of an action form patterns and (roughly) repeat those patterns after an interval.

Fig. 3 shows the graphical user interfaces of the apps developed for smartphones and laptops to commence the data collection process. The smartphone app has two modes of operation — they can send the recorded data to the connected laptop (Fig. 1(c)), or they can store them at the local storage of the smartphone, as shown in Fig. 3(a). The first mode is useful to visualize the data using the PC software, as shown in Fig. 3(b). However, the other mode comes in handy to record specific activities, such as running and climbing up the stairs, when the participant needs to move away from the

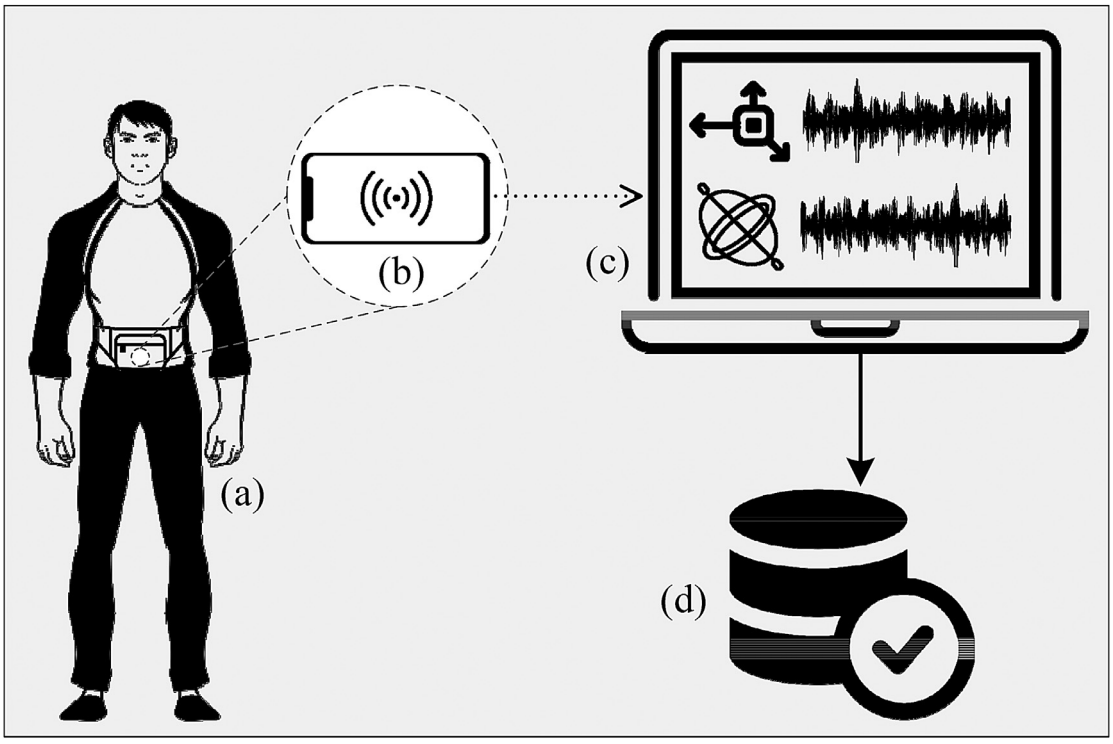


Fig. 2. A schematic diagram of the data-collection process showing (a) a subject, (b) a smartphone transmitting sensor data, (c) a laptop wirelessly connected to the phone and receiving sensor data, and (d) a permanent database to store HAR data for further processing.

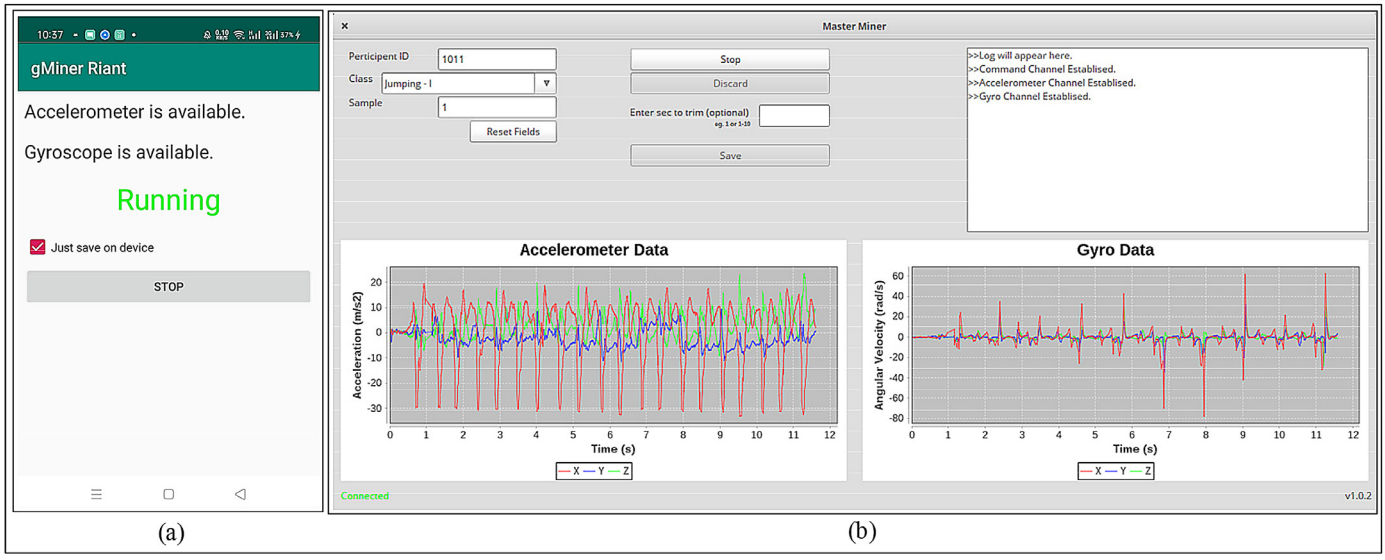


Fig. 3. The interface of the (a) smartphone app and (b) PC app used for HAR data collection.

laptop, which may disrupt the wireless connection. The PC software provides some other functionalities as well. While provided with a Participant ID, a Class ID, and a sample no., it stores the samples collected from that participant with those pieces of information. The software also has the option to trim the collected sample from the start, in case there was a time difference between the beginning of the recording and the start of the activity.

3. Description of the dataset

The KU-HAR dataset available online contains three variants of the collected HAR data. Two of them have the raw time-domain signals as they were collected from the participants using accelerometer and gyroscope sensors. The third variant contains the subsamples that were extracted from the original samples. These variants are uploaded as separate compressed files. Their contents are described below:

3.1. Raw time-domain data

This file contains the collected raw HAR signals without any form of prior processing or modification. As mentioned earlier, a total of 1945 activity samples were collected from the 90 participants involved in the study. These samples are provided individually as comma-separated value (.csv) files in 18 different directories. Each .csv file contains the tri-axial time-series sensors' data

Table 3
The contents of a .csv file containing raw data.

Column 1:	the exact time (elapsed since the start) when the accelerometer output was recorded (in ms)
Column 2:	acceleration along the X-axis (in m/s^2)
Column 3:	acceleration along the Y-axis (in m/s^2)
Column 4:	acceleration along the Z-axis (in m/s^2)
Column 5:	the exact time (elapsed since the start) when the gyroscope output was recorded (in ms)
Column 6:	rate of rotation around the X-axis (in rad/s)
Column 7:	rate of rotation around the Y-axis (in rad/s)
Column 8:	rate of rotation around the Z-axis (in rad/s)

corresponding to the activity sample. Table 3 outlines the contents of these .csv files.

3.2. Trimmed interpolated raw data

In reality, several factors force the data collection procedure to deviate from the theory and contaminate the collected samples with noise and interference. One such phenomenon is the time delay between the start of the recording and the beginning of the activity. Because of this mismatch, the data collected in the first few seconds may not come from the corresponding activity at all. The same thing can happen at the end of the sample as well. Keeping these parts in the sample can mislead the learning algorithms. Hence, we performed a trimming operation on the affected signals to remove these portions. All the collected samples were checked manually to find such occurrences and omit them. Also, we wanted to collect the samples at a constant rate (100 Hz). However, because of the differences in the smartphones' capacity, processing power, and the sensors' quality and build, the rate often fluctuated from 100 Hz. We saved each data point's exact recording time (columns 1 and 5 of each .csv file) to detect such occurrences. In this step, we used the elapsed time information to perform a one-dimensional data interpolation to make the sampling rate of all the signals a constant 100 Hz. The resultant activity samples have been provided in the corresponding .csv files. The contents of these .csv files follow the format presented in Table 3.

3.3. Time-domain subsamples

Samples collected from the participants were segmented to create subsamples to increase the number of instances in each activity class. This step is flexible and offers many options, as samples can be segmented in numerous ways using various windowing techniques and window length. We put 3 seconds' data of an activity sample into each subsample extracted from it. No windowing technique was employed during this process, which means each subsample is unique and does not exactly resemble any other portion of the original sample. Table 2 presents the number of subsamples extracted from the collected samples of each class. They amount up to a total of 20,750 activity subsamples. These subsamples have been provided as a single .csv file with relevant information on the subsamples and class IDs (defined in Table 2). Table 4 outlines the information useful to interpret this file. Each row of the .csv file represents an activity subsample. Each subsample is comprised of six different time-domain signals (tri-axial signals of two sensors). The corresponding class ID, the original length of each subsample, and a serial no. are appended at the end of each row. We encourage future researchers to use these subsamples to train ML models so that their performance is directly comparable. However, if they want to direct the segmentation operation based on other ideas or for different applications, they can use the previous sets of data.

The reason behind providing the same activity data in multiple subsets is to ensure maximum customizability of the acquired HAR data. Based on their affinity, researchers can choose to use any of these subsets for their task and decide how to process them further. Before moving to the next section, we present a quantitative comparison of the constructed dataset with some of the existing ones in Table 5.

Table 4
Details on the .csv file containing activity subsamples.

Column no.	Information
1 – 300	Accelerometer X-axis readings
301 – 600	Accelerometer Y-axis readings
601 – 900	Accelerometer Z-axis readings
901 – 1200	Gyroscope X-axis readings
1201 – 1500	Gyroscope Y-axis readings
1501– 1800	Gyroscope Z-axis readings
1801	Class ID (given in Table 2)
1802	Length of each axial data
1803	Serial no. of the subsample

4. HAR classification results and discussion

This paper's primary goal is to describe the proposed KU-HAR dataset in detail and explain how to use it. However, since the dataset was built to encourage the development of new HAR models, we want to provide some classification results involving a simple classification framework. But before moving on to the classification stage, let us take a look at Fig. 4(a), which illustrates the t-Distributed Stochastic Neighbor Embedding (t-SNE) graph of the time-domain subsamples of the KU-HAR dataset described in Section 3.3. t-SNE is a dimensionality-reduction algorithm widely used to observe data distribution in a two or three-dimensional space [17]. Suppose the same class samples are closer together, and the samples of different classes are further away from each other — in that case, the situation is considered to be ideal for classification. However, as seen from Fig. 4(a), that is not the case for this dataset. Time-domain signals are often noisy and less useful while mining for vital information, which is why certain noise removal techniques are applied to suppress the effect of noise and interference present in them. Signals recorded using smartphone sensors contain high-frequency noise components and requires to be passed through a high-order low-pass filter [6]. In general, accelerometer outputs are an aggregation of gravity acceleration and body acceleration components, the first of which is unwanted in this regard and require another filtering operation to get rid of [18]. However, in this study, the app used for data collection actively eliminated the gravity components immediately after recording the accelerometer data.

We did not perform any filtering or noise removal operation as well — it can be done following different principles or based on the requirement of the task or the preference of the researchers. We performed Fast Fourier Transformation (FFT) on the derived subsamples to extract frequency-domain features from them. This transformation separates most of the noise elements from the real activity samples. FFT was applied separately on each of the six signals of a subsample. t-SNE of the dataset's subsamples based on

Table 5
Quantitative comparison with other benchmark HAR datasets.

Dataset name	Year published	Number of classes	Number of participants	Sampling rate (Hz)	Sensors used	Number of samples
UCI HAR [6,7]	2012	6	30	50	Accelerometer and gyroscope	10,299
UCI HAPT [8,9]	2015	12	30	50	Accelerometer and gyroscope	10,929
WISDM [10,19]	2012	6	29	20	Accelerometer and gyroscope	5424
DU-MD [14,15,20]	2018	10	50	30	Accelerometer and gyroscope	5000
HASC Challenge [5]	2011	6	540	100	Accelerometer	6700
WARD [4]	2009	13	20	30	Accelerometer and gyroscope	1298
USC-HAD [12]	2012	12	14	100	Accelerometer and gyroscope	840
Opportunity [11]	2012	20	12	–	Accelerometer, gyroscope, and magnetometer	3371
UMAFall [13]	2017	11	17	200	Accelerometer, gyroscope, and magnetometer	531
Garcia-Gonzalez et al. [16,21]	2020	4	18	Varies	Accelerometer, gyroscope, magnetometer, and GPS	29,126,810
KU-HAR (proposed)	2020	18	90	100	Accelerometer and gyroscope	20,750

"–" denotes that the information was not found in the associated paper

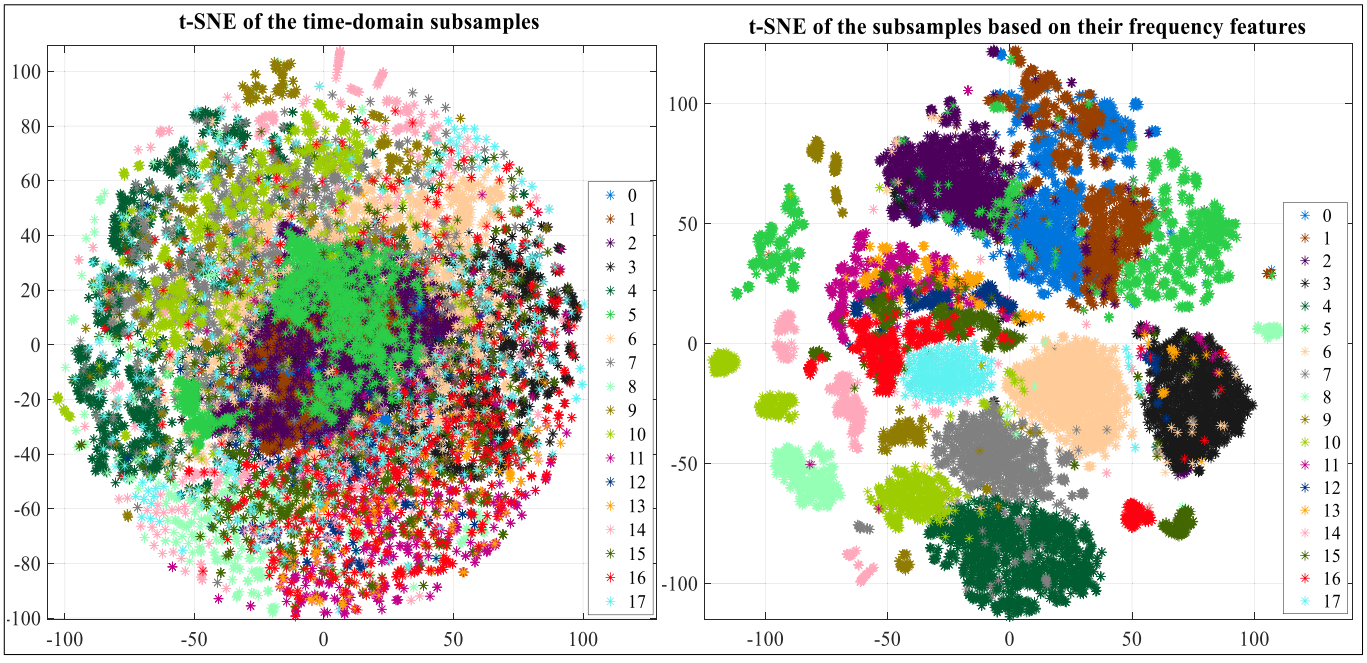
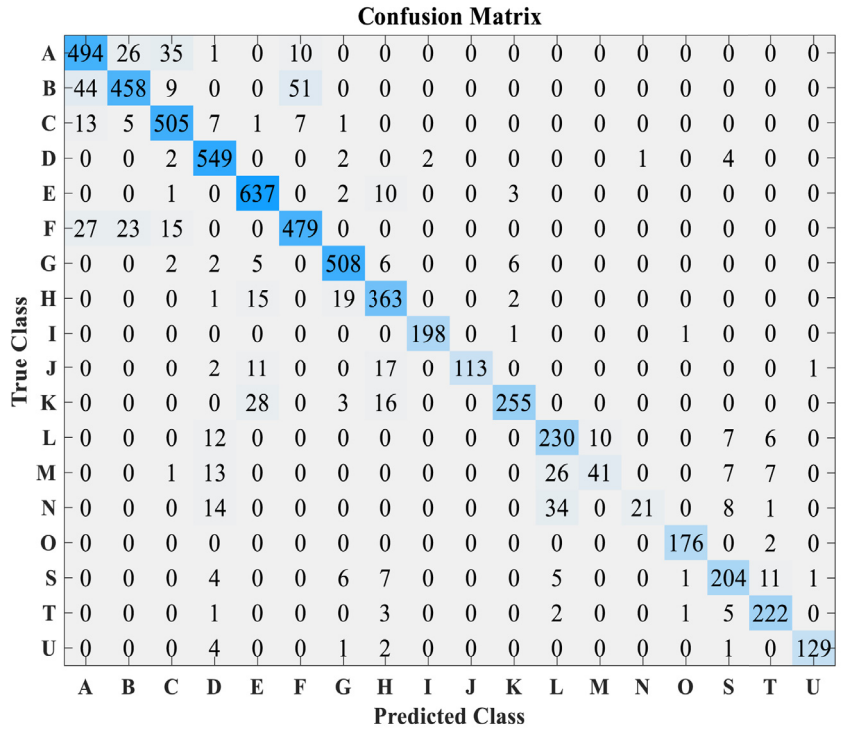


Fig. 4. The t-SNE graph of the subsamples based on (a) time-domin features and (b) frequency-domain features.

their frequency-domain features has been presented in Fig. 4(b). As seen from the figure, most of the homogenous samples are clustered together, creating a more favorable situation for the learning algorithm than the initial one. However, there are areas where samples of different classes overlap; these samples will be harder to categorize correctly by the classifier.

As mentioned in Section 1, HAR is an extensively-researched field of ML. Numerous state-of-the-art HAR classification methods are available, incorporating many shallow-learning, tree-based learning, and complex deep-learning algorithms [22]. However, in this study, we performed the classification of the activity classes using the Random Forest (RF) algorithm, a widely used supervised learning algorithm used to solve classification and regression problems. RF is a powerful ensemble learning technique that combines the outcomes of multiple Decision Trees while making class predictions following the principles of "bagging" [23]. We randomly selected 70% of the HAR subsamples (14,525 subsamples) of the devised dataset to train the RF-based classification model. Then we tested it on the remaining 6225 subsamples (30% of the dataset). The described classification model provided an 89.67% classification accuracy on the tested subsamples, which is a relatively good accuracy score considering the classification model's simplicity.

As our dataset is imbalanced in terms of class sample sizes (mentioned in Table 2), F1-score, which is not affected by class imbalance, would be a better parameter to judge its performance than the accuracy score. The model provided an F1-score of 87.59%, which closely trails the classification accuracy score, implying the achieved classification performance's reliability. Fig. 5(a) presents the confusion matrix of the preformed classification using RF. This matrix conveys more in-depth information on the classifier's success in identifying individual activity classes. The matrix shows that despite having high success rates in most of the regions, the classifier found it difficult to categorize some samples that belong to classes *Stand, Sit, Talk-sit,* and *Lay*. This happened because none of these four activities involve much mobility. A similar phenomenon occurred among the classes related to walking (i.e., *Walk, Walk-backward,* and *Walk-circle*). However, the classifier performed very well while identifying other classes' samples. The class-wise performance has been summarized in Fig. 5(b). As seen from the figure, the average precision and recall of all the 18 classes are over 91% and 84.7%, respectively. Only three classes' recall and



Class	Recall	Precision	F1-socre
Stand	0.873	0.855	0.864
Sit	0.815	0.895	0.853
Talk-sit	0.937	0.886	0.911
Talk-stand	0.980	0.900	0.938
Stand-sit	0.975	0.914	0.944
Lay	0.881	0.876	0.878
Lay-stand	0.960	0.937	0.949
Pick	0.908	0.856	0.881
Jump	0.990	0.990	0.990
Push-up	0.785	1.000	0.879
Sit-up	0.844	0.955	0.896
Walk	0.868	0.774	0.819
Walk-backward	0.432	0.804	0.562
Walk-circle	0.269	0.955	0.420
Run	0.989	0.983	0.986
Stair-up	0.854	0.864	0.859
Stair-down	0.949	0.892	0.919
Table-tennis	0.942	0.985	0.963
Average	**0.847**	**0.907**	**0.877**

Fig. 5. (a) Confusion matrix and (b) class-wise performace of the classification operation.

one class's precision are lower than 80%. Improvements have to be made in these classes in order to build better HAR models. Researchers are welcome to experiment with the dataset, extract diverse and more powerful features from its samples and subsamples, and develop cogent recognition algorithms.

5. Conclusion and future work

This paper introduced and described the KU-HAR dataset, which holds information on 18 different daily life activities accumulated using smartphone sensors. This article provided key information regarding the construction and formulation of the dataset. It also delineated all the subsets of the dataset, which is available online and free to use. Moreover, it presented a brief overview of the existing HAR datasets and provided a comprehensive discussion. Finally, we submitted the outcomes of a classification framework that involves frequency-domain features extracted from the datasets' subsamples and RF classifier. The obtained results show that the model is capable of providing an 89.67% classification accuracy and an 87.59% F1-score. We hope that other researchers will find this dataset useful, incorporate it in their studies, and explore it in new and different ways. We plan to continue collecting activity data to increase the number of samples of specific classes and incorporate new activities performed in everyday lives. Upon processing, these data will be added in the upcoming versions of the KU-HAR dataset.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the contribution of Ibrahim Rafi (student, Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208), who developed the apps (for PC and smartphone) used in this project. We cordially thank Kangkan Bhakta, Md. Al-Masrur Khan, Md. Sanaullah Chowdhury, Muslima Sultana, Faozia Rashid Taimy, Shahriar Sikder, and Md. Rashed Jaowad Khan (students, Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208) for assisting this data collection project. And lastly, we would like to express our gratitude to all the participants who helped us to construct this dataset.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2021.02.024.

References

[1] D. Georgiev, 60+ Revealing Statistics about Smartphone Usage in 2020, (2020). https://techjury.net/blog/smartphone-usage-statistics/ (accessed May 27, 2020).
[2] D. Metev, 39+ Smartphone Statistics You Should Know in 2020, (2020). https://review42.com/smartphone-statistics/ (accessed May 27, 2020).
[3] M.A. Labrador, O.D. Lara Yejas, Human activity recognition : using wearable sensors and smartphones, CRC Press, 2013.
[4] A.Y. Yang, R. Jafari, S.S. Sastry, R. Bajcsy, Distributed recognition of human actions using wearable motion sensor networks, J. Ambient Intell. Smart Environ. 1 (2009) 103–115, doi:10.3233/AIS-2009-0016.
[5] N. Kawaguchi, N. Ogawa, Y. Iwasaki, K. Kaji, T. Terada, K. Murao, S. Inoue, Y. Kawahara, Y. Sumi, N. Nishio, HASC challenge, Gathering large scale human activity corpus for the real-world activity understandings, ACM Int. Conf. Proceeding Ser. (2011), doi:10.1145/1959826.1959853.
[6] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, A Public Domain Dataset for Human Activity Recognition Using Smartphones, 2013. http://www.i6doc.com/en/livre/?GCOI=28001100131010. (accessed May 29, 2020).

[7] J. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, X. Parra, UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set, (2012). https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones (accessed April 20, 2020).
[8] J.L. Reyes-Ortiz, L. Oneto, A. Ghio, A. Samá, D. Anguita, X. Parra, Human activity recognition on smartphones with awareness of basic activities and postural transitions, Lect. Notes Comput. Sci. (Including Subser.Lect. Notes Artif.Intell. Lect. Notes Bioinformatics). 8681 LNCS (2014) 177–184, doi:10.1007/978-3-319-11179-7_23.
[9] J.L. Reyes-Ortiz, D. Anguita, L. Oneto, X. Parra, UCI Machine Learning Repository: Smartphone-Based Recognition of Human Activities and Postural Transitions Data Set, (2015). http://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions (accessed May 20, 2019).
[10] J.R. Kwapisz, G.M. Weiss, S.A. Moore, Activity recognition using cell phone accelerometers, ACM SIGKDD Explor. Newsl. 12 (2011) 74–82, doi:10.1145/1964897.1964918.
[11] R. Chavarriaga, H. Sagha, A. Calatroni, S.T. Digumarti, G. Tröster, J.D.R. Millán, D. Roggen, The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition, Pattern Recognit. Lett. 34 (2013) 2033–2042, doi:10.1016/j.patrec.2012.12.014.
[12] M. Zhang, A.A. Sawchuk, USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors, (2012) 1036. https://doi.org/10.1145/2370216.2370438.
[13] E. Casilari, J.A. Santoyo-Ramón, J.M. Cano-García, UMAFall: A Multisensor Dataset for the Research on Automatic Fall Detection, ProcediaComput. Sci. 110 (2017) 32–39, doi:10.1016/j.procs.2017.06.110.
[14] S.S. Saha, S. Rahman, M.J. Rasna, A.K.M. Mahfuzul Islam, M.A. RahmanAhad, DU-MD: An Open-Source Human Action Dataset for Ubiquitous Wearable Sensors, in: 2018 Jt. 7th Int. Conf. Informatics, Electron. Vis. 2018 2nd Int. Conf. Imaging, Vis. Pattern Recognit., IEEE, 2018, pp. 567–572, doi:10.1109/ICIEV.2018.8641051.
[15] S.S. Saha, S. Rahman, M.J. Rasna, T. Bin Zahid, A.K.M.M. Islam, M.A.R. Ahad, Feature Extraction, Performance Analysis and System Design Using the DU Mobility Dataset, IEEE Access 6 (2018) 44776–44786, doi:10.1109/ACCESS.2018.2865093.
[16] D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, M.R. Luaces, A public domain dataset for real-life human activity recognition using smartphone sensors, Sensors (Switzerland) 20 (2020), doi:10.3390/s20082200.
[17] Laurens van der Maaten, G. Hinton, Visualizing Data using t-SNE Laurens, J. Mach. Learn. Res. 9 (2008) 2579–2605, doi:10.1007/s10479-011-0841-3.
[18] S.S. Saha, S. Rahman, M.J. Rasna, T. Bin Zahid, A.K.M.M. Islam, M.A.R. Ahad, Feature Extraction, Performance Analysis and System Design Using the DU Mobility Dataset, IEEE Access 6 (2018) 44776–44786, doi:10.1109/ACCESS.2018.2865093.
[19] WISDM Lab: Dataset, (2020). http://www.cis.fordham.edu/wisdm/dataset.php (accessed May 26, 2020).
[20] M.A.R. Ahad, Mobility Dataset: Sensor-based Activity Dataset, (2018). http://aa.binbd.com/mobility.html (accessed May 10, 2019).
[21] D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, M.R. Luaces, A Public Domain Dataset For Real-life Human Activity Recognition Using Smartphone Sensors, (2020). http://lbd.udc.es/research/real-life-HAR-dataset/ (accessed May 27, 2020).
[22] N. Sikder, M.S. Chowdhury, A.S.M. Arif, A.-A. Nahid, Human Activity Recognition Using Multichannel Convolutional Neural Network, in: 2019 5th Int. Conf. Adv. Electr. Eng., IEEE, 2019, pp. 560–565, doi:10.1109/ICAEE48663.2019.8975649.
[23] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, doi:10.1023/A:1010933404324.