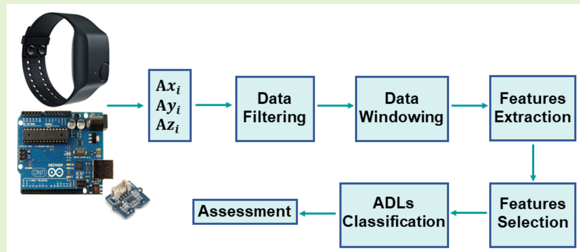
Impact of Wearable Measurement Properties and Data Quality on ADLs Classification Accuracy

(Article begins on next page)

Impact of Wearable Measurement Properties and Data Quality on ADLs Classification Accuracy

A. Poli, G. Cosoli, L. Scalise *Member, IEEE*, S. Spinsante *Senior Member, IEEE*

***Abstract*— In the field of automatic recognition and classification of Activities of Daily Living (ADLs), a paramount role to determine the classification accuracy is played by sensor technologies, as the algorithms' performance is highly affected by the nature and quality of the collected measurement data. This work aims to investigate the influence of the wearable device characteristics and measurement uncertainty on the classification accuracy. For this study, two wearables devices are considered: a top-quality smartwatch (Empatica E4) and a low-cost Arduino-based wristband prototype. These devices have been used to measure the acceleration signal at the dominant wrist of subjects performing some relevant activities in real-life conditions. The experimental evaluation of some ADLs classification algorithms shows that their accuracy fluctuates depending on the choice of the sensor, which in turn affects the amount and type of relevant features to process. As such, the combination of features' domain, i.e. time or frequency, number and type, which leads to the best classification accuracy has to be tuned on a specific sensor basis, despite the same type of signal, i.e. acceleration, is measured and processed under identical circumstances and processed. Accuracy values of 50-99% and 66-95% in the ADLs classification, are obtained for Empatica E4 and Arduino-based prototype, respectively; the best performance among classifiers is obtained with J48 and Random Forest, confirming that, with an appropriate configuration, satisfactory accuracy may be attained, even by resorting to the use of simple sensors.**



***Index Terms*— Accelerometers, Activities of Daily Living, data quality, measurement uncertainty, wearable devices.**

I. Introduction

In the last few years, research on human activity recognition (HAR), which is defined as the capability to automatically track and recognize human activity patterns in real life settings [1]), has been increasing both in terms of number of publications and application fields. Indeed, the automatic recognition and classification of the so-called Activities of Daily Living (ADLs) is gaining more and more relevance in several areas: Active and Assisted Living (AAL) [2]–[5], to support ageing population in leading an active and independent lifestyle, also offering cognitive assistance [6]; rehabilitation centers, to monitor physical activities and encourage physical exercises, providing a feedback to the caregiver [7], [8]; Smart Cities, to improve the people's quality of life [9]; human computer interaction (HCI), to recognize human body gestures in gaming and tell the machine to complete specific tasks [10]; Industry 4.0, to control and evaluate the workers' conditions, as well as to provide information about the performance [11], [12]; surveillance systems, to prevent crimes and dangerous activities in public busy environments [13], [14]. It is worthy to note how the monitoring of physical activities has gained particular relevance with the advent of tele-medicine and home care, requiring suitable sensing devices and an adequate intelligence to provide valuable parameters without the need of too many hardware components, that are often difficult to install and sometimes cumbersome and expensive.

A paramount role in HAR is obviously played by sensor technologies, which for this area can be grouped in three main categories: RGB-D camera-based, ambient-sensor based, and wearable-based [15]. The last category is gaining more and more consensus, since wearable devices have evolved and become user-friendly, relatively cheap, easy-to-use [16], naturally close to the measure and they can be applied to manifold domains, from sport medicine to AAL, with a high reliability [17]. This type of devices, typically equipped with Micro Electro-Mechanical Systems (MEMS) accelerometers and gyroscopes, can generate great amounts of data that are often difficult to manage and process to derive valuable indicators. Artificial Intelligence (AI) technologies have unquestion-

This paragraph of the first footnote will contain the date on which you submitted your paper for review. The support of the *More Years Better Lives JPI* and the Italian Ministero dell'Istruzione, Università e Ricerca (CUP: I36G17000380001), for this research activity carried out within the project PAAL - Privacy-Aware and Acceptable Lifelogging services for older and frail people, (JPI MYBL award number: PAAL_JTC2017) is gratefully acknowledged.

A. Poli and S. Spinsante are with Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Ancona, 60131 Italy, a.poli,s.spinsante@staff.univpm.it.

G. Cosoli and L. Scalise are with Dipartimento di Ingegneria Industriale e Scienze Matematiche, Università Politecnica delle Marche, Ancona, 60131 Italy, l.scalise,g.cosoli@staff.univpm.it.

ably tackled this problem and they make it possible to derive meaningful health-related figures and indicators, assisting the decision-making process and supporting the whole healthcare system [18]. The indicators provided through AI can help the patients' assistance also at home, promoting their well-being and improving their quality of life, thanks to the regular and long-term monitoring capability, and the enabled preventative interventions. Machine Learning (ML) can be considered as the most tangible manifestation of AI and it can allow to "obtain more with less", enabling the minimization of the required hardware thanks to the continuous growth in digital technology [19]. Hence, ML algorithms offer a valid solution for ADLs recognition and classification, whose performance and accuracy will inevitably depend on the metrological characteristics of the wearable device used for acquiring the data [20], which need to be considered together with the intra- and inter-subject variability impacting on the system usability [21].

Among the quantities commonly assessed by wearable devices, many studies in literature related to HAR deal with the acceleration signal [22]–[25]. In fact, accelerometers quantify motion across three dimensions [26], being at the same time small, lightweight, and inexpensive sensors, thus appearing particularly suitable for the monitoring, recognition and tracking of ADLs. It is worthy to note that the choice of device type and the sensor positioning influences the HAR performance [27]; an example is reported by Stisen *et al.* [28], which investigated 13 different devices usable as accelerometer readings. The final outcomes demonstrate that devices heterogeneities affect the performance of ML algorithms, impairing HAR tasks significantly. Regarding the device positioning, many researchers compared different devices locations to find the optimal solution, among which Cleland *et al.* [29]. By considering seven ambulatory activities, Cleland *et al.* found significant differences in classification accuracy provided by data from different positions, such as wrist and foot (even if the wrist seems to be the most preferred placement [30]). A possible solution that can mitigate these problems is to consider more than one sensor to collect a huge variety of data. Indeed, the classification accuracy is arguably improved by fusion with data generated by other sensors, such as gyroscopes [31]. Finally, the use of multiple body-worn accelerometers can improve the classification accuracy [32], but the user's comfort should always be accounted for, in particular to ensure the compliance in the prolonged use of wearable sensors for long-term monitoring [32]–[34].

Concerning ML algorithms, different approaches have been extensively applied in the literature. Among the others, Zubair *et al.* [35] performed activity classification using Random Forest (RF) and Decision Tree (DT) in connection with Adaboost ensemble method, obtaining an accuracy of 99.8% and 99.9%, respectively, in five different activities, measured by means of one body worn accelerometer. RF validity has been confirmed also by Sztyler *et al.* [36] reporting a performance of 84% in the classification of eight different activities measured with the wearable device worn in seven different body positions. Xu *et al.* [37] state that activity similarity can improve the results obtainable with RF used alone, reporting an accuracy increase from 92.31% to 95.59%. Fullerton *et al.* [32] report an accuracy of 97.60% in the recognition of six different ADLs with the k-Nearest Neighbours (kNN) classifier; moreover, they state that Support Vector Machine (SVM) approaches are accurate (96.7%, comparable to ensemble classifiers, reporting an accuracy of 96.4%) but quite slow, and therefore not recommended for free-living monitoring. The better performance of kNN with respect to SVM in physical activities classification is confirmed also in [38]. Hsu *et al.* [39] tested Least Squares SVM recognizer (LS-SVM) with the Nonparametric Weighted Feature Extraction by Principal Component Analysis (NWFE+PCA) reduction method in the classification of 10 ADLs and 11 sport activities, reporting an accuracy of 99.65%.

However, the performance of such algorithms is highly affected by the nature and quality of the collected dataset [40]. Such awareness results in challenges during the study, especially in the phase of experimental setup and subsequent data analysis. Among the challenges which highly affect the quality of the recognition, the most common ones are how to collect the data in the real-life conditions, and how to select and extract the features to be computed [41]. Following the preliminary findings presented in [42], the aim of this study is to investigate the influence of the wearable sensing device characteristics and measurement uncertainty on the accuracy of ADLs classification. In particular, two different measuring devices have been considered (Empatica E4 smartwatch [43] and an Arduino-based wristband prototype) to measure the acceleration signal at the dominant wrist of the subject performing different ADLs in real-life conditions. The Waikato Environment for Knowledge Analysis (Weka [44]) toolbox has been used to implement different learning algorithms and to evaluate their performance in HAR, in relation to the sensing device and the measurement data accuracy.

The paper is organized as follows. Section II introduces the experimental setup adopted and presents the materials and methods used to develop the proposed study. The related classification outcomes, and the main evaluation indicators to assess the performance of six different ML algorithms working on ADLs recognition, are discussed in Section III. Finally, the main findings are summarized and discussed in Section IV, followed by the conclusion of the work in Section V.

II. Materials and Methods

In this study, the measurement of the acceleration signal at the wrist of a subject performing different ADLs has been carried out by means of two different wearable sensing devices, namely the Empatica E4, and an Arduino-based wristband prototype.

The Empatica E4, shown in Fig. 1, is a wrist-worn top-quality sensor device that can be considered a Class IIa Medical Device according to CE Crt. No. 1876/MDD (93/42/EEC Directive).

The E4 is equipped with multiple sensors, allowing to continuously measure and then assess the physiological quantities in free-living conditions, namely: photoplethysmographic and EDA sensors, triaxial MEMS accelerometer, and optical infrared thermometer (whose characteristics are reported in Table I). The E4 can be used in two different modalities:

TABLE I
SENSORS INTEGRATED IN EMPATICA E4 AND ARDUINO-BASED PROTOTYPE DEVICES

Devices	Sensors	Measured parameters	Applications	Metrological characteristics
Empatica E4	Photoplethysmographic sensor (PPG)	Blood volume pulse (BVP)	Analysis of cardiac related-parameters (e.g., heart rate (HR) and heart rate variability (HRV))	Sampling frequency: 64 Hz
	Electrodermal activity sensor (EDA)	Changes in certain electrical properties of the skin (i.e., Impedance)	Investigation on the psychological conditions of the subject, such as stress or anxiety	Alternating current (8 Hz with max 100 μA, in compliance to IEC 60601-1:2005); Sampling frequency: 4 Hz; Resolution: 900 pS; Range: 0.01-100 μS
	Triaxial MEMS accelerometer	Acceleration signal	Analysis of motion-based activities	Sampling frequency: 32 Hz; Resolution: 0.015 g (8 bit); Range: ± 2 g (± 4 g and ± 8 g are available only with a custom on-demand firmware)
	Optical infrared thermometer	Skin Temperature	Detection of fever states	Sampling frequency: 4 Hz; Accuracy:± 0.2 °C; Range: 36÷39°C
Arduino-based prototype	I2C Grove 6-Axis Accelerometer Gyroscope (LSM6DS3 chip by ST^{TM} Microelectronics)	Acceleration	Analysis of motion-related activities	Sampling frequency: 32 Hz; Sensitivity: (0.061/0.122/0.244/0.488) mg/LSB; Range: $\pm 2/\pm 4/\pm 8/\pm 16g$

Fig. 1. Empatica E4 wrist-band device.

streaming and recording mode. In the former mode, measurement data are displayed in real-time through a dedicated mobile app (named E4 realtime), whereas in the latter data are stored in an internal memory. The battery duration declared by the manufacturer is up to 24 hours in the former modality, 48 hours in the latter. In the streaming mode, measurement data generated by the sensors onboard the E4 wristband is also sent to a secure cloud-based repository (called Empatica Connect), where the user shall setup an account, to which the device is associated by means of its serial number. The cloud-based repository provides to each registered user the list of recorded sessions, and a dashboard for data visualization; each session, saved in the form of a .zip archive, includes separate .csv files, each one pertaining to a specific sensor. In particular, the acceleration signals are organized in a table format, where the rows report the sample time, whereas the columns show the acceleration components (A_x, A_y, and A_z) along the three axes (X, Y, and Z). Each acceleration sample includes the three components. It is worthy to note that the conversion factor between raw samples and true acceleration values is attained by g/64 (where g = 9.81 m/s^2), given that according to E4 manual a sample component value of 64 corresponds to 1 g, in the measurement range [-2g, 2g] m/s^2.

On the other hand, the Arduino-based wristband prototype shown in Fig. 2 has been manufactured by the authors using an Arduino UNO Microcontroller, to which a Grove 3D digital accelerometer and gyroscope (whose characteristics are reported in Table I) has been connected through a four wires-connector. For consistency with the acceleration data format of the E4, the acceleration measurements acquired by Arduino are saved in a file with a 4-column format, containing the timestamp and the acceleration components' values (expressed as a multiple of g m/s^2) along the three axes (A_x, A_y, A_z).

A. Experimental campaign: ADLs and test population

This section describes the experiment design to collect the whole dataset, and the participants involved. Experimental tests were performed by making the test subjects wearing at the same time both the devices, i.e. the Empatica E4 and the Arduino-based prototype. Each subject was asked to perform a series of ADLs, which can be categorized as personal hygiene and housekeeping related activities, specifically: washing hands (WH), brushing teeth (BT), brushing

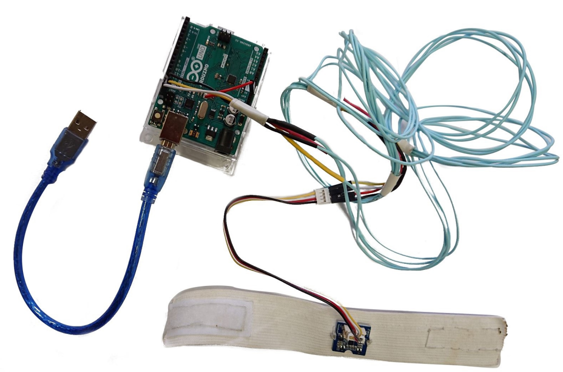


Fig. 2. Arduino-based prototype.

hair (BH), dusting (D), ironing (I), and washing dishes (WD). Each activity was recorded continuously for 5 minutes (to have a quite long acquisition and facilitating, for example, the spectral analysis of the signal thanks to a better resolution) and repeated three times by each subject (in order to have statistically significant data). The acquired signals can be considered realistic because the activities were all performed in free conditions and in a real-world scenario, namely the home environment, by means of real tools, and in standard daily hours. Only the activity duration is sometimes modified (e.g., for WH or BT), with respect to what usually happens in daily life, for the reasons mentioned above. The experiment required the participants to wear the measurement devices on the dominant wrist, executing the previously defined ADLs without any instructions or guidance. However, in order to annotate and label the data collected, the participants were supervised during the recorded sessions, to collect reference information for the activity classification task. The subject population considered in this study was composed by 36 healthy subjects (18 females, 18 males), aged 29.5 $\pm$ 3.4 years (mean $\pm$ standard deviation); 11 subjects were left-handed and 25 right-handed.

B. Data processing and feature extraction

Before moving to the classification of the ADLs performed during the tests, the acquired acceleration data were processed in MATLAB environment. Pre-processing may improve data quality and consequently increase the reliability of outcomes, since raw data are generally affected by noise, maybe linked to movement artefacts or a not adequate device-wrist contact. The main methods considered for refining noisy datasets were a 4^{th} order low-pass Butterworth filter with a cut-off frequency equal to 15 Hz, to preserve human motion while eliminating noise [40], [45], and a 3^{rd} order median filter to remove abnormal spikes. Regarding the sensor's calibration, note that E4 calibrates itself during the initial 15 s of each session, which, for this reason are removed (the subject does not move during this interval). On the other hand, the accelerometer in the Arduino-based prototype performs calibration at the start of the measurement exploiting the acceleration due to gravity; in particular, the sensor is maintained fixed in three different positions and orientations (aligning each sensor axis with the gravitational pull corresponding one) for 30 s, hence the collected data can be processed.

Before any further computation is performed, the segmentation approach is widely used to divide the signal into a series of windows segments. The segmentation depends on the type and the duration of the activities, besides the sampling rate, and it will influence the number of instances available for the following classification algorithms. Therefore, the size of windows must be both sufficient to include details about the activity and to generate stable predictions. In this study, prior to the classification, the acceleration signals are sliding windowed to equal-sized parts, namely into 3 s non-overlapping windows [46], since the performed ADLs are not rapidly time-variant.

For each window, features are extracted as peculiar of the different activities, where acceleration data significantly

TABLE II
LIST OF FEATURES EXTRACTED IN TIME AND FREQUENCY DOMAIN

Domain	Features	Computation
Time	Mean	X, Y, Z axes, SMV
	Median	X, Y, Z axes, SMV
	Standard Deviation	X, Y, Z axes, SMV
	Maximum	X, Y, Z axes, SMV
	Minimum	X, Y, Z axes, SMV
	Range	X, Y, Z axes, SMV
	Axis Correlation	XY, YZ, ZX axes
	Signal Magnitude Area	SMV
	Coefficient of Variation	X, Y, Z axes, SMV
	Median Absolute Deviation	X, Y, Z axes, SMV
	Skewness	X, Y, Z axes, SMV
	Kurtosis	X, Y, Z axes, SMV
	Zero Crossing	X, Y, Z axes, SMV
	Autocorrelation	X, Y, Z axes, SMV
	Percentiles (20^{th} - 50^{th} - 80^{th})	SMV
	No. of Peaks	SMV
	Peak - Peak Amplitude	SMV
Frequency	Spectral Energy	SMV
	Spectral Entropy	SMV
	Spectral Centroid	SMV

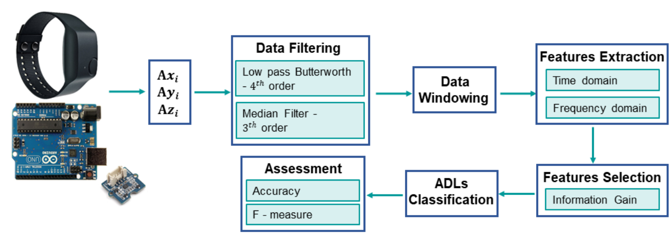


Fig. 3. Workflow of the proposed HAR approach.

vary among several subjects. In particular, ML algorithms are able to realize an indirect measurement method by the functional mapping $f(\hat{x})$ of features x [47]. As reported in the literature [45]–[48], relevant time features were derived from the acceleration signals, namely the directional components ($A_{x,i}$, $A_{y,i}$, $A_{z,i}$, where i is the index of the sample) and the Signal Magnitude Vector (SMV), which can be defined as a vector of N elements a_i, with $i = 1 \ldots N$, being N the total number of samples in a given acquisition (i.e., the product between the acquisition duration and the sampling frequency): $SMV_i = \sqrt{A_{x,i}^2 + A_{y,i}^2 + A_{z,i}^2}$. The SMV was used to reduce the impact of sensor orientation on the activity discrimination. For what regards spectral features, they were extracted from the magnitude of the discrete Fast Fourier Transform (FFT) of each signal. Table II summarizes the whole set of time and frequency features considered.

As shown in Fig. 3, the classification algorithms extract the information related to the collected activities from the computed features.

Therefore, a dataset including only time features (D1), one composed only by frequency features (D2) and a third one containing both time and frequency features (D3), extracted in each case from the SMV, were created. In the same manner, two additional datasets were established considering time only (D4), and then both time and frequency domain (D5), features extracted from both directional acceleration components and

SMV. At this point, after having organized the data from E4 and Arduino-based prototype in the described datasets, ADLs classification was performed. The performance of different ML classifiers was estimated by means of the Weka tool, which allowed to compute different evaluation metrics and also to identify the features most suitable for the activities classification.

C. Feature selection and automatic classification of ADLs

According to the literature, the selected features have a large impact on the classifiers performance. For this reason, in order to select the most discriminant features and exclude the redundant ones, a filter feature selection method was used. Among the feature selection methods supported in Weka, the Information Gain (IG) approach was selected, as validated in a previous work [40]. This approach evaluates the amount of information for each feature by measuring the information gain with respect to the class of activity [44], namely:

$$IG = H(Class) - H(Class|Feature) \tag{1}$$

where *H(Class)* is the entropy of the class of activity and *H(Class|Feature)* is the conditional entropy, which represents how the considered feature is consistent to identify a particular class of activity. More specifically, IG assesses a specific rank for each feature; low ranks are associated to redundant or non essential features to identify the class, whereas the higher ranks are related to features carrying significant information for the classification. Regarding the approaches for automatic classification, supervised learning approaches are generally preferred in HAR systems, since they have proved to show good performances in learning relationships among input attributes, features, target attributes, and labeled classes [49]. In this experimentation, the authors have tested six different ML algorithms and evaluated their performance in the ADLs classification. These approaches can be grouped in three categories, based on the estimation criterion for adjusting the parameters of the classification method [50], [51]:

- Generative approaches (e.g., Naïve-Bayes, NB), which are statistical approaches to the pattern recognition problem. Although they need a huge amount of training data, these techniques are flexible and able to deal effectively with the data uncertainty. Moreover, the generative techniques learn from both labeled and unlabeled data;
- Discriminative approaches (e.g., DT, kNN, RF, SVM, and Artificial Neural Networks - ANNs), which create the decision boundaries in the feature space, thus learning the feature mapping to activity labels. Discriminative techniques face the problem of over-fitting and the large amount of labeled data required for training;
- Heuristic approaches, which is a hybrid modality using a combination of generative and discriminative approaches. Generally, it was found to reach the best predictive performance with respect to its single use.

The authors tested all the above-mentioned approaches in classifying the considered ADLs; among the DT ones, the J48 algorithm was used. Since it is well-known that cross-validation results improve by repeating the operation multiple times, the 10-fold cross-validation method was adopted for the evaluation of the classifiers [52]. Each fold consists of a training and a testing set; in this study, 70% of all the labeled data is used for the training, whereas the remaining 30% for the testing. The performance was measured by two different validation metrics, namely the accuracy and f-measure.

D. Evaluation Metrics

According to the literature, there are a wide range of performance metrics commonly used to evaluate the classification algorithms. Here, for each tested approach we focused on two evaluation measures that are the attained accuracy and the f-measure. The former is defined as the number of correctly classified ADLs instances over the total number of instances considered [48], or:

$$accuracy = (1 - ErrorRate) \times 100 \tag{2}$$

where:

$$ErrorRate = \frac{|N_{cci} - N_{ti}|}{N_{ti}} \tag{3}$$

being N_{cci} the number of correctly classified instances, and N_{ti} the number of total instances considered by each classification algorithm. Contrarily, f-measure refers to a weighted harmonic mean of the recall and precision, as an indicator of the overall effectiveness of the activity classification. According to the definitions [48], the f-measure can be calculated as per (4):

$$f - measure = \frac{(1+\beta)^2 \times Recall \times Precision}{\beta^2 \times Recall + Precision}. \tag{4}$$

Assuming that β is the weight coefficient, *Recall* (or *Sensitivity*) is the ratio of correctly classified instances to the total number of identified instances that are true, calculated as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

whereas *Precision* is the ratio of correctly classified instances to relevant instances, defined as:

$$Precision = \frac{TP}{TP + FP}. \tag{6}$$

Hence, f-measure values depends on the abovementioned performance indicators: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). An f-measure closer to 1 means high values of accuracy.

