Computer Methods and Programs in Biomedicine

Towards monitoring medical adherence using a wristband and machine learning

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Section/Category:	Classification and interpretation
Keywords:	Hand-gesture classification; monitoring medical adherence; wearable sensors; machine learning; deep learning; CNN-LSTM
Corresponding Author:	Enrico Gianluca Caiani Politecnico di Milano Milan, MIlan ITALY
First Author:	Sara Moccia, MS, PhD
Order of Authors:	Sara Moccia, MS, PhD
	Sarah Solbiati, MS
	Mahshad Khornegah, MS
	Federica Francesca Stefania Bossi, MS
	Enrico Gianluca Caiani
Manuscript Region of Origin:	Europe
Abstract:	Background: The demographic shift generated by the ageing of the world's population is having important consequences in the rise of chronic diseases. People with chronic diseases require long-term care and commitment to adhere to the prescribed medications, e.g. taking pills on a daily basis. Poor medication adherence is a common problem in patients with chronic diseases, possibly leading to hospital readmissions and medical complications, with increased healthcare expenses. Methods: Towards objectively monitoring medication adherence, we propose a method to automatically recognize hand gestures in daily living. The method relies on a commercially available wristband sensor (MMR, MbientLab Inc.) integrating tri-axial accelerometer and gyroscope. Both machine (ML) and deep-learning (DL) algorithms were evaluated for 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 MMR signals. Twenty-two participants were involved in the experimental analysis, performing a 10-minute 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: Our 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.
Suggested Reviewers:	Ahmad Lotfi Professor, Nottingham Trent University School of Science and Technology ahmad.lotfi@ntu.ac.uk Corresponding author of paper [34]: Eating and drinking gesture spotting and recognition using a novel adaptive segmentation technique and a gesture discrepancy measure, Expert Systems with Applications (2020) 140:112888. doi:10.1016/j.eswa.2019.112888
	Ryan McGinnis University of Vermont ryan.mcginnis@uvm.edu Author of review [27]: Estimating biomechanical time-series with wearable sensors: A

	systematic review of machine learning techniques, Sensors (2019) 19(23):5227. doi:10.3390/s19235227
Opposed Reviewers:	

Dipartimento di Elettronica, Informazione e Bioingegneria

Politecnico di Milano

20133 Milano (Italia)
Piazza Leonardo da Vinci, 32

Prof. Filippo Molinari, PhD.
Editor *Computer Methods and Programs in Biomedicine*

Department of Electronics and Telecommunications, Politecnico di Torino,
Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
E-mail: filippo.molinari@polito.it

Milan, 1st May 2021

Dear Prof. Molinari,

We would like to submit the manuscript *"Towards monitoring medical adherence using a wristband and machine learning"* as **Original Paper** to **Computer Methods and Programs in Biomedicine**.

This work focuses on the gesture activity measured using a wrist device and opportunely classified (in both a multiclass or a binary class context) by different Machine learning and Deep learning approaches, compared to each other to establish the best one, in order to provide a potential indirect measurement of medication adherence.

This application is novel and shows potential of using wearable wrist inertial sensors, compared to proper processing, for gesture classification.

We confirm that the manuscript has not been published previously, that it is not under consideration for publication elsewhere, that its publication is approved by all authors, and that, if accepted, it will not be published elsewhere in the same form, without the written consent of the copyright-holder.

We sincerely hope that our work could be considered for review and possible publication on your prestigious Journal.

Yours sincerely,

Prof. Enrico G Caiani
on behalf of all co-authors.

Highlights

- Poor medication adherence threatens patient's outcome, urging for monitoring
- Wristband's inertial data can support automatic monitoring of adherence
- We propose a learning-based pipeline for hand-gesture classification
- Multi-gesture and 'drink' gesture recognition resulted in 90.5% and 92.5% f1-scores
- These results pave the way for new methods of medical adherence monitoring

Towards monitoring medical adherence using a wristband and machine learning

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] *Electronics, Information and Bioengineering Dpt., Politecnico di Milano, Milan, Italy*
[d] *Institute of Electronics, Computer and Telecommunication Engineering (IEIIT), National Research Council of Italy (CNR), Milan, Italy*

Abstract

Background: The demographic shift generated by the ageing of the world's population is having important consequences in the rise of chronic diseases. People with chronic diseases require long-term care and commitment to adhere to the prescribed medications, e.g. taking pills on a daily basis. Poor medication adherence is a common problem in patients with chronic diseases, possibly leading to hospital readmissions and medical complications, with increased healthcare expenses.

Methods: Towards objectively monitoring medication adherence, we propose a method to automatically recognize hand gestures in daily living. The method relies on a commercially available wristband sensor (MMR, MbientLab Inc.) integrating tri-axial accelerometer and gyroscope. Both machine (ML) and deep-learning (DL) algorithms were evaluated for 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 MMR signals. Twenty-two participants were involved in the experimental analysis, performing a 10-minute 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: Our 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.

* Corresponding Author Enrico G Caiani.
Email address: enrico.caiani@polimi.it

Keywords: Hand-gesture classification, monitoring medical adherence, wearable sensors, machine learning, deep learning, CNN-LSTM

1. Introduction

The remarkable improvements in healthcare of the past century have led to an increase in life expectancy. Population aging is associated with the rise of chronic diseases, such as heart diseases, stroke and diabetes, which are the most frequent conditions affecting the elderly [1]. Patients with chronic conditions require long-term care, which includes home nursing, assisted living and long-stay hospitalization. Chronic diseases further imply long-term therapy and several medicine prescriptions, posing issues relevant to poor medication adherence [2]. 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 [3], as well as to increased emergency-room visits, hospitalization and hospital readmissions [4]. Studies have shown that 26% of hospitalizations involving older adults are related to poor or wrong medication adherence [5]. 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 [6].

In this complex scenario, several approaches for medication adherence monitoring have been developed, including both direct and indirect measurements. The former are based on the direct observation of medication intake, including the detection of drugs in biologic fluid (i.e., blood or urine), or the use of medications manufactured with an ingestible sensor embedded in the pill, emitting an electric signal upon digestion. These methods provide accurate estimation of adherence, but they are invasive, costly and time consuming. Conversely, indirect measurements of adherence monitoring include self-reporting, pharmacy refill rates assessment, as well as smartphone reminders applications. Requiring, in most cases, an individual's interaction, those methods lack reliability for long-term monitoring [7].

In order to attenuate such issues, electronic medication packaging devices have been proposed. Those are adherence-monitoring devices incorporated into the packaging of a prescription medication, which can record dosing events, provide audio-visual reminders for the next dose and provide feedbacks on adherence performance [8]. Among them, the medication events monitoring systems (MEMS) are the most commonly used in medication adherence studies. MEMS include pill containers equipped with a microprocessor that registers the opening event as a possible removal of a pill [9,10]. These devices are relatively cheap, easy to use, and safe. However, they can be

easily deceived, as they are assuming the pill intake for every opening, thus without taking into account possible accidental actuation of the container [8]. As an alternative solution, the use of wearable devices has been proposed. In [11], a wrist-worn device equipped with a tri-axis accelerometer was used to detect gestures as drinking water, picking pills, holding pills and taking them to the mouth. The work relies on extracting signal features from an accelerometer and gyroscope, both embedded in a commercial smartwatch, to detect the gestures of twisting the pill bottle cap and turning the palm upward to take the pill, respectively. Similarly, Wang et al. [8] 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.

Given the described potential of wearable devices in addressing the problem of poor medication adherence by recognizing hand gestures, we hypothesised that machine (ML) and deep learning (DL) methods, which are currently employed in human activity recognition [12,13], can be used to process wearable-device signals and provide accurate hand gesture recognition for the purpose of medication adherence monitoring.

Specifically, our aim was to develop and test ML/DL methods for recognizing eight common hand activities from accelerometer and gyroscope signals acquired by a commercially available wristband, with a particular focus on the classification of the drinking gesture, to be part of a novel solution for supporting drug adherence currently under development.

2. Materials and methods

This section describes the utilized wrist device (Sec. 2.1), the data acquisition protocol (Sec. 2.2) and gesture classification pipeline (Sec. 2.3), together with the experimental protocol for evaluation and testing (Sec 2.4).

2.1 MetaMotionR wrist monitoring device

The MetaMotionR wrist wearable device (MMR), developed by MbientLab (MBIENTLAB INC, San Francisco, CA, USA), was used. 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.

The MMR device (Fig. 1) embeds a tri-axial accelerometer, a tri-axial gyroscope, an ambient light sensor, and a humidity sensor. 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°/sec. 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 IOS devices.

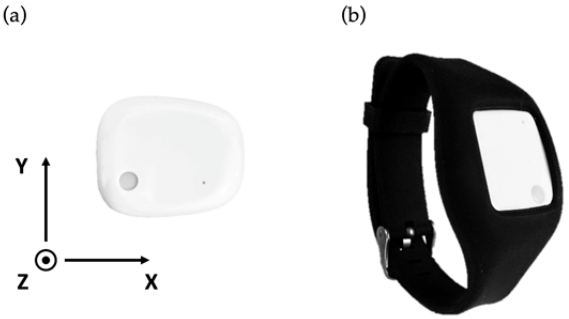

Fig. 1. MetaMotionR device (MMR) (MBIENTLAB INC, San Francisco, CA, USA): (a) Axis orientation of the inertial sensors embedded in the MMR; (b) Sensor integrated in MBIENTLAB provided rubber WatchBand.

2.2 Study population and acquisition protocol

The study was approved by the Ethical Committee of Politecnico di Milano. Twenty-two healthy subjects, both men and women (mean ± SD, 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 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) while performing a set of eight common daily hand gestures (Fig. 2), chosen among the most common studied in the literature [14,15,16]:

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 by the DH and puts the cap on the table.
4. Pour water: the subject takes an opened bottle, pours an amount of water in a glass, and puts the bottle back on the table using only the DH.
5. Typing: the subject types at least ten characters on a computer keyboard with the DH using the index finger.

6. Answering the phone: the subject takes the phone from the table using the DH, raises it up to the ear and holds it for 3-5 seconds, then puts it back on the table.
7. Combing hair: 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.

All the objects were already present on the table.

The subjects were asked to perform the eight gestures in a random order during a 10-minute acquisition (Protocol 1). They were asked to perform one gesture in a 30-second interval (timing was monitored by the supervising researcher), 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. There were no restrictions in the modality of performing the activities. During the acquisition session, the sequence of actions performed by the subjects was annotated by the supervising researcher. Additionally, to increase the dataset size, seventeen subjects out of the enrolled twenty-two accepted to perform also a second 2-minute acquisition (Protocol 2), which included the action of drinking 4 different quantities 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

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 their performance during the tests.

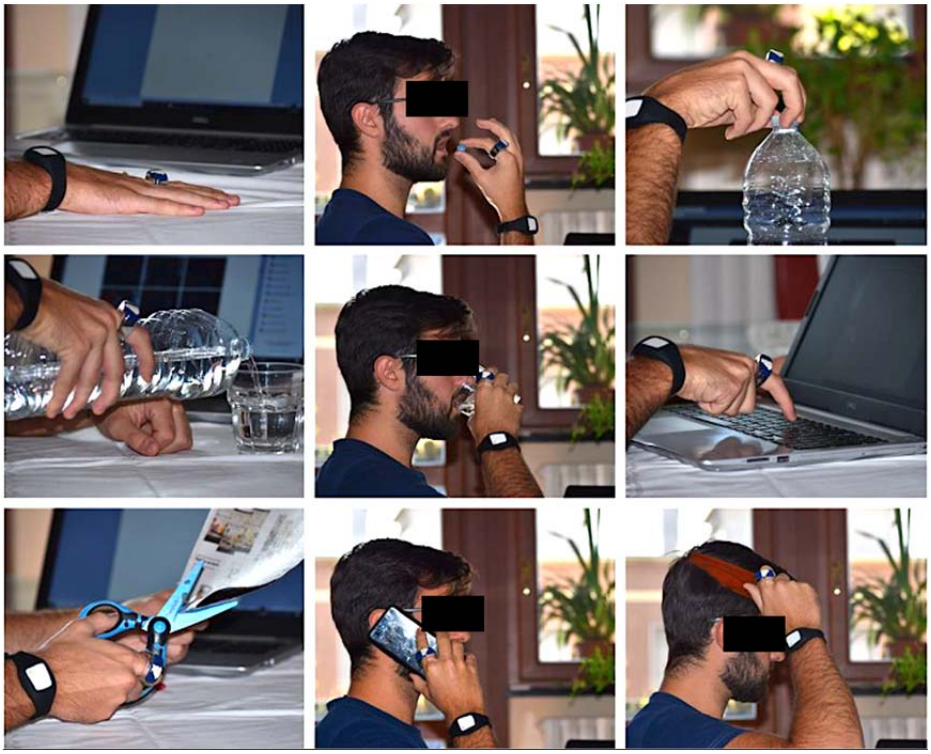

Fig. 2. Hand gestures studied in this work. From left to right/ up to down: Hand idle (performed between gestures), eating, opening a bottle, filling a glass, drinking water, typing, cutting, answering the phone and combing.

2.3 Gestures classification

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

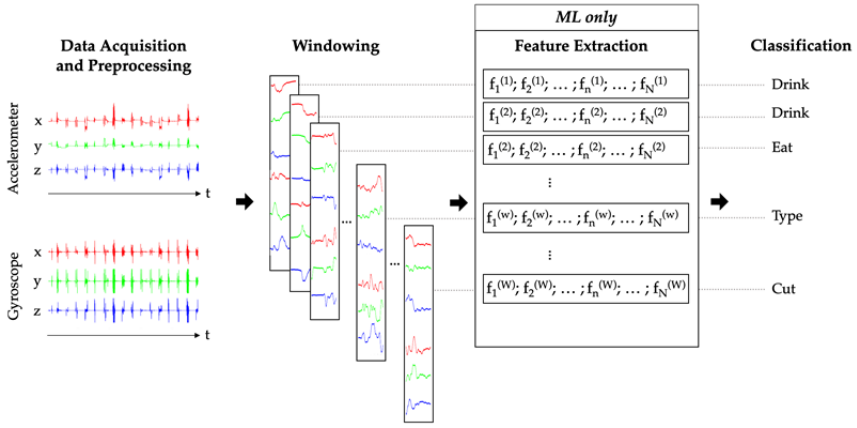

Fig. 3. The proposed gesture classification pipeline includes: data acquisition and pre-processing, signal windowing, feature extraction, which is performed for machine-learning (ML) methods only, and classification. In feature extraction, $f_n(w)$ represents the n-th feature of the feature vector f obtained for the window w.

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 our signals, which was below 5-6 Hz, as verified by Fourier power spectrum analysis. Pre-processing also included raw signal standardization: each signal acquired with the MMR was centred to have zero mean and standard deviation equal to one. After filtering, the portion of the signals relevant to the idle gesture was manually removed.

The raw signal was split into fixed size window segments, referred to as temporal windows, as commonly performed in the literature [17]. Each window included one gesture only. To select the proper window length, three window lengths of 2s, 3s and 6s were evaluated. After preliminary testing, the window size of 3s (corresponding to 150 data points with the sampling frequency of 50Hz) was selected. Window overlap was 0.75s, corresponding to 75% overlap.

The feature extraction step was performed for ML algorithms only, as they require handcrafted features. Here, a combination of time and domain features (Table 1) mostly used in research work for activity recognition [12,13,18] was used. Both for the accelerometer and gyroscope signals, each feature was computed for the three axes 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:

- Classification of all the eight gestures (multi-gesture classification problem)
- Classification of the of the drinking gesture versus all the other gestures (binary classification problem)

The classification was performed testing both ML and DL approaches.

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

2.3.1 Gestures classification with ML

As for 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; the hyperparameters of each classifier were tuned using grid search with 5-fold cross-validation (Table 2), according to the highest f1-score. The Least Absolute Shrinkage and Selection Operator (LASSO) [23] algorithm was used for feature selection. The ML classification process was implemented in Python using the open-source machine learning library Scikit-learn (http://scikit-learn.org/stable/index.html).

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.

2.3.2 *Gestures classification with DL*

For gesture classification with DL, both a convolutional neural network (CNN) and a hybrid model combining CNN and long short-term memory (LSTM) were investigated.

The architecture of the proposed CNNs for both multi-gesture and binary gesture classifications are shown in Table 3: it is based on the one proposed in [24], but the input shape and number of layers have been adapted to our signals. The input shape of both 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 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 used for preventing overfitting. For the multi-gesture and binary classification CNNs, a fully connected layer with 8 (where 8 is the number of gestures) and 2 neurons was used, respectively. In the latter case, a further fully connected layer with dropout was added to attenuate overfitting issues when moving from 500 to 2 neurons.

The CNN-LSTM models (Table 4) were built by adding an LSTM layer on top of the convolutional part of the CNNs in Table 3. In particular, the models take inspiration from [25], adapting the number of layers and the input shape to our signals. 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. The architecture of the CNN in the CNN-LSTM model was kept as previously described, in order to make a fair comparison between the CNN and CNN-LSTM.

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.

2.4 *Validation protocol*

Leave-one-subject-out cross-validation was used to evaluate both the ML and DL approaches. As performance metrics, precision, recall, f1-score and the precision-recall (PR) curve were computed. In addition, the balanced accuracy was calculated.

2.5 Statistical analysis

With the aim to compare methods' performance, the non-parametric Friedman test ($p<0.05$) was applied to compare 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.

Table 3

Architecture of the CNN for gesture classification. Both the multi-gesture and binary classification CNNs 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	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* Layer 8 - *Binary*	Dropout	_	(None, 500)	(None, 500)	1	1
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* Layer 9 - *Binary*	Drop out	_	(None, 500)	(None, 500)	1	1
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.

3. Results

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

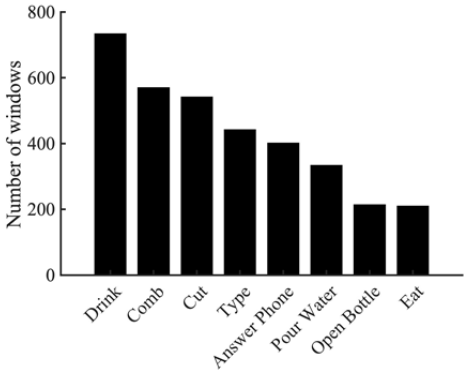

Fig. 4. Data distribution for each gesture in the available dataset.

Fig. 4 shows the final number of windows available for 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 with 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 allowed to obtain significantly higher results 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 compared to KNN (82.5 [75.5; 87.0]% and 71.0 [60.0; 81.0]%, respectively), with SVM and CNN-LSTM resulting with the highest values of balanced accuracy (96.3 [92.6; 97.5]% and 96.3 [91.1; 99.2]%, respectively).

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.

The PR curves obtained with CNN-LSTM are reported in Fig. 5. 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 in the binary classification.

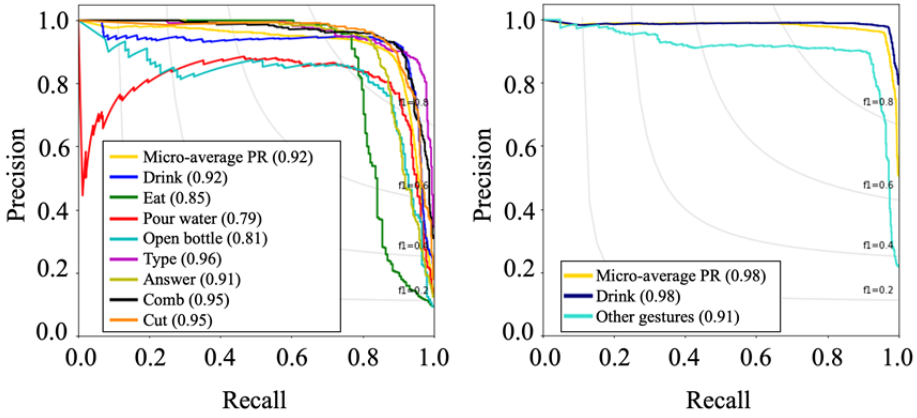

Fig. 5. Data distribution for each gesture in the available dataset.

4. Discussion

In this paper, an innovative learning-based framework for gesture classification from hand activity captured 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 dataset collected for this study included similar movements, thus challenging the proper discrimination of each activity.

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 allowed to achieve 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 to process 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 that is 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 [28,29,30] or heart failure patients [31]. Accordingly, a binary classification problem was addressed 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 [15], where the same number of subjects was studied, our SVM results were superior, both in terms of f1-score and accuracy. Also, our results outperformed [22] in the overall recall, while the precision was slightly lower; however, it is worth noticing that only 2 subjects were studied in [22]. On the contrary, [32] achieved higher values of precision and recall 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 [33] 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 our results with RF, though remaining inferior to the results we obtained with DL models. Instead, Ortega-Anderez and colleagues [34], in the 2-class classification of eating/drinking versus other gestures obtained with the RF model a better performance compared to our study, both considering multi-gesture and binary (drink versus non-drink). This outcome possibly depends on their choice to consider eating and drinking gestures as a single class.

Table 6

Comparison of the results obtained with ML models with the state of the art.

Ref	# subjects	Sensor Type	Sensor placement	Activities	ML models	Balanced Accuracy	Precision	Recall	f1-score
Our 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	Finger Wrist	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%	-
[32]	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%	-

[33]	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
[34]	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 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, our results with CNN and CNN-LSTM were comparable to [34] for the 3-class classification problems, and outperformed their 5-class classification. On the contrary, our performance appears slightly inferior to [26], 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 [25], our work showed higher values of f1-score in both CNN and CNN-LSTM models.

Table 7

325 Comparison of the results obtained with DL models with the state of the art.

Ref	# subjects	Sensor type	Sensor placement	Activities	ML models	Balanced Accuracy	Precision	Recall	f1-score
Our 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, and magnetometer	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,	CNN DBN	CNN 95% DBN 84%	-	-	CNN 89.6% DBN 76%

				forehand, backhand and smash					
[34]	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; CNN = Convolutional Neural Network; LSTM = Long-Short Term Memory; DC = Deep Convolutional; DBN = Deep Belief Network; ANN = Artificial Neural Network.

4.1 Limitations

In the acquisition protocol, only one activity was performed in a 30-second 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 obtaining 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 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 solutions for passive monitoring of medical adherence.

Conflict of interest statement

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

References

[1] J.L. Wolff, B. Starfield, G. Anderson, Prevalence, expenditures, and complications of multiple chronic conditions in the elderly, Archives of Internal Medicine (2002) 162(20):2269-76. doi:10.1001/archinte.162.20.2269

[2] M. Lemstra, C. Nwankwo, Y. Bird, J. Moraros, Primary nonadherence to chronic disease medications: a meta-analysis. Patient preference and adherence (2018) 12: 721–731. doi:10.2147/PPA.S161151

[3] E. Sabaté, Adherence to long-term therapies: evidence for action. World Health Organization (2003).

[4] 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, International Journal of Clinical Pharmacy (2020) 42:1243-51. doi:10.1007/s11096-020-01040-1

[5] T.H. Lim, A.H. Abdullah, Medication Adherence using Non-intrusive Wearable Sensors, EAI Endorsed Transactions on Ambient Systems (2017) 4(16). doi:10.4108/eai.19-12-2017.153484

[6] H. Kalantarian, N. Alshurafa, M. Sarrafzadeh, Detection of gestures associated with medication adherence using smartwatch-based inertial sensors, IEEE Sensors Journal (2016) 16(4):1054-61. doi:10.1109/JSEN.2015.2497279

[7] M. Aldeer, M. Javanmard, R.P. Martin, A review of medication adherence monitoring technologies, Applied System Innovation (2018) 1(2):14. doi:10.3390/asi1020014

[8] 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 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2014) 4168-4171. doi:10.1109/EMBC.2014.6944542

[9] T.L. Hayes, J.M. Hunt, A. Adami, J.A. Kaye, An electronic pillbox for continuous monitoring of medication adherence. In International Conference of the IEEE Engineering in Medicine and Biology Society (2006) 6400-6403. doi:10.1109/IEMBS.2006.260367

[10] M. Aldeer, R.P. Martin, R.E. Howard, PillSense: designing a medication adherence monitoring system using pill bottle-mounted wireless sensors. In IEEE International Conference on Communications Workshops (2018) 1-6.

[11] 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.

[12] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors. IEEE communications surveys & tutorials (2012) 15(3):1192-209. doi:10.1109/SURV.2012.110112.00192

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

[14] 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

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

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

[17] 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 preprint arXiv:1806.05226 (2018).

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

[19] Z. He, L. Jin, Activity recognition from acceleration data based on discrete consine transform and SVM, In IEEE International Conference on Systems, Man and Cybernetics (2009) 5041-5044. doi:10.1109/ICSMC.2009.5346042

[20] K.M. Chathