Computer Methods and Programs in Biomedicine 219 (2022) 106753

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Automated classification of hand gestures using a wristband and machine learning for possible application in pill intake monitoring

Sara Moccia[a,b], Sarah Solbiati[c,d], Mahshad Khornegah[c], Federica FS Bossi[c], Enrico G Caiani[c,d,*]

[a] The Biorobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy
[b] Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy
[c] Department of Electronics, Information and Biomedical Engineering, Politecnico di Milano, P.zza L. da Vinci 32, Milan 20133, Italy
[d] Institute of Electronics, Computer and Telecommunication Engineering (IEIIT), National Research Council of Italy (CNR), Milan, Italy

ARTICLE INFO

Article history:
Received 1 May 2021
Revised 9 March 2022
Accepted 11 March 2022

Keywords:
Hand-gesture classification
Monitoring medical adherence
Wearable sensors
Machine learning
Deep learning
CNN-LSTM

ABSTRACT

Background: Thanks to the increased interest towards health and lifestyle, a larger adoption in wearable devices for activity tracking is present among the general population. Wearable devices such as smart wristbands integrate inertial units, including accelerometers and gyroscopes, which can be utilised to perform automatic classification of hand gestures. This technology could also find an important application in automatic medication adherence monitoring. Accordingly, this study aims at comparing the performance of several Machine-Learning (ML) and Deep-Learning (DL) approaches for the automatic identification of hand gestures, with a specific focus on the drinking gesture, commonly associated to the action of oral intake of a pill-packed medication.

Methods: A method to automatically recognize hand gestures in daily living is proposed in this work. The method relies on a commercially available wristband sensor (MetaMotionR, MbientLab Inc.) integrating tri-axial accelerometer and gyroscope. Both ML and DL algorithms were evaluated for both multi-gesture (drinking, eating, pouring water, opening a bottle, typing, answering a phone, combing hair, and cutting) and binary gesture (drinking versus other gestures) classification from wristband sensor signals. Twenty-two participants were involved in the experimental analysis, performing a 10 min acquisition in a laboratory setting. Leave-one-subject-out cross validation was performed for robust performance assessment.

Results: The highest performance was achieved using a convolutional neural network with long- short term memory (CNN-LSTM), with a median f1-score of 90.5 [first quartile: 84.5; third quartile: 92.5]% and 92.5 [81.5;98.0]% for multi-gesture and binary classification, respectively.

Conclusions: Experimental results showed that hand gesture classification with ML/DL from wrist accelerometers and gyroscopes signals can be performed with reasonable accuracy in laboratory settings, paving the way for a new generation of medical devices for monitoring medical adherence.

1. Introduction

Thanks to the advancement in technology and circuit miniaturization, multiple sensors are nowadays embedded into a number of mobile and wearable devices, such as smartphones, smart watches and fitness bands. Specifically, inertial units including accelerometer and gyroscope sensor, are usually integrated in such devices and used in multiple tasks related to health. For example, inertial sensors are utilised for tracking human activities, such as the time spent in bed, the distance walked or the number of steps climbed [1]. In addition, the potential of inertial sensors for performing hand gestures classification, specifically in the field of human computer interaction [2–4], as well as for monitoring changes in daily behaviour in elders [5], have been explored. To these aims, Machine Learning (ML) methods have been extensively adopted for the classification of human activities [6,7] and hand gestures using signals acquired from the wearables devices.

In the field of healthcare, following the trend of increased adoption of wearable devices for activity tracking, due to the growing attention of people towards health and lifestyle [8,9], also few solutions to monitor medication adherence by recognizing hand gestures were proposed in the past, but without using ML methods for

* Corresponding author at: Department of Electronics, Information and Biomedical Engineering, Politecnico di Milano, P.zza L. da Vinci 32, Milan 20133, Italy.
E-mail address: enrico.caiani@polimi.it (E.G. Caiani).

https://doi.org/10.1016/j.cmpb.2022.106753
0169-2607/

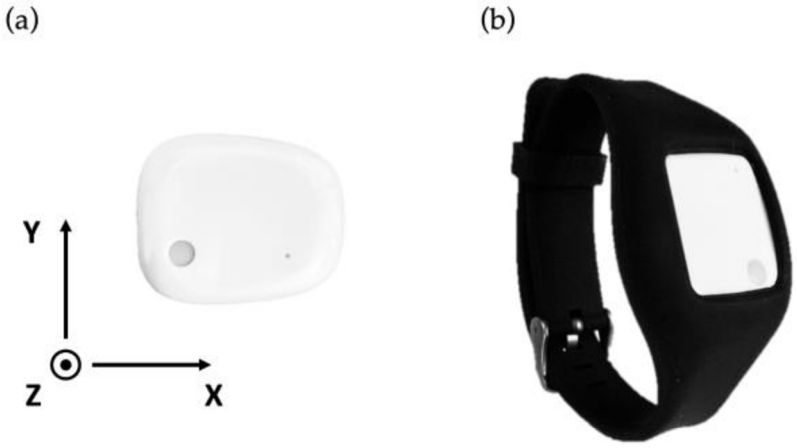

Fig. 1. MetaMotionR device (MMR) (MBIENTLAB INC, San Francisco, CA, USA): (a) Axis orientation of the inertial sensors embedded in the MMR; (b) Sensor integrated in the provided rubber WatchBand.

analysis. In [10], a wrist-worn commercial smartwatch was used to detect gestures as drinking water, picking pills, holding pills and taking them to the mouth, relying on extracting signal features from an embedded tri-axis accelerometer.

Similarly, Wang et al. [11] proposed a method based on dynamic time warping analysis of the data generated by accelerometers embedded in two wristwatches, worn one on each hand, to detect the gesture of taking empty gelatine capsules, drinking water and wiping mouth. However, in both of these solutions, only accelerometer data was used, and number of gestures or subjects investigated was very limited.

Medication non-adherence constitutes indeed a substantial issue worldwide. The World Health Organization reports that, in developed countries, approximately 50% of patients suffering from one or more chronic diseases does not take medications as prescribed, ultimately leading to increased morbidity and mortality [12], as well as to increased emergency-room visits, hospitalization and hospital readmissions [13,14]. This contributes in increasing the financial burden on the health care system. Lack of adherence has been estimated to provoke about 125,000 deaths in the United States with associated costs for the healthcare system being between $100 billion and $289 billion per year [15].

We hypothesized that wrist-worn devices allowing acquisition of both accelerometer and gyroscope data, together with their analysis using ML methods, including also Deep Learning (DL) techniques, could offer novel opportunities for gesture identification oriented towards the implementation of solutions to support drug adherence.