III. Results

A. Evaluation of classifiers performance

In this section, the authors evaluate the performance of the different classifiers by considering as inputs the identified datasets. The results from the 10-fold cross-validation are reported as average values of the above-mentioned evaluation metrics. In terms of accuracy, the preliminary performances of the six different algorithms tested with the same set of features for both the Empatica E4 and the Arduino-based prototype signals are summarized in Table III. Regarding the

TABLE III
AVERAGE PERCENT ACCURACY OF THE TESTED CLASSIFIERS ON ACCELERATION MEASUREMENT DATA FROM EMPATICA E4 AND ARDUINO-PROTOTYPE

Device	Classifier	D1	D2	D3	D4	D5
Empatica E4	J48	99.55	31.36	99.51	99.39	99.23
	RF	98.55	30.77	95.53	99.12	97.67
	NB	56.87	20.80	48.65	50.66	54.20
	SVM	70.65	17.84	71.11	83.39	86.12
	ANNs	93.28	24.24	92.67	96.25	93.52
	kNN	80.65	27.78	81.41	89.82	89.75
Arduino-prototype	J48	83.49	31.63	80.37	90.04	87.05
	RF	89.86	39.08	88.70	95.06	93.81
	NB	54.37	16.73	55.93	66.54	67.23
	SVM	65.17	18.81	65.97	81.24	81.57
	ANNs	75.31	23.80	72.79	91.70	91.28
	kNN	79.03	40.03	78.28	92.75	92.23

TABLE IV
SPECTRAL ENERGY AND ENTROPY VALUES (MEAN ± STANDARD DEVIATION) COMPUTED FOR EACH DEVICE AND ADL

ADL	Spectral Entropy		Spectral Energy	
	E4	Arduino	E4	Arduino
WH	4.28±3.79	2.78±1.80	0.51±0.22	0.61±0.19
BT	3.91±2.84	4.74±3.05	0.44±0.15	0.55±0.16
BH	2.31±1.61	4.44±3.81	0.58±0.18	0.53±0.23
D	5.42±5.44	4.41±3.61	0.46±0.23	0.49±0.19
I	5.75±5.09	7.33±9.21	0.44±0.19	0.31±0.19
WD	2.44±1.27	3.12±2.01	0.55±0.14	0.54±0.17

TABLE V
PARTIAL PERCENT ACCURACY AND AVERAGE PERCENT ACCURACY (AVG.) OF THE TESTED CLASSIFIERS (CL.) – EMPATICA E4

CL.	ADLs						Avg.
	WH	BT	BH	D	I	WD	
J48	99.09	98.55	100.00	99.81	99.45	99.40	99.39
RF	98.55	98.91	100.00	98.73	99.46	99.27	99.12
ANNs	96.19	97.46	99.45	96.90	96.24	91.84	96.25
kNN	90.39	90.57	99.45	92.39	88.94	77.17	89.82
NB	90.94	85.68	17.93	34.42	44.21	29.34	50.66
SVM	86.59	90.99	99.45	58.51	89.31	76.81	83.39

TABLE VI
PARTIAL PERCENT ACCURACY AND AVERAGE PERCENT ACCURACY (AVG.) OF THE TESTED CLASSIFIERS (CL.) – ARDUINO PROTOTYPE

CL.	ADLs						Avg.
	WH	BT	BH	D	I	WD	
J48	92.32	91.38	89.51	88.57	87.07	91.38	90.04
RF	94.38	96.82	96.63	94.38	93.82	94.38	95.06
ANNs	90.45	95.88	94.57	90.07	89.88	89.32	91.70
kNN	93.44	95.50	97.00	92.13	88.01	90.45	92.75
NB	71.72	82.21	82.78	52.06	64.80	45.70	66.54
SVM	83.52	91.76	85.39	76.03	84.64	66.10	81.24

Empatica E4 data, it is possible to observe that accuracy is >90% for D1 and D3 datasets analyzed with J48, RF, and ANNs classifiers, whereas the value decreases to the range 70-80% when SVM and kNN are used; the worst performance is reported for NB classifier (i.e., 56.78% and 48.65% for D1 and D3 respectively). Similar results are obtained with D4 and D5 datasets, whereas the accuracy of all the classifiers is <40% if D2 is considered. Such results confirm that the frequency domain features used alone are not suitable for the automatic classification of the considered ADLs, based on acceleration signals collected on the wrist. By the computation of the mean ± standard deviation values of the Spectral Energy and Spectral Entropy features, reported in Table IV, it is possible to see that Spectral Energy values are very similar between the two measurement devices, whereas the Spectral Entropy is typically higher and much more dispersed for the Arduino-based device than the E4. Similar values of the Spectral Energy, either for different devices and activities, explain why the use of such feature would not improve the accuracy of the classification algorithm in a significant fashion. On the other hand, Spectral Entropy is a measure of uncertainty, and we can explain the differences among the devices based on the fact that E4 runs a proprietary firmware that is able to identify unreliable or noisy samples and automatically remove them, while this capability is not available from the Arduino-base device. So, we can assume that Spectral Entropy may be a relevant feature when the target of the classification is to identify the type of device used (E4 or Arduino), and not the activity, as its value is similar among the six ADLs considered.

On the other hand, if we consider the results attained from the Arduino-based prototype, the highest accuracy values ($> 80\%$) are obtained for D1 and D3 datasets processed by J48 and RF classifiers. D4 and D5 datasets (including also the features computed over the three directions of the acceleration signals, namely A_x, A_y and A_z) allow to obtain better performance for all the tested classifiers: accuracy values $> 90\%$ are obtained for RF, ANNs, and kNN. Again, the overall worst performance is reported for NB classifier (i.e., 16.73%) and poor results are obtained when D2 dataset is chosen, whatever classifier is used.

Being D4 the dataset allowing very good performance, the accuracy values provided on it by the different classifiers are detailed and summarized in Table V and Table VI, for Empatica E4 and Arduino-based prototype, respectively. In particular, the Tables show how accurately the six activities performed by subjects (i.e., Washing Hands, Brushing Teeth, Brushing Hair, Dusting, Ironing, Washing Dishes) are classified, depending on the different ML classifiers selected.

As stated above, for measurement data collected by wearing the Empatica device, J48, RF and ANNs outperformed the other classifiers, whereas NB was able to correctly recognize 50% of the activities. The same activities registered by wearing Arduino-based prototype were well-recognized by RF, ANNs and kNN algorithms, reaching a total accuracy greater than 91%. It is worthy to note that J48 and RF achieved 100% accuracy for BH activity recorded by using the Empatica E4 device. However, their performance decreases when processing data collected with the Arduino-based device. Similarly, ANNs and SVM were efficient in recognizing all the activities from the acceleration signals collected using Empatica E4, but they were less accurate in classifying data from the Arduino-based tested device. On the contrary, kNN classifier was more suitable in analyzing data from Arduino-based tested device,

TABLE VII
F-MEASURE PERFORMANCE OF THE TESTED CLASSIFIERS ON ACCELERATION MEASUREMENT DATA

	Arduino - based device			Empatica E4		
Dataset	J48	RF	ANNs	J48	RF	ANNs
D1	0.83	0.90	0.75	0.99	0.98	0.93
D2	0.32	0.38	0.20	0.31	0.31	0.23
D3	0.99	0.98	0.89	0.99	0.96	0.92
D4	0.90	0.89	0.92	0.99	0.99	0.96
D5	0.87	0.94	0.91	0.99	0.98	0.93

TABLE VIII
SENSITIVITY OF THE TESTED CLASSIFIERS ON ACCELERATION MEASUREMENT DATA

	Arduino - based device			Empatica E4		
Dataset	J48	RF	ANNs	J48	RF	ANNs
D1	0.83	0.89	0.75	0.99	0.98	0.93
D2	0.31	0.39	0.24	0.31	0.31	0.24
D3	0.98	0.97	0.89	0.99	0.96	0.93
D4	0.90	0.89	0.92	0.99	0.98	0.96
D5	0.87	0.94	0.91	0.99	0.98	0.93

reaching an accuracy greater than J48 one, for five activities, excepting the WD activity. NB classifier seems to provide a relatively low accuracy, since apparently the highest number of mis-classifications happened when using it. Specifically, the lowest accuracy value (i.e., 17.93%) was achieved working on data pertaining to the BH activity and collected using the Empatica E4.

In order to validate the findings presented above and compare the performance of different learning algorithms, further evaluations were carried out. Besides the accuracy, the f-measure was computed for the tested classifiers with the highest accuracy of the model for both Empatica E4 and Arduino-based devices (i.e., J48, RF and ANNs), as reported in Table VII. Among the classifiers that achieved the highest accuracy, three ML algorithms were selected (i.e., J48, RF and ANNs) and the f-measure was calculated for each single dataset, as shown in Table VII. It is interesting to note how very high values were achieved on D1, D3, D4 and D5 by using Arduino-based device and Empatica E4. Firstly, the high values of f-measure confirm the accuracy trend, achieving simultaneously the highest accuracy and f-measure with the same ML approaches. Secondly, the low f-measure values in D2, which is the dataset including frequency-domain features only, demonstrate that the features computed in the time domain are essential to reach a good classification performance.

Further analyses addressed the sensitivity and the precision of each tested approach, as shown in Tables VIII and IX. According to the definitions (5) and (6), the results in Tables VIII and IX show that J48 and RF achieved the highest true positive rates in almost all the datasets. By comparing the five different datasets, the three ML approaches had unstable accuracy for any dataset configuration selected. Anyway, the classifier that reached the highest values was J48, especially when fed on measurement data acquired by means of Empatica. Such consideration is confirmed in terms of sensitivity and

TABLE IX
PRECISION OF THE TESTED CLASSIFIERS ON ACCELERATION MEASUREMENT DATA

	Arduino - based device			Empatica E4		
Dataset	J48	RF	ANNs	J48	RF	ANNs
D1	0.84	0.89	0.75	0.99	0.98	0.93
D2	0.32	0.38	0.22	0.31	0.31	0.23
D3	0.99	0.97	0.89	0.99	0.96	0.93
D4	0.90	0.89	0.92	0.98	0.98	0.96
D5	0.87	0.93	0.91	0.99	0.98	0.93

TABLE X
SUBSET OF FEATURES SELECTED BASED ON THE INFORMATION GAIN EVALUATION

Arduino - based device	Empatica E4
Autocorrelation	Signal Magnitude Area
Maximum	Mean
Mean	Percentiles 80^{th}
Median	Median
Percentiles 80^{th}	Autocorrelation
Percentiles 20^{th}	Percentile 20^{th}
Range	Maximum
Standard Deviation	Variance
Median Absolute Deviation	Standard Deviation
Interquartile	Range

precision. More in general, both metrics show high rates for all classifiers, with the exception of D2. The reason is that the ML algorithms require a larger amount of data for training, which was not possible with D2, being composed exclusively by features in the frequency domain.
From all these findings, RF clearly showed an accuracy better than the other classifiers in terms of good performance measures, in classifying different activities, based on data collected from different wearable devices.

B. Evaluation of filter-selected features

Identifying the appropriate features is a crucial phase to develop a successful automatic classification system based on machine learning. Indeed, a high correlation among the features may affect the classification performance in classifying the activities [53]. As such, the IG filter was applied on D3 dataset, since it includes the whole set of features for SMV signals, from both Arduino and E4. In order to have the same number of features in both datasets, and according to the number of features selected in previous works [54], [55], the 10 features with the highest IG ranking score were selected to compose a new optimal subset, and test again the classifiers performances. For both the devices used, the selected 10 features are listed in Table X, arranged in descending order of their IG value.

It is interesting to notice how the features selected by the IG filter are almost the same but with different ranking, for the two sensing devices. Median absolute deviation with interquartile, and signal magnitude area with variance are the two couples of features that uniquely characterize the subset of selected features for the specific Arduino-based prototype and the E4 device, respectively. As a result, the selected

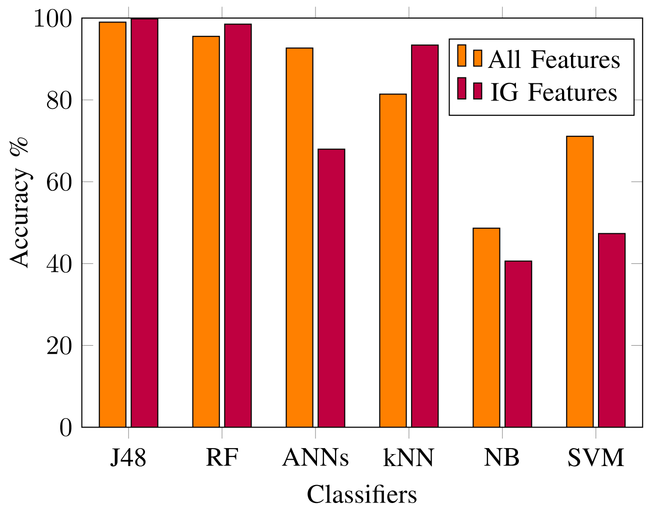


Fig. 4. Comparison of accuracy before and after applying the Information Gain - Empatica E4 acceleration measurement data.

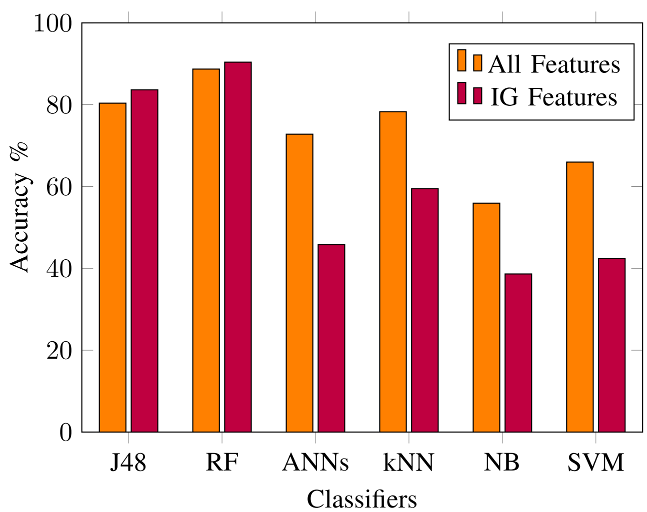


Fig. 5. Comparison of accuracy before and after applying the Information Gain - Arduino-based device acceleration measurement data.

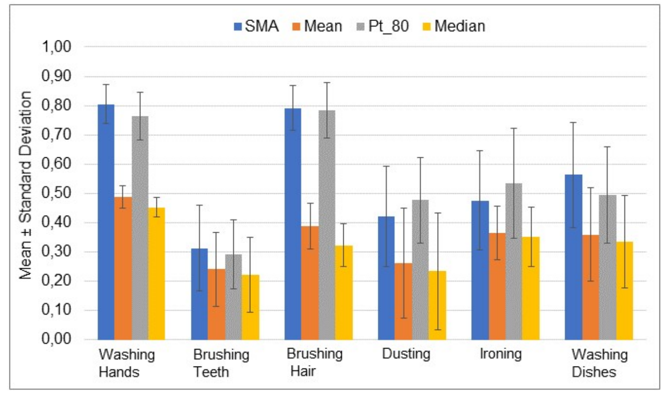


Fig. 6. Mean and standard deviation of the four most relevant features normalized and averaged over the population performing the six ADLs - Empatica E4.

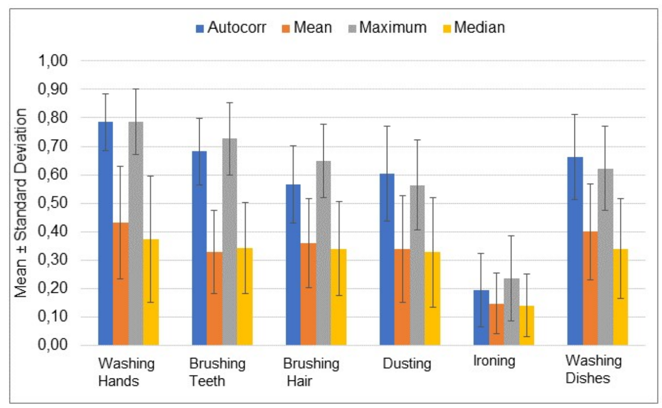


Fig. 7. Mean and standard deviation of the four most relevant features normalized and averaged over the population performing the six ADLs - Arduino based prototype.

features were used as input to the six classifiers to compare the performance obtained when using all features versus a smaller set, as shown in Fig.4 and Fig. 5.

By examining the results obtained from E4 device, among the learning algorithms, the RF, kNN, and J48 improved their performance, reaching the highest values, as shown in Fig.4. In details, following the features selection, J48 provided the highest accuracy of 99.81%. Compared to the performance of RF, kNN, and J48, the classification results from NB, SVM, and ANNs cannot be considered satisfactory. Looking at the results obtained by analyzing data collected with Arduino-base prototype (Fig. 5), the highest accuracy of 89,39% is achieved by RF. On the other hand, the accuracy obtained by NB and SVM decreased from 56% to 40% and from 66% to 42%, respectively. ANNs contributed with a wider range of accuracy values, namely from 72% to 45%.

By comparing the general performance of using all the features versus using less features, it is possible to notice that some algorithms slightly increased their performance when processing a smaller set of features, namely J48 and RF. In contrast, other classifiers drastically decreased their performance when using the reduced subset of features, especially to classify data from Arduino-based device; indeed, this is the case of ANNs, kNN, NB, and SVM, reporting a substantial decrease in terms of accuracy (the maximum decrease, i.e., 27.02% obtained for ANNs). Regarding the classification of activities performed by wearing E4, the results indicate that three out of six classifiers (i.e., ANNs, NB, and SVM) obtained lower accuracy percentages (i.e., 67.96%, 40.61%, and 47.34% for ANNs, NB, and SVM, respectively). This is reasonably due to the nature of each classifier, which can be differently affected by irrelevant and redundant features.

Therefore, the variability among the activities was assessed resorting to the four best features selected by IG (Table X). Fig. 6 and Fig. 7 show the mean values and the related standard deviation for each selected feature, normalized and then averaged over all the test users. Regarding E4, signal magnitude area, mean, percentile 80^{th}, and median were compared among the six activities, as shown in Fig. 6. Instead, in Fig. 7 the autocorrelation, mean, maximum and median values were selected and compared, to estimate the variability among the activities recorded by wearing the Arduino prototype. Looking at the two Figures we obtain information about the characteristics of the collected signal, and consequently about the activities performed by the users.

Although almost all the error bars are wider for the Arduino-based prototype measurement data, there are some exceptions in which the standard deviation values are quite similar to those obtained from E4. A remarkable example is given by the features computed for the ”Dusting” and ”Washing Dishes” activities. An appropriate selection of the features can reduce the high correlation between them and the dimension of the feature space as well. This would improve the algorithm performance, also in terms of latency, which is important to consider in (almost) real-time applications, such as AAL or healthcare-related ones, where a trade-off between performance and efficiency is mandatory.

IV. Discussion

In this work, the authors evaluated automatic HAR by means of six different classifiers (i.e., J48, RF, ANNs, kNN, NB, and SVM) on six ADLs (i.e., Washing Hands, Brushing Teeth, Brushing Hair, Dusting, Ironing, and Washing Dishes) performed in realistic daily life conditions by 36 subjects, and for which acceleration signals were collected from the wrist. Two wearable devices were worn at the same time by the experiment participants, in order to evaluate the effect of the measurement device choice: Empatica E4, an expensive research device with certified features, and an Arduino-based prototype, developed on purpose in the lab, from relatively cheap commercial components. In order to compare both the devices, data quality and classification accuracy were investigated. In particular, the acceleration signals only were considered for the ADLs classification.

The classification performance can fluctuate, depending on the type of both activity and classifier (Table V and Table VI). Identifying the appropriate features can be a chance to improve the recognition performance. Therefore, five different datasets combining features were arranged and tested; the best classification performance in terms of accuracy is obtained when time and frequency domain features are considered from the acceleration SMV together with the signal components along the three directions (i.e., A_x, A_y, and A_z), that is D4 and D5 datasets.

The performance of classification using Empatica E4 was higher than that obtained using the Arduino-based prototype. Indeed, it was observed that the ML classifiers obtained higher values of accuracy, f-measure, specificity, and precision, when processing data collected from Empatica E4 than from Arduino, over almost all classifiers. As a result, we can infer that the measurement device characteristics affect the data quality and, as a consequence, the activities classification accuracy. Such influence is strictly connected to the reliability of activity classification in critical application fields, such as assistive technologies for ageing people. Indeed, accuracy values are in the range of 50%-99% for Empatica E4, instead values of 66%-95% are obtained for the Arduino-based prototype. The best performance is reached by using J48 and RF classifiers for both the wearable devices used in this study.

Another point of interest is the type of activity performed, and consequently the signal quality. Acceleration signals are sensitive to the movements of the subjects. Accordingly, in this work some activities are classified with a better performance, whereas there are activities more difficult to discriminate. In particular, the quality of data declines very drastically in the case of non predictable physical activities. For example, participants differently washed dishes and washed hands, since these activities are characterized by personal and casual movements, which are extremely subjective. In fact, from Table V, it is observed that activities such as ”Washing Hands” and ”Washing Dishes” featured low classification rates, with high misclassification, mostly due to the difficulty of distinguishing the two activities characterized by highly random gestures. Similarly, ”Dusting activity” can be considered a quite random activity, depending on the cleaning surface, the tool used to dust and the performed hand gestures. On the other hand, activities such as ”Brushing Teeth”, ”Brushing Hair” and ”Ironing” are performed quite similarly among the users. In this case, the movements are repetitive, linear and performed mainly along one spatial dimension. Nevertheless, the best performing classifiers were the RF and the DT algorithms, that outperformed other classifiers in most cases.

The performance of classifiers was evaluated also considering the most meaningful features. The number of relevant features for a good classification of ADLs varies both with the sensing device and the classifier, denoting the importance of an appropriate choice of both hardware and software components of a monitoring system, especially when high accuracy and reliability are required, such as in AAL and health-related applications, sometimes also supporting decision-making processes in the definition of therapeutic strategies. This means that a HAR system should consider the diversity among the activities repeated by the same person; in particular, in order to have a good performance, a classifier should be able to effectively deal with uncertainty and variability of the input data. An accurate HAR system, reaching a compromise between performance and computation load, could be a valid tool in many applications. Especially for ageing people, this tool can support as long as possible independent and safe living, thus aiding the healthcare system in remotely managing decentralized care, in home environments.

V. Conclusion and Future works

The results of this study show the potential implications that measurement device and data quality may have on HAR classification accuracy. This study would analyze the consequences of making decisions in HAR system in real-life, investigating some possible solutions to improve the results. As a future work, the classification approaches herein considered can be trained and tested on data gathered from older adults, verifying the classification performance not only with respect to the choice of device and the selection of features, but also in terms of the influence of age-related patterns in the acceleration signals.

References

[1] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE Pervasive Comput.*, vol. 9, no. 1, pp. 48–53, Jan.-Mar. 2010, 10.1109/MPRV.2010.7.

[2] P. Climent-Pérez, S. Spinsante, A. Mihailidis, and F. Florez-Revuelta, "A review on video-based active and assisted living technologies for automated lifelogging," *Expert Syst. Appl.*, vol. 139, p. 112847, Jan. 2020, Accessed on: May 20, 2020, DOI: 10.1016/j.eswa.2019.112847, [Online].

[3] I. Pires, N. Garcia, N. Pombo, F. Flórez-Revuelta, and S. Spinsante, "Approach for the development of a framework for the identification of activities of daily living using sensors in mobile devices," *Sensors*, vol. 18, no. 2, p. 640, Feb. 2018, Accessed on: May 20, 2020, DOI: 10.3390/s18020640, [Online].

[4] S. Spinsante, S. Porfiri, and L. Scalise, "Accuracy of heart rate measurements by a smartwatch in low intensity activities," in *Proc. MeMeA*, Istanbul, Turkey, Jun. 2019, pp. 1–6.

[5] M. Rawashdeh, M. G. A. Zamil], S. Samarah, M. S. Hossain, and G. Muhammad, "A knowledge-driven approach for activity recognition in smart homes based on activity profiling," *Future Gener. Comput. Syst.*, vol. 107, pp. 924 – 941, 2020, Accessed on: May 20, 2020, DOI: 10.1016/j.future.2017.10.031, [Online].

[6] M. Z. Uddin and M. M. Hassan, "Activity recognition for cognitive assistance using body sensors data and deep convolutional neural network," *IEEE Sens. J.*, vol. 19, no. 19, pp. 8413–8419, Oct. 2019, Accessed on: May 20, 2020, DOI:10.1109/JSEN.2018.2871203, [Online].

[7] Y.-J. Chang, S.-F. Chen, and J.-D. Huang, "A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," *Res. Dev. Disabil.*, vol. 32, no. 6, pp. 2566–2570, 2011, Accessed on: May 20, 2020, 10.1016/j.ridd.2011.07.002, [Online].

[8] D. González-Ortega, F. Díaz-Pernas, M. Martínez-Zarzuela, and M. Antón-Rodríguez, "A kinect-based system for cognitive rehabilitation exercises monitoring," *Comput. Methods Prog. Biomed.*, vol. 113, no. 2, p. 620–631, Feb 2014, Accessed on: May 20, 2020, DOI: 10.1016/j.cmpb.2013.10.014, [Online].

[9] M. N. Malik, M. A. Azam, M. Ehatisham-ul Haq, W. Ejaz, and A. Khalid, "ADLauth: Passive authentication based on activity of daily living using heterogeneous sensing in smart cities," *Sensors*, vol. 19, p. 2466, May 2019, Accessed on: May 20, 2020, DOI: 10.3390/s19112466, [Online].

[10] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Trans Cybern*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013, Accessed on: May 20, 2020, DOI: 10.1109/TCYB.2013.2265378, [Online].

[11] L. C. Günther, S. Kärcher, and T. Bauernhansl, "Activity recognition in manual manufacturing: Detecting screwing processes from sensor data," in *Proc. CIRP CMS*, Ljubljana, Slovenia, 2019, pp. 1177–1182.

[12] W. Tao, Z.-H. Lai, M. C. Leu, and Z. Yin, "Worker activity recognition in smart manufacturing using imu and semg signals with convolutional neural networks," in *Proc. SME's NAMRC*, vol. 26, Texas, USA, 2018, pp. 1159–1166.

[13] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *Vis. Comput.*, vol. 29, p. 983–1009, Sep. 2012, Accessed on: May 20, 2020, DOI: 10.1007/s00371-012-0752-6, [Online].

[14] S.-R. Ke, H. Thuc, Y. Lee, J. Hwang, J.-H. Yoo, and K. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, Jun. 2013, Accessed on: May 20, 2020, DOI: 10.3390/computers2020088, [Online].

[15] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey," https://arxiv.org/pdf/2004.08821.pdf, Accessed on: May 29, 2020.

[16] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 2013, Accessed on: May 20, 2020, DOI: 10.1109/SURV.2012.110112.00192, [Online].

[17] G. Cosoli, S. Spinsante, and L. Scalise, "Wrist-worn and chest-strap wearable devices: Systematic review on accuracy and metrological characteristics," *Measurement*, vol. 159, p. 107789, Jul. 2020, Accessed on: May 20, 2020, DOI: doi.org/10.1016/j.measurement.2020.107789, [Online].

[18] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Br. Med. J.*, vol. 2, pp. svn–2017, Jun. 2017, Accessed on: May 20, 2020, DOI: 10.1136/svn-2017-000101, [Online].

[19] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," *J. Glob. Health*, vol. 8, no. 2, Oct. 2018, Accessed on: May 20, 2020, DOI: 10.7189/jogh.08.020303, [Online].

[20] L. Scalise and G. Cosoli, "Wearables for health and fitness: Measurement characteristics and accuracy," in *Proc. I2MTC*, Houston, TX, USA, 2018, pp. 1–6.

[21] A. Akbari and R. Jafari, "Personalizing activity recognition models with quantifying different types of uncertainty using wearable sensors," *IEEE. Trans. Biomed. Eng.*, pp. 1–1, Jan. 2020, Accessed on: May 20, 2020, DOI: 10.1109/TBME.2019.2963816, [Online].

[22] C. V. C. Bouten, K. T. M. Koekkoek, M. Verduin, R. Kodde, and J. D. Janssen, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," *IEEE. Trans. Biomed. Eng.*, vol. 44, no. 3, pp. 136–147, Mar. 1997, Accessed on: May 20, 2020, DOI: 10.1109/10.554760, [Online].

[23] D. Karantonis, M. Narayanan, M. Mathie, N. Lovell, and B. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 156–167, Jan. 2006, Accessed on: May 20, 2020, DOI: 10.1109/TITB.2005.856864, [Online].

[24] J. Lu and K. Tong, "Robust single accelerometer-based activity recognition using modified recurrence plot," *IEEE Sens. J.*, vol. 19, no. 15, pp. 6317–6324, Aug. 2019, Accessed on: May 20, 2020, DOI: 10.1109/JSEN.2019.2911204, [Online].

[25] M. Mathie, A. Coster, N. Lovell, and B. Celler, "Accelerometry: Providing an integrated, practical method for long-term, ambulatory monitoring of human movement," *Physiol. Meas.*, vol. 25, pp. R1–20, May 2004, Accessed on: May 20, 2020, DOI: 10.1088/0967-3334/25/2/R01, [Online].

[26] A. R. Javed, U. Sarwar, S. Khan, C. Iwendi, M. Mittal, and N. Kumar, "Analyzing the effectiveness and contribution of each axis of tri-axial accelerometer sensor for accurate activity recognition," *Sensors*, vol. 20, p. 2216, Apr. 2020, Accessed on: May 20, 2020, DOI: 10.3390/s20082216, [Online].

[27] L. Atallah, B. Lo, R. King, and G. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 4, pp. 320–329, Aug. 2011, Accessed on: May 20, 2020, DOI: 10.1109/TBCAS.2011.2160540, [Online].

[28] A. Stisen, H. Blunck, S. Bhattacharya, T. Prentow, M. Kjærgaard, A. Dey, T. Sonne, and M. Jensen, "Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition," in *Proc. SenSys-ACM*, Seoul, South Korea, 2015, pp. 127–140.

[29] I. Cleland, B. Kikhia, C. Nugent, A. Boytsov, J. Hallberg, K. Synnes, S. McClean, and D. Finlay, "Optimal placement of accelerometers for the detection of everyday activities," *Sensors (Basel)*, vol. 13, no. 7, pp. 9183–200, Jul. 2013, Accessed on: May 20, 2020, DOI: 10.3390/s130709183, [Online].

[30] J. Bergmann, V. Chandaria, and A. Mcgregor, "Wearable and implantable sensors: The patient's perspective," *Sensors (Basel)*, vol. 12, no. 12, pp. 16 695–709, Dec. 2012, Accessed on: May 20, 2020, DOI: 10.3390/s121216695, [Online].

[31] A. Wang, G. Chen, J. Yang, S. Zhao, and C. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sens. J.*, vol. 16, no. 11, pp. 4566–4578, Jun. 2016, Accessed on: May 20, 2020, DOI: 10.1109/JSEN.2016.2545708, [Online].

[32] E. Fullerton, B. Heller, and M. Munoz-Organero, "Recognizing human activity in free-living using multiple body-worn accelerometers," *IEEE Sens. J.*, vol. 17, no. 16, pp. 5290–5297, Aug. 2017, Accessed on: May 21, 2020, DOI: 10.1109/JSEN.2017.2722105, [Online].

[33] Z. Wang, D. Wu, J. Chen, A. Ghoneim, and M. A. Hossain, "A triaxial accelerometer-based human activity recognition via eemd-based features and game-theory-based feature selection," *IEEE Sens. J.*, vol. 16, no. 9, pp. 3198–3207, May 2016, Accessed on: May 21, 2020, DOI: 10.1109/JSEN.2016.2519679, [Online].

[34] O. C. Kurban and T. Yıldırım, "Daily motion recognition system by a triaxial accelerometer usable in different positions," *IEEE Sens. J.*, vol. 19, no. 17, pp. 7543–7552, Sep. 2019, Accessed on: May 21, 2020, DOI: 10.1109/JSEN.2019.2915524, [Online].

[35] M. Zubair, K. Song, and C. Yoon, "Human activity recognition using wearable accelerometer sensors," in *Proc. ICCE*, Seoul, South Korea, 2016, pp. 1–5.

[36] T. Sztyler, H. Stuckenschmidt, and W. Petrich, "Position-aware activity recognition with wearable devices," *Pervasive Mob. Comput.*, vol. 38, Jan. 2017, Accessed on: May 21, 2020, DOI: 10.1016/j.pmcj.2017.01.008, [Online].

[37] H. Xu, Y. Pan, J. Li, L. Nie, and X. Xu, "Activity recognition method for home-based elderly care service based on random forest and activity similarity," *IEEE Access*, vol. 7, pp. 16 217–16 225, 2019, Accessed on: May 21, 2020, DOI: 10.1109/ACCESS.2019.2894184, [Online].

[38] A. Jain and V. Kanhangad, "Human activity classification in smartphones using accelerometer and gyroscope sensors," *IEEE Sens. J.*, vol. 18, no. 3, pp. 1169–1177, Feb. 2018, Accessed on: May 21, 2020, DOI: 10.1109/JSEN.2017.2782492, [Online].

[39] Y. Hsu, S. Yang, H. Chang, and H. Lai, "Human daily and sport activity recognition using a wearable inertial sensor network," *IEEE Access*, vol. 6, pp. 31 715–31 728, 2018, Accessed on: May 21, 2020, DOI: 10.1109/ACCESS.2018.2839766, [Online].

[40] A. Poli, S. Spinsante, C. Nugent, and I. Cleland, *Improving the Collection and Understanding the Quality of Datasets for the Aim of Human Activity Recognition*. 1st ed., Cham, Switzerland: Springer, 2020, ch. 7, pp. 147–165. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-030-25590-9_7

[41] S. Slim, A. Atia, M. M.A., and M.-S. Mostafa, "Survey on human activity recognition based on acceleration data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 3, Jan. 2019, Accessed on: May 21, 2020, DOI: 10.14569/IJACSA.2019.0100311, [Online].

[42] A. Poli, L. Scalise, S. Spinsante, and A. Strazza, "ADLs monitoring by accelerometer-based wearable sensors: Effect of measurement device and data uncertainty on classification accuracy," in *15th Int. Symp. on Medical Measurements and Applications*, vol. To appear, Jun. 2020.

[43] *E4 WristBand from Empatica User's Manual*, Empatica Inc., MI, IT, 2018.

[44] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10–18, Nov. 2009, Accessed on: May 21, 2020, DOI: 10.1145/1656274.1656278, [Online].

[45] I. Cleland, M. P. Donnelly, C. D. Nugent, J. Hallberg, M. Espinilla, and M. Garcia-Constantino, "Collection of a diverse, realistic and annotated dataset for wearable activity recognition," in *Proc. PerCom Workshops*, Athens, Greece, 2018, pp. 555–560.

[46] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, Dec 2015, Accessed on: May 21, 2020, DOI: 10.3390/s151229858, [Online].

[47] M. Gubian and D. Petri, "Uncertainty analysis of learning-from-examples algorithms," in *Proc. AMUEM*, Trento, Italy, 2008, pp. 34–39.

[48] D. Cook and N. Krishnan, *Activity learning: Discovering, recognizing, and predicting human behavior from sensor data*. Hoboken, New Jersey, USA: John Wiley & Sons, Ltd, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119010258

[49] F. Li, K. Shirahama, M. Nisar, L. Köping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 3, p. 679, Feb. 2018, Accessed on: May 21, 2020, DOI: 10.3390/s18020679, [Online].

[50] J. Gordon and J. M. Hernández-Lobato, "Combining deep generative and discriminative models for bayesian semi-supervised learning," *Pattern Recognit.*, vol. 100, p. 107156, Apr. 2020, Accessed on: May 21, 2020, DOI: 10.1016/j.patcog.2019.107156, [Online].

[51] T. Jebara and M. Meila, "Machine learning: Discriminative and generative," *Math Intell.*, vol. 28, no. 1, pp. 67–69, Mar. 2006, Accessed on: May 21, 2020, DOI: 10.1007/BF02987011, [Online].

[52] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, p. 2839–2846, Sep. 2015, Accessed on: May 21, 2020, DOI: 10.1016/j.patcog.2015.03.009, [Online].

[53] D. Hemavathi and H. Srimathi, "Effective feature selection technique in an integrated environment using enhanced principal component analysis," *J. Ambient Intell. Human. Comput.*, pp. 1–10, Jan. 2020, Accessed on: May 21, 2020, DOI: 10.1007/s12652-019-01647-x, [Online].

[54] S. Fan, Y. Jia, and C. Jia, "A feature selection and classification method for activity recognition based on an inertial sensing unit," *Information*, vol. 10, no. 10, p. 290, Sep. 2019, Accessed on: May 21, 2020, DOI: 10.3390/info10100290, [Online].

[55] N. Capela, E. Lemaire, and N. Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients," *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Accessed on: May 21, 2020, DOI: 10.1371/journal.pone.0124414, [Online].

Angelica Poli was born in San Benedetto del Tronto, Italy, in 1993. She received the Master degree in Biomedical Engineering at Università Politecnica delle Marche, Ancona, Italy, in 2019. She started her Ph.D in Electronic Instrumentation and Measurements in 2019 at the Information Engineering Department (DII) of the same university. Her research is focused on automatic processing and assessment of human biometric measurement data and the evaluation of accuracy in the recognition of ADLs for lifelogging applications.

Gloria Cosoli was born in Chiaravalle, Ancona, AN, Italy in 1989. She received the B.S. degree in biomedical engineering and M.S. degree in electronic engineering from the Università Politecnica delle Marche (UNIVPM), Ancona, in 2011 and 2013, respectively. She received the Ph.D. degree in mechanical engineering from the same university in 2017. From 2016 to date, she has been a Postdoctoral Research Fellow with the Department of Industrial Engineering and Mathematical Sciences (DIISM) of UNIVPM. She is the author/coauthor of 7 articles and 22 conference proceedings. Her research interests include non-invasive physiological measurements, numerical modelling, mechanical measurements, signal processing and NDT. Dr. Cosoli was a recipient of IEEE MeMeA 2015 Best Poster award.

Lorenzo Scalise (M'10) was born in Siena, SI, Italy in 1971. He received the M.S. degree in electronic engineering from Università degli Studi di Ancona, Ancona, in 1996, and the Ph.D. degree in mechanical measurements from Università degli Studi di Padova, Padova, in 1999. In 1999, he was a visiting researcher at the University of Twente (The Netherlands). From 2000 to 2015, he has been Research Assistant with the Department of Industrial engineering and Mathematical Sciences, at the Marche Polytechnic University, Ancona (Italy). Since 2015, he is Associate Professor in the same department, teaching Mechanical Measurement and Biomedical Instrumentation, Applied Measurement Techniques, Electronic Systems in Radiology and Electrical Measurement, at the Faculties of Engineering and Medicine. He is the author of more than 50 papers, 130 conference proceedings, 13 chapters in international books and 4 international patents. His research interests are in the field of measurement techniques (mechanical and electronic), with special focus on sensing systems, biomedical instrumentation, assistive technologies, e-health, and characterization of systems and materials. He is member of IEEE (Instrumentation and Measurement Society), of the International Society for Optics and Photonics (SPIE), of the International Academy of Laser medicine and Surgery (IALMS) and of the Society of Experimental Mechanics (SEM).

Susanna Spinsante (M'03, SM'13) is currently a tenured Assistant Professor in Electrical and Electronic Measurements at the Information Engineering Department (DII) of Università Politecnica delle Marche. She received her PhD in Electronics and Telecommunications Engineering in 2005 from the same University, where she spent several years working in signal processing for telecommunications and video applications. Since 2012 her research interests are focused on the use of RGBD sensors for applications related to human monitoring, motion-related measurements, action recognition, and Active and Assisted Living. She co-authored more than 180 papers in international peer reviewed journals and conference proceedings. She is involved in two European projects, co-funded by AAL JP and MYBL JP, focused on the use of RGBD and wearable sensors for human lifelogging and action understanding. She is the Chair of the Technical Committee on RGBD sensors within the IEEE Sensors Council Italy Chapter.