Accordingly, our aim was to test and compare the performance of several ML methods, including also two DL techniques, in both binary and multi-class classification of common hand gestures from both accelerometer and gyroscope signals acquired by a commercially-available wristband, in a relatively large (i.e., compared to similar studies) number of subjects using a specifically designed protocol. Specific attention was given to the classification of the drinking gesture, as drinking is commonly associated to the action of oral intake of a pill-packed medication and, as observed in literature, it is accurately identifiable among other activities using wrist-worn devices [11,15–17]. In addition, other hand gestures implying the movement of the hand towards the head, such as eating, combing hair and answering the phone, were included in the protocol to verify the robustness of the proposed methods in not confusing them with the drinking action.

2. Materials and methods

2.1. Wrist monitoring device

The MetaMotionR (MMR) wrist wearable device (MBIENTLAB INC, San Francisco, CA, USA) was used to track human gestures. The device is light and comfortable, with a USB rechargeable battery, and can be easily used during daily activities. It features ultra-low power consumption, providing energy efficient smartphone communication and central processing, and it embeds a tri-axial accelerometer, a tri-axial gyroscope, an ambient light sensor, and a humidity sensor. In Fig. 1, the MMR device is shown together with the schematization of the orientation of the axis of its embedded inertial sensors.

In this work, only the accelerometer and gyroscope signals were used. The accelerometer has a maximum resolution of 16 bit, and the gyroscope of 2000°/s. A sampling frequency of 50 Hz was selected for both sensors. The acquired signals and the corresponding timestamp were stored in the memory of a smartphone to which the MMR was connected through Bluetooth by using the MetaBase App (MBIENTLAB INC, San Francisco, CA, USA), available for both Android and iPhone Operating System devices.

2.2. Study population and acquisition protocol

The study was approved by the Ethical Committee of Politecnico di Milano. Twenty-two healthy subjects including both men and women (mean ± standard deviation, 29 ± 12 years, age range 22÷61) voluntarily participated in the experiment after signing a written informed consent form.

The acquisition protocol was designed to investigate the problem of automated classification of hand activities, with a specific focus on actions that could be related to the oral intake of a pill-packed medication, from the acquired signals in a laboratory setting. The subjects were asked to sit in a comfortable position and wear the MMR wrist monitor on their dominant hand (DH).

The main protocol (Protocol 1) was chosen to include eight gestures, some of which could be related to pill- intake, such as drinking, opening a bottle and pouring water (as preparatory actions). Other gestures implying the movement of the hand towards the head, such as eating, combing hair and answering the phone, were included in the protocol to verify the robustness of the proposed methods in not confusing them with the drinking action. Finally, two additional gestures (typing a keyboard and cutting paper) were added as traditionally included in such protocols [5,18,19], to

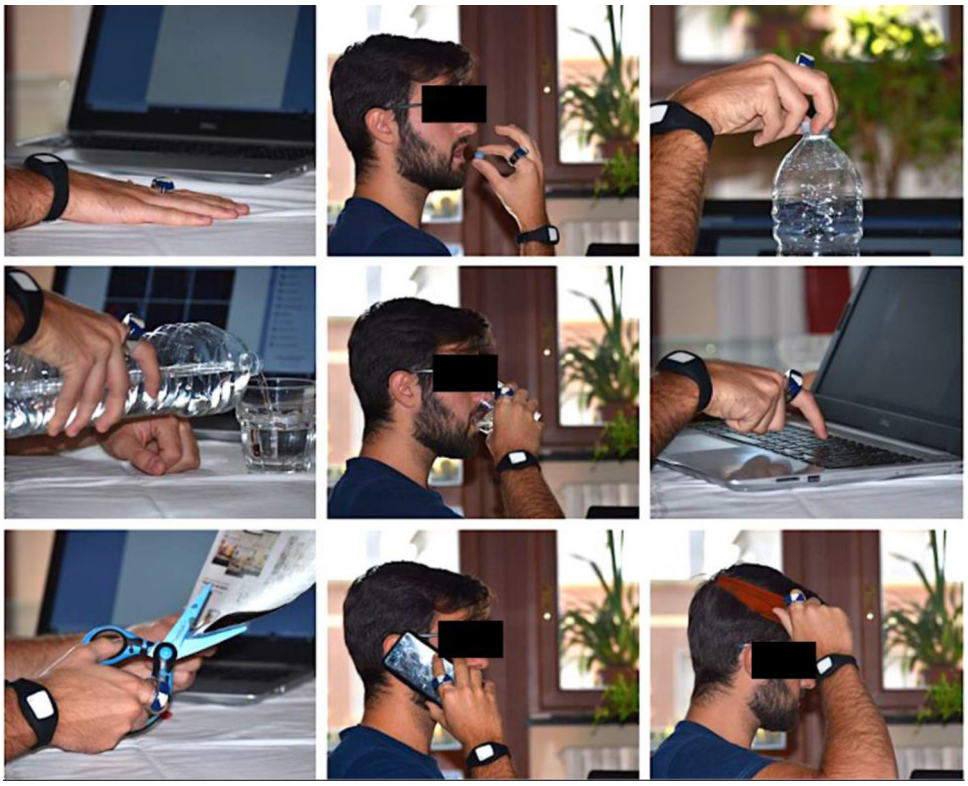

Fig. 2. Hand gestures studied in this work. From left to right/ up to down: Hand idle (performed between gestures), eating, opening a bottle, filling a glass, drinking water, typing, cutting, answering the phone and combing.

possibly allow comparisons with the performance of other studies. In Fig. 2, a picture describing each of the eight gestures, plus idle, is shown:

1. *Drinking*: the subject takes a glass of water by his/her DH, drinks an amount of water, and then puts it back on the table.
2. *Eating*: the subject takes an almond and brings it to the mouth, to simulate pill taking.
3. *Opening a bottle*: the subject opens a bottle cap using the DH and puts the cap on the table.
4. *Pour water*: using always the DH, the subject takes an opened bottle, pours an amount of water in a glass, and puts the bottle back on the table.
5. *Typing*: the subject types at least ten characters on a computer keyboard using the index finger of the DH.
6. *Answering the phone*: the subject picks up the phone from the table using the DH, takes it up to the ear and holds it for 3–5 s, then puts it back on the table.
7. *Combing hair*: with the DH, the subject picks up a comb from the table, combs hair for a few seconds and then puts the comb back on the table.
8. *Cutting*: the subject takes a piece of paper from the table, while holding a pair of scissors with the DH. Then he/she cuts the paper for about 3 or 4 times, and puts the scissors and the paper back on the table.

The subjects were asked to perform the eight gestures in a random order during a 10 min acquisition, in which each gesture lasted for a 30 s interval (timing was monitored by the supervising researcher), and keeping their hand still in an idle position between two consecutive gestures. The protocol further required the subjects to perform each gesture at least once, and to drink at least twice, in the 10 min acquisition. There were no restrictions in the modality of performing the activities. During the acquisition session, the sequence of actions chosen by the subjects was annotated by the supervising researcher.

A second protocol (Protocol 2), focused on increasing the numerosity of the dataset in particular for the drinking gesture, was also designed: in a 2 min acquisition, the subject needed to repeat four times the drinking gesture, each time drinking from a cup a different quantity of water as follows:

60 ml of water in eight sips.
45 ml of water in six sips.
30 ml of water in four sips.
15 ml of water in two sips.

Seventeen subjects out of the enrolled twenty-two accepted to perform also this second protocol.

At the end of the acquisition session, each subject was asked whether the device or the environmental factors (i.e., laboratory-controlled settings, presence of a supervisor) affected performance during the tests.

2.3. Gestures classification

The pipeline for gesture classification involved pre-processing, signal windowing, feature extraction (for ML methods only) and classification, as schematized in Fig. 3.

The pre-processing step consisted of a fourth-order low-pass Butterworth filtering with a cut-off frequency of 5 Hz. Such frequency was chosen due to the frequency content of the acquired signals, which was below 5–6 Hz, as verified by Fourier power spectrum analysis. The pre-processing also included raw signal standardization: each signal acquired with the MMR sensors was centered to have zero mean and standard deviation equal to one.

After the filtering procedure, the portions of signals containing the idle gesture were automatically identified and removed from the signal, thus obtaining a new signal in which the gestures appeared concatenated, with no idle in- between. Subsequently, a sliding window technique was applied, using fixed-size window segments, referred to as temporal windows, as commonly per-

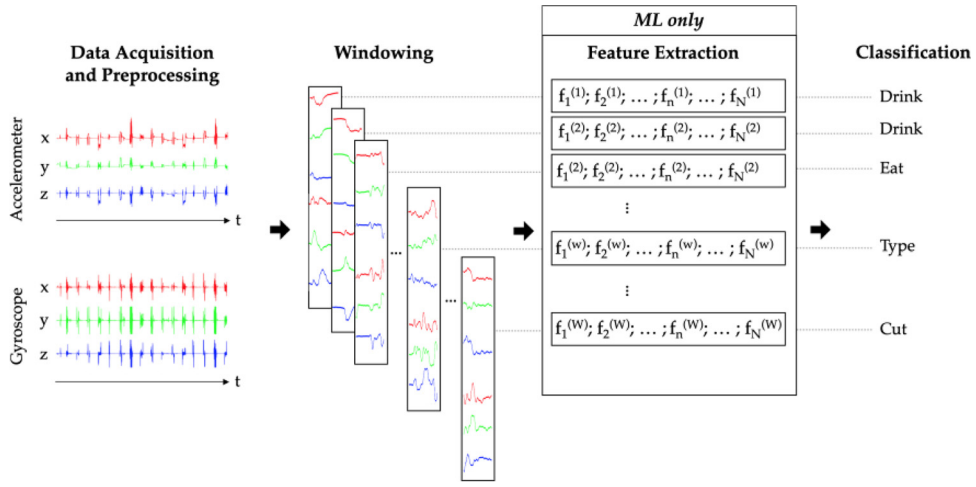

Fig. 3. Schematization of the proposed gesture classification pipeline, including: data acquisition and pre-processing, signal windowing, feature extraction (for ML methods only), and classification. In feature extraction, $f_n(w)$ represents the n-th feature of the feature vector f obtained for each window w.

Table 1
Features in time and frequency domains used for gesture recognition.

	Features
Time domain	Root mean square (RMS)
	Variance
	Mean absolute deviation (MAD)
	Kurtosis skewness
	Interquartile range (IQR)
Frequency domain	Energy
	Spectral entropy
	Mean frequency of power spectrum
	Median frequency of power spectrum

Table 2
Tuned hyperparameters for each classifier with corresponding grid-search values.

Classifier	Hyperparameter(s)	Grid-search values
KNN	Number of neighbours	[5,10]
SVM	Gaussian kernel size	$[10^{-7}, 10^{-3}]$
	Regularization parameter	$[10^{-3}, 10^{3}]$
RF	Number of trees	50, 100, 150
	Maximum depth of each tree	10, 15, 20, 40

Abbreviations: KNN = K-Nearest Neighbour; SVM = Support Vector Machine; RF = Random Forest.

formed in the literature [20]. Previous works on the classification of human activities adopted a 1 s window length [21,22] with 40 to 50% overlap between consecutive windows [21–24]. Such studies included the classification of different activities, ranging from drinking water to walking, watching TV and picking objects. In this work, in order to select the most appropriate window length, the lengths of 2 s, 3 s and 6 s were evaluated. After preliminary testing, the window size (i.e., the length of the data segments used for classification) of 3 s (corresponding to 150 data points with the sampling frequency of 50 Hz) was chosen. Moreover, being short-duration hand gestures included in the protocol, a window overlap of 75% was considered. Since only the windows containing a portion of the signal relevant to one single gesture were considered, while portions containing the transition between two gestures were automatically discarded, the selected overlap was useful in limiting the possibility to miss large signal portions over the transition between two consecutive activities, and thus possibly improving classification accuracy.

Afterwards, for each window, features extraction was performed to be utilized in ML algorithms, as they require handcrafted features. In Table 1, the list of the selected time and frequency domain features, which are mostly used in research work concerning activity recognition [6,7,25], is presented.

Both for the accelerometer and gyroscope signals, each feature was computed for the three axes separately, and for their modulus. The goal of the classification step was to assign a label to each temporal window. Two different classification problems were considered:

Simultaneous classification of all the eight gestures (multi-gesture classification problem)

Classification of the of the drinking gesture among all the other gestures (binary classification problem).

The classification problems were investigated by both ML and DL approaches.

2.4. Gestures classification with ML

Concerning ML approaches, inspired by human activity recognition work in the literature, the following classifiers were evaluated: Support Vector Machine (SVM) [19,20], Random Forest (RF) [21] and K-Nearest Neighbour (KNN) [22]. These ML approaches processed the features listed in Table 1. Grid search-based 5-fold cross-validation was used for hyperparameter tuning, as detailed in Table 2, using the highest f1-score as decision metric. The Least Absolute Shrinkage and Selection Operator (LASSO) [23] linear model with iterative fitting along a regularization path was used for automated feature selection [28]. The ML classification process was implemented in Python using the open-source machine learning library Scikit-learn (http://scikit-learn.org/stable/index.html).

2.5. Gestures classification with DL

Two DL models were additionally investigated for the gesture classification task. First, a convolutional neural network (CNN) based on the model proposed in [29] was implemented. The input shape and the number of layers of the CNNs for the multi-gesture and binary gesture classifications have been adapted to the signals acquired for this study. The relevant architecture is detailed in Table 3. The input shape of both multi-gesture and binary CNNs was 150 (i.e., the number of data points in a window) times 6 (number of channels, corresponding to the 3 axes of the

Table 3
Architecture of the CNN for gesture classification. Both the multi-gesture and binary classification CNNs share the same architecture up to layer 7. The different top layers for the multi-gesture and binary classification are highlighted in italics. TdC: Time distributed Convolution; FC: Fully Connected.

Layer	Type	Feature maps	Input shape	Output shape	k	s
Layer 1	Convolution + ReLU	100	(None, 150,6)	(None, 148, 100)	3	1
Layer 2	Convolution + ReLU	150	(None, 148, 100)	(None, 146, 150)	3	1
Layer 3	Convolution + ReLU	150	(None, 146, 150)	(None, 144, 150)	3	1
Layer 4	Dropout	_	(None, 144, 150)	(None, 144, 150)	1	1
Layer 5	Max pooling	_	(None, 144, 150)	(None, 48, 150)	3	3
Layer 6	Flatten	_	(None, 48, 150)	(None, 7200)	1	1
Layer 7	FC + ReLU	_	(None, 7200)	(None, 1000)	1	1
Layer 8 - *Multi-gesture*	Dropout	_	(None, 1000)	(None, 1000)	1	1
Layer 9 - *Multi-gesture*	FC + ReLU	_	(None, 1000)	(None, 500)	1	1
Layer 10 - *Multi-gesture*	Dropout	_	(None, 500)	(None, 500)	1	1
Layer 8 - *Binary*						
Layer 11 - *Multi-gesture*	FC + Softmax	_	(None, 500)	(None, 8)	1	1
Layer 9 - *Binary*	FC + ReLU	_	(None, 500)	(None, 200)	1	1
Layer 10 - *Binary*	Drop out	_	(None, 200)	(None, 200)	1	1
Layer 11 - *Binary*	FC + ReLU	_	(None, 200)	(None, 100)	1	1
Layer 12 - *Binary*	Drop out	_	(None, 100)	(None, 100)	1	1
Layer 13 - *Binary*	FC + Softmax	_	(None, 100)	(None, 2)	1	1

Abbreviations: ReLU = Rectifying Linear Unit; k = kernel size; s = stride.

Table 4
Architecture of the CNN-LSTM for gesture classification. Both the multi-gesture and binary classification CNN-LSTMs share the same architecture until layer 7. The different top layers for the multi-gesture and binary classification are highlighted in italics. TdC: Time distributed Convolution; FC: Fully Connected.

Layer	Type	Feature maps	Input shape	Output shape	k	s
Layer 1	TdC + ReLU	100	(None,None, 30, 100)	(None,None, 28, 100)	3	1
Layer 2	TdC + ReLU	150	(None,None, 28, 100)	(None,None, 26, 150)	3	1
Layer 3	TdC + ReLU	150	(None,None, 26, 150)	(None,None, 24, 150)	3	1
Layer 4	Drop out	_	(None,None, 24,150)	(None,None, 24, 150)	1	1
Layer 5	Max pooling	_	(None,None, 24, 150)	(None,None, 8, 150)	3	3
Layer 6	Flatten	_	(None,None, 8, 150)	(None,None, 1200)	1	1
Layer 7	LSTM	_	(None,None, 1200)	(None, 150)	1	1
Layer 8 - *Multi-gesture*	FC + ReLU	_	(None, 150)	(None,1000)	1	1
Layer 8 - *Binary*	FC + ReLU	_	(None, 150)	(None, 500)		
Layer 9 - *Multi-gesture*	Drop out	_	(None,1000)	(None,1000)		
Layer 10 - *Multi-gesture*	FC + ReLU	_	(None,1000)	(None, 500)		
Layer 11 - *Multi-gesture*	Drop out	_	(None, 500)	(None, 500)	1	1
Layer 9 - *Binary*						
Layer 12 - *Multi-gesture*	FC +Softmax	_	(None, 500)	(None, 8)	1	1
Layer 10 - *Binary*	FC + ReLU		(None, 500)	(None, 200)		
Layer 11 - *Binary*	Drop out		(None, 200)	(None, 200)	1	1
Layer 12 - *Binary*	FC + ReLU	_	(None, 200)	(None, 100)	1	1
Layer 13 - *Binary*	Drop out	_	(None, 100)	(None, 100)	1	1
Layer 14 - *Binary*	FC + Softmax	_	(None, 100)	(None, 2)	1	1

Abbreviations: ReLU = Rectifying Linear Unit; LSTM = Long-Short Term Memory; k = kernel size; s = stride.

accelerometer and gyroscope). Both CNNs shared the same backbone, where each convolutional layer was activated by a rectifying linear unit (ReLU) function. Temporal max pooling and dropout (with probability = 0.5) were utilized to prevent overfitting. For the multi-gesture classification CNN, a fully connected layer with 8 neurons was used, where 8 is the number of different gestures to be classified. On the contrary, for the binary classification CNN, a fully connected layer with 2 neurons was used, in addition to a further fully connected layer with dropout, to attenuate the overfitting issues expected when moving from 500 to 2 neurons.

In addition to the CNN model, a second hybrid model inspired by [30] was investigated, in which CNN and long short-term memory (LSTM) are combined Table 4. reports the architecture of the CNN-LSTM models implemented for the multi-gesture and binary gesture classification tasks. Specifically, those were built by adding an LSTM layer on top of the convolutional part of the CNNs described in Table 3. Indeed, maintaining the architecture of the CNN part unchanged, a fair comparison between the CNN and CNN-LSTM models was possible. The input of the CNN- LSTM consisted of a temporal window split into 5 sequences of equal length. For processing such sequences, time-distributed convolution was used.

For both the CNN and CNN-LSTM, Adam optimizer was used, with the cross-entropy loss function. The best model among epochs was chosen according to the highest accuracy on the validation set (20% of the training set). The Python Keras library (https://keras.io/) was used for training and testing the DL models.

2.6. Validation protocol

Fig. 4 summarizes the flow of how the original data were used in the testing procedure. Specifically, leave- one-subject-out cross-validation was used to validate both the ML and DL approaches. As performance metrics, precision, recall, f1-score and the precision-

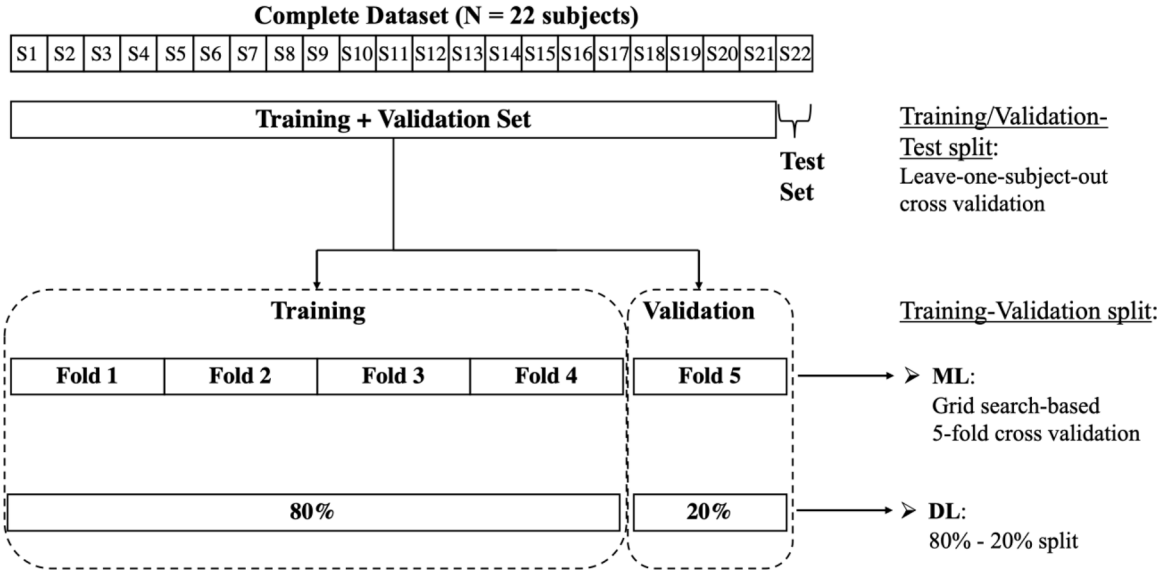

Fig. 4. Schematization of the validation procedure. The complete dataset, composed of 22 subjects, was divided into training/validation and test sets using leave-one-subject-out cross validation. As concerns ML methods, the training/validation dataset was split into 5 folds, and grid search-based 5-fold cross validation was used for tuning the hyperparameters. As concerns DL methods, validation was performed on 20% of the training/validation set.

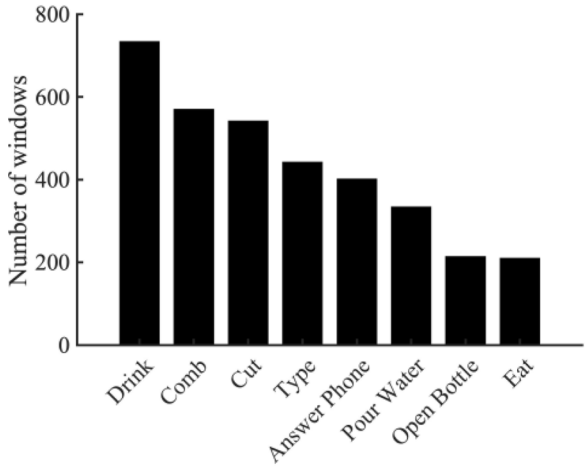

Fig. 5. Data distribution for each gesture in the available dataset.

recall (PR) curve were computed. In addition, the balanced accuracy was calculated.

2.7. Statistical analysis

With the aim to compare methods' performance, the non-parametric Friedman test ($p < 0.05$) was applied to the values of f1-score, precision and recall obtained for each classifier (H0: no differences among methods). In case the null hypothesis was rejected, the post-hoc Wilcoxon Signed Rank test with Bonferroni correction was performed for additional paired comparisons.

3. Results

All the enrolled subjects declared that neither the laboratory setting nor the wristband device influenced their gesture performance during the test.

Fig. 5 shows the final number of 3 s duration time windows acquired representing each gesture, with 'drink', 'comb' and 'cut' being the most frequent. Both ML and DL methods provided good classification results in terms of f1-score, precision, recall and balanced accuracy, as shown in Table 5. Among ML methods, SVM resulted in higher classification outcomes, significantly outperforming KNN in terms of f1-score (multi-gesture: 83.5 [78.0; 91.5]% versus 82.0 [76.5; 89.0]%; binary: 87.5 [79.5; 93.5]% versus 82.5 [75.5; 87.0]%) and precision (binary: 83.0 [75.3; 90.5]% versus 71.0 [60.0; 81.0]%).

In multi-gesture classification, CNN and CNN-LSTM performed better than ML methods, with CNN-LSTM resulting in the highest balanced accuracy (89.0 [84.0; 92.8]%). Remarkably, the CNN-LSTM resulted in significantly better performance compared to the CNN in terms of precision (92.0 [88.0; 93.3]% versus 91.0 [85.8; 92.3]%) and recall (90.0 [85.0; 92.5]% versus 88.0 [82.5; 92.3]%).

In the binary classification, the highest f1-score and precision values were obtained with CNN (92.5 [86.0;97.5]% and 94.0 [82.3; 100]%, respectively) and CNN-LSTM (92.5 [81.5; 98.0]% and 94.0 [83.0; 97.0]%, respectively), especially when compared to KNN (82.5 [75.5; 87.0]% and 71.0 [60.0; 81.0]%, respectively), with SVM and CNN- LSTM resulting in the highest values of balanced accuracy (96.3 [92.6; 97.5]% and 96.3 [91.1; 99.2]%, respectively).

The PR curves obtained with CNN-LSTM are reported in Fig. 6. In the multi-gesture classification problem, the gestures "Type", "Comb" and "Cut" resulted with the highest area under the PR curve (0.96, 0.65 and 0.95, respectively), which was equal to 0.92 for the "Drink" gesture. Interestingly, this value increased up to 0.98 for the binary classification.

4. Discussion

In this paper, a comparison of the performance of several ML and DL methods for gesture classification from hand activity captured through inertial sensors by a wrist wearable device was presented, with possible applications in the context of measuring medication adherence by recognizing gestures that are related to the pill intake. The gestures in the experimental protocol included both actions related to pill intake (opening a bottle, pouring water and drinking) as well as possible confounding movements (eating, combing hair and answering the phone) and other gestures (typing a keyboard and cutting paper). In this way, the considered ML and DL methods were challenged to properly discriminate each activ-

Table 5
Performance metrics for the tested classifiers for multi-gesture and binary classification. Median is reported with 1st and 3rd quartile in brackets.

	Classifier	f1-score (%)	Precision (%)	Recall (%)	Balanced Accuracy
Multi-gesture	SVM	83.5 [78.0;91.5]	85.5 [82.3;93.0]	84.5 [78.5;91.3]	81.5 [77.5;91.5]
	KNN	82.0 [76.5;89.0] *	85.5 [80.0;89.5]	83.5 [76.0;89.0]	81.0 [75.3;85.8]
	RF	78.5 [75.0;90.0]	82.0 [79.0;91.0]	79.5 [74.8;89.3][a]	79.0 [73.5;87.5]
	CNN	88.5 [82.0;92.0][a,b]	91.0 [85.8;92.3]*,a,b	88.0 [82.5;92.3]*,b	87.5 [80.5;91.0][a]
	CNN-LSTM	90.5 [84.5;92.5]*,b	92.0 [88.0;93.3][c]	90.0 [85.0;92.5]*,a,c	89.0 [84.0;92.8]*,a,b
Binary	SVM	87.5 [79.5;93.5]	83.0 [75.3;90.5]	100.0 [86.5;100.0]	96.3 [92.6;97.5]
	KNN	82.5 [75.5;87.0] *	71.0 [60.0;81.0]*	100.0 [93.3;100.0]	93.2 [90.5;95.3]
	RF	88.0 [77.0;91.0]	83.0 [72.5;93.0][a]	91.5 [84.3;97.5]*,a	93.9 [82.6;95.5]*
	CNN	92.5 [86.0;97.5] [a]	94.0 [82.3;100.0]*,a,b	96.0[88.5;100.0]	95.7 [92.4;98.7]
	CNN-LSTM	92.5 [81.5;98.0] [a]	94.0 [83.0;97.0]*,a	95.5[85.0;100.0]	96.3 [91.1;99.2]

Results of the post-hoc Bonferroni test ($p<0.05/n$, with $n = 10$) performed for f-score, Precision and Recall are reported as:
* vs SVM;
[a] vs KNN;
[b] vs RF;
[c] vs CNN.

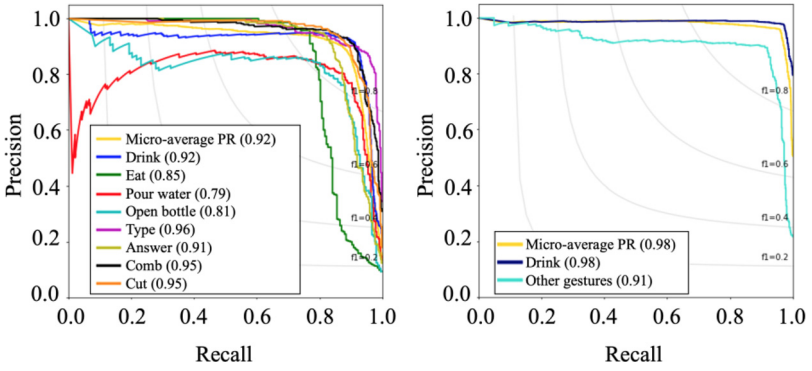

Fig. 6. Data distribution for each gesture in the available dataset.

ity in a multiclass classification task, as well as to correctly classify drinking versus all the other gestures in a binary classification task.

ML classifiers included SVM, KNN and RF, among which SVM performed slightly better, while the other two had comparable performance. Successful performance of SVM was also observed in previous work, such as [19] and [22], in which it outperformed KNN and Naïve Bayes. In both multi-gesture and binary classification problems, DL-based approaches, which included a CNN and a CNN-LSTM, outperformed the ML ones. This result is in line with other similar research [26,27], and possibly explained by the ability of neural networks to extract relevant features different from the manual hand-crafted ones, as CNNs could learn the internal relationships present in the dataset.

While no significant difference between the two CNN models was highlighted in the binary problem, CNN-LSTM hybrid model achieved a higher performance in multi-gesture classification, and a higher balanced accuracy in the binary problem. This result could be attributed to its ability to handle high-dimensional feature-space (which was high if compared with the number of subjects in the dataset), as well as to its robustness in tackling the noise components of the accelerometer and gyroscope signals. Also, LSTM allows the processing of the temporal information naturally encoded in the signals and thus improving CNN results, as observed in [25]. Among the analysed gestures, some were classified particularly well, such as 'drinking', 'typing', 'combing' and 'cutting', while others ('pouring water' and 'opening the bottle') were classified with less precision.

In this work, specific attention was given to the 'drinking' gesture, as the natural action closer to the possible pill intake, but that could have its own interest also in the context of remote monitoring of hydration conditions in elderly [33–35] or heart failure patients [36]. Accordingly, a binary classification problem was introduced in order to test the ability of the proposed methods to distinguish the 'drinking' gesture from all the other gestures. Particularly, both DL models and SVM outperformed the KNN and RF. The introduction of the binary classification came from the observation that, in the multi-gesture classification problem, the drinking gesture appeared as the best-classifiable medication adherence-related gesture, and it represents an element of novelty compared to the current literature on medication adherence monitoring.

In Table 6, the results obtained in this work and in similar studies using ML methods for solving a multi-classification problem are reported. Compared to [5], where the same number of subjects was studied, the results obtained with the SVM model developed in this work were superior, both in terms of f1-score and accuracy. Also, the present work outperformed the results of [21] in the overall recall, while the precision was slightly lower; however, it is worth noticing that only 2 subjects were studied in [21]. On the contrary, in [22] higher values of precision and recall were achieved using the same wearable device (MMR wrist monitor) for the classification problem: this could be attributed to the development of a novel multi-step refinement with the aim of improving the classification accuracy, as well as to the lower complexity in terms of lower number of subjects ($n = 6$), and to the different kinds of activities classified, including standing, sitting and walking. In a recent study, Chun and collaborators [37] performed a classification of the drinking gesture versus non-drinking, obtaining the best results using the RF model. Their outcomes, in terms of recall and f1-score, were comparable to the results obtained in the present study as regards RF, though remaining inferior to the results we obtained with DL models. On the contrary, Ortega-Anderez and colleagues [38], in the 2-class classifica-

Table 6
Comparison of the results obtained using conventional ML models in the current state of the art.

Reference	# subjects	Sensor Type	Sensor placement	Activities	ML models	Balanced Accuracy	Precision	Recall	f1-score
This work	20	Acc Gyr	Wrist	Eat, drink, open a bottle, pour water, type, answer a phone, combing hair, cutting by scissors	SVM KNN RF	SVM 90% KNN 89% RF 89%	SVM 84% KNN 82% RF 82%	SVM 84% KNN 82% RF 82%	SVM 84% KNN 82% RF 82%
[15]	20	Acc Gyr	FingerWrist	Eat, drink, answer a phone, brush the teeth, brush hair, use a hiar dryer	SVM DT	SVM Wrist65% Both 92% DT Wrist67% Both 89%	–	–	SVM Wrist62% Both 91% DT Wrist67% Both 88%
[22]	2	Acc Gyr	Wrist, outer side of lower arm, outer side of upper arm	Opening and closing a window, watering a plant, turning book pages, Drinking from a bottle, cutting with a knife, chopping with a knife, stirring in a bowl, forehand, backhand and smash	SVM KNN NB	–	SVM 88.9% KNN 76.2% NB 75.7%	SVM 66.5% KNN 44.2% NB 56.6%	–
[31]	6	Acc	Wrist	Hand washing, Teeth brushing, Standing, Sitting, Picking up an object from the floor, Walking upstairs, Walking downstairs	SVM RF KNN	99.28%	94.43%	93.22%	–
[32]	30	Acc	Left and right wrist	Drink gesture versus non drinking (including watching a movie, eating, talking, brushing teeth, folding laundry, walking, browsing the news)	HMM, KNN, RF	–	RF 90.3%	RF 91.0%	RF >75.0% in all participants >90.0% in 20 out of 30 participants
[33]	6	Acc Gyr	Wrist	2-class: Null, Drinking or Eating	KNN, RF, SVM	2-class: RF 97.4%	2-class: RF 97.2%	2-class: RF 96.3%	–

Abbreviations: Acc = Accelerometer; Gyr = Gyroscope; SVM = Support Vector Machine; KNN = K-Nearest Neighbour; RF = Random Forest; DT = Decision Tree; NB = NaiveBayes; HMM = Hidden Markov Models.

Table 7. Comparison of the results obtained using DL models in the current the state of the art.

Reference	# subjects	Sensor type	Sensor placement	Activities	ML models	Balanced Accuracy	Precision	Recall	f1-score
This work	20	Acc Gyr	Wrist	Eat, drink, open a bottle, pour water, type, answer a phone, combing hair, cutting by scissors	CNN CNN-LSTM	CNN 92% CNN-LSTM 93%	CNN 88% CNN-LSTM 89%	CNN 87% CNN-LSTM 89%	CNN 87% CNN-LSTM 89%
[25]	4	Acc Gyr Mag	Upper Arms, wrists, hands, back, hip, knee	Open and close door, open and close fridge, open and close dishwasher, open and close drawer, clean table, drink from cup, Toggle switch, Groom, prepare coffee, Drink coffee, prepare Sandwich, eat sandwich, Clean up	CNN DC-LSTM	- 	- 	- 	CNN 78% DC-LSTM 86%
[26]	2	Acc Gyr	Wrist, outer side of lower arm, outer side of upper arm	Opening and closing a window, watering a plant, turning book pages, Drinking from a bottle, cutting with a knife, chopping with a knife, stirring in a bowl, forehand, backhand and smash	CNN DBN	CNN 95% DBN 84%	- 	- 	CNN 89.6% DBN 76%
[33]	6	Acc Gyr	Wrist	3-class: Null, Drinking, Eating 5-class: Null, Drinking, Spoon, Fork, Hand	ANN	3-class: ANN 98.2% 5-class: ANN 97.8%	3-class: ANN 95.7% 5-class: ANN 88.7%	3-class: ANN 95.0% 5-class: ANN 85.8%	-

Abbreviations: Acc = Accelerometer; Gyr = Gyroscope; Mag = Magnetometer; CNN = Convolutional Neural Network; LSTM = Long-Short Term Memory; DC = Deep Convolutional; DBN = Deep Belief Network; ANN = Artificial Neural Network.

tion of eating/drinking versus other gestures, obtained with the RF model a better performance compared to this study, both considering multi-gesture and binary (drink versus non-drink). This outcome could possibly depend on their choice to consider eating and drinking gestures as a single class.

Table 7 shows the comparison of the results obtained in this work and in similar studies using DL methods for solving multi-classification problems. From this analysis, the results obtained in this study with CNN and CNN- LSTM were comparable to [38] for the 3-class classification problems, and outperformed their 5-class classification. On the contrary, in this study the performance appears slightly inferior to [31], in which three sensors were used, placed in different positions along the two experimental subjects' arms, thus possibly improving the activity recognition accuracy. On the other hand, when compared to [30], this work showed higher values of f1-score in both CNN and CNN-LSTM models.

It is worth noticing that the number of subjects in our study was considerably larger than those in the previous works.

4.1. Limitations

In the acquisition protocol, only one activity was performed in a 30 s interval, with the hand being still between two consecutive gestures. This represents a simplification of a real-life scenario that would probably bring additional challenges. However, this study was conceived as a first feasibility study to test and compare the performance of different methods in multi-class and binary classification problems from the acquired signals from the wrist device. Future studies will tackle these more complex experimental conditions on the basis of the lesson learned and trained algorithms.

As a second limitation, all the subjects enrolled in the experiments were right-handed; for higher generalization; future studies should consider including left-handed subjects as well. Similarly, a larger number of subjects in different age ranges should be considered to avoid introducing possible biases.

5. Conclusion

In this work, the problem of automated classification of eight hand gestures using a wearable wrist-worn device was investigated. Both multi-gesture classification as well as binary classification of drinking against all the other gestures were taken into consideration. Three ML models (SVM, RF and KNN) commonly used in human activity recognition were tested using temporal and frequency features, with SVM resulting in the one achieving the best performance. In addition, two DL-based methods (CNN and CNN-LSTM) were applied. All the models showed good performances in classifying each activity, with the DL models outperforming the ML ones, and in particular the CNN-LSTM being the best performing model (median f1-score = 90.5% for the multi-gesture classification). All the models showed better performance for the binary classification of the 'drinking' gesture.

These results represent a promising step in the direction of developing automated solutions for gesture tracking with possible applications in specific healthcare domain in which the drinking action constitutes a possible indicator of clinically relevant patient's behaviour, such as pill intake. In addition to gesture tracking, such solutions could result in novel opportunities to support and promote behavioural changes through self-monitoring and personalized feedbacks, thus increasing user's motivation and engagement [39–41].

Declaration of Competing Interest

The authors declare no conflict of interest relevant to this work.

References

[1] I. Sim, Mobile devices and health, N. Engl. J. Med. 381 (10) (2019) 956–968, doi:10.1056/NEJMra1806949.

[2] S. Jiang, B. Lv, W. Guo, C. Zhang, H. Wang, X. Sheng, P.B. Shull, Feasibility of wrist-worn, real-time hand, and surface gesture recognition via sEMG and IMU sensing, IEEE Trans. Ind. Inf. 14 (8) (2017) 3376–3385, doi:10.1109/TII.2017.2779814.

[3] M. Kim, J. Cho, S. Lee, Y. Jung, IMU sensor-based hand gesture recognition for human-machine interfaces, Sensors 19 (18) (2019) 3827, doi:10.3390/s19183827.

[4] H. Abualola, H. Al Ghothani, A.N. Eddin, N. Almoosa, K. Poon, Flexible gesture recognition using wearable inertial sensors, in: Proceedings of the IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS), 2016, pp. 1–4, doi:10.1109/MWSCAS.2016.7870143.

[5] A. Moschetti, L. Fiorini, D. Esposito, P. Dario, F. Cavallo, Recognition of daily gestures with wearable inertial rings and bracelets, Sensors 16 (8) (2016) 1341, doi:10.3390/s16081341.

[6] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Commun. Surv. Tutor. 15 (3) (2012) 1192–1209, doi:10.1109/SURV.2012.110112.00192.

[7] S. Rosati, G. Balestra, M. Knaflitz, Comparison of different sets of features for human activity recognition by wearable sensors, Sensors 18 (12) (2018) 4189, doi:10.3390/s18124189.

[8] Deloitte, Global mobile consumer trendsMobile Continues Its Global Reach Into All Aspects of consumers', 2nd edition, Deloitte, 2017.

[9] W.R. Thompson, WR, Worldwide survey of fitness trends for 2020, ACSM's Health Fit. J. 23 (6) (2019) 10–18, doi:10.1249/FIT.0000000000000526.

[10] T. Putthaprasat, D. Thanapatay, J. Chinrungrueng, N. Sugino, Medicine intake detection using a wearable wrist device accelerometer, in: Proceedings of the International Conference on Computer Engineering and Technology, 2012, pp. 4–5.

[11] R. Wang, Z. Sitová, X. Jia, X. He, T. Abramson, P. Gasti, K.S. Balagani, A. Farajidavar, Automatic identification of solid-phase medication intake using wireless wearable accelerometers, in: Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 4168–4171, doi:10.1109/EMBC.2014.6944542.

[12] E. Sabaté, Adherence to Long-Term therapies: Evidence For Action, World Health Organization, 2003.

[13] A.E. Linkens, V. Milosevic, P.H. van der Kuy, V.H. Damen-Hendriks, C. Mestres Gonzalvo, K.P. Hurkens, Medication-related hospital admissions and readmissions in older patients: an overview of literature, Int. J. Clin. Pharm. 42 (2020) 1243–1251, doi:10.1007/s11096-020-01040-1.

[14] T.H. Lim, A.H. Abdullah, Medication adherence using non-intrusive wearable sensors, EAI Endorsed Trans. Ambient Syst. 4 (16) (2017), doi:10.4108/eai.19-12-2017.153484.

[15] H. Kalantarian, N. Alshurafa, M. Sarrafzadeh, Detection of gestures associated with medication adherence using smartwatch-based inertial sensors, IEEE Sens. J. 16 (4) (2016) 1054–1061, doi:10.1109/JSEN.2015.2497279.

[16] H. Kalantarian, N. Alshurafa, E. Nemati, T. Le, M. Sarrafzadeh, A smartwatch-based medication adherence system, in: Proceedings of the IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks, 2015, pp. 1–6, doi:10.1109/BSN.2015.7299348.

[17] K. Serdaroglu, G. Uslu, S. Baydere, Medication intake adherence with real time activity recognition on IoT, in: Proceedings of the IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications, 2015, pp. 230–237, doi:10.1109/WiMOB.2015.7347966.

[18] G. Laput, C. Harrison, Sensing fine-grained hand activity with smartwatches, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–13, doi:10.1145/3290605.3300568.

[19] D. Gomes, I. Sousa, Real-time drink trigger detection in free-living conditions using inertial sensors, Sensors 19 (9) (2019) 2145, doi:10.3390/s19092145.

[20] A. Jordao, A.C. Nazare Jr, J. Sena, W.R. Schwartz, Human activity recognition based on wearable sensor data: a standardization of the state-of-the-art, arXiv Computer Science (2018), preprint arXiv:1806.05226.

[21] A. Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors, ACM Comput. Surv. 46 (3) (2014) 33 (CSUR), doi:10.1145/2499621.

[22] D. Ortega-Anderez, A. Lotfi, C. Langensiepen, K. Appiah, A multi-level refinement approach towards the classification of quotidian activities using accelerometer data, J. Ambient Intell. Humaniz. Comput. 10 (11) (2019) 4319–4330, doi:10.1007/s12652-018-1110-y.

[23] L. Bao, S.S. Intille, Activity recognition from user-annotated acceleration data, in: In Proceedings of the International conference on pervasive computing, 2004, pp. 1–17, doi:10.1007/978-3-540-24646-6_1.

[24] S.J. Preece, J.Y. Goulermas, L.P. Kenney, D. Howard, A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data, IEEE Trans. Biomed. Eng. 56 (3) (2008) 871–879, doi:10.1109/TBME.2008.2006190.

[25] Y. Zhang, Y. Zhang, Z. Zhang, J. Bao, Y. Song, Human activity recognition based on time series analysis using U-Net, arXiv Computer Science (2018), arXiv preprint arXiv:1809.08113.

[26] Z. He, L. Jin, Activity recognition from acceleration data based on discrete consine transform and SVM, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2009, pp. 5041–5044, doi:10.1109/ICSMC.2009.5346042.

[27] K.M. Chathuramali, R. Rodrigo, Faster human activity recognition with SVM, in: Proceedings of the International Conference on Advances in ICT for Emerging Regions, 2012, pp. 197–203, doi:10.1109/ICTer.2012.6421415.

[28] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. R. Stat. Soc. Series B Methodol. 58 (1) (1996) 267–288, doi:10.1111/j.2517-6161.1996.tb02080.x.

[29] C.A. Ronao, S.B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks, Expert Syst. Appl. 59 (2016) 235–244, doi:10.1016/j.eswa.2016.04.032.

[30] F.J. Ordóñez, D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115, doi:10.3390/s16010115.

[31] J.B. Yang, M.N. Nguyen, P.P. San, X.L. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, in: Proceedings of the 24th International Conference on Artificial Intelligence, 2015, pp. 3995–4001.

[32] R.D. Gurchiek, N. Cheney, R.S. McGinnis, Estimating biomechanical time-series with wearable sensors: a systematic review of machine learning techniques, Sensors 19 (23) (2019) 5227, doi:10.3390/s19235227.

[33] J. Mentes, Oral hydration in older adults: greater awareness is needed in preventing, recognizing, and treating dehydration, Am. J. Nurs. 106 (6) (2006) 40–49, doi:10.1097/00000446- 200606000-00023.

[34] C.M. Sheehy, P.A. Perry, S.L. Cromwell, Dehydration: biological considerations, age-related changes, and risk factors in older adults, Biol. Res. Nurs. 1 (1) (1999) 30–37, doi:10.1177/109980049900100105.

[35] C. Lecko, Improving hydration: an issue of safety, Nurs. Resid. Care 10 (3) (2008) 149–150, doi:10.12968/nrec.2008.10.3.28593.

[36] P. Pellicori, K. Kaur, A.L. Clark, Fluid mManagement in patients with chronic heart failure, Card. Fail. Rev. 1 (2) (2015) 90–95, doi:10.15420/cfr.2015.1.2.90.

[37] K.S. Chun, A.B. Sanders, R. Adaimi, N. Streeper, D.E. Conroy, E. Thomaz, Towards a generalizable method for detecting fluid intake with wrist-mounted sensors and adaptive segmentation, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 80–85, doi:10.1145/3301275.3302315.

[38] D. Ortega-Anderez, A. Lotfi, A. Pourabdollah, Eating and drinking gesture spotting and recognition using a novel adaptive segmentation technique and a gesture discrepancy measure, Expert Syst. Appl. 140 (2020) 112888, doi:10.1016/j.eswa.2019.112888.

[39] J.A. Naslund, K.A. Aschbrenner, S.J. Bartels, Wearable devices and smartphones for activity tracking among people with serious mental illness, Ment. Health Phys. Act. 10 (2016) 10–17, doi:10.1016/j.mhpa.2016.02.001.

[40] L. Laranjo, D. Ding, B. Heleno, B. Kocaballi, J.C. Quiroz, H.L. Tong, B. Chahwan, A.L. Neves, E. Gabarron, K.P. Dao, D. Rodrigues, Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression, Br. J. Sports Med. 55 (8) (2021) 422–432, doi:10.1136/bjsports-2020-102892.

[41] E.C. Nelson, T. Verhagen, M.L. Noordzij, Health empowerment through activity trackers: an empirical smart wristband study, Comput. Human Behav. 62 (2016) 364–374, doi:10.1016/j.chb.2016.03.065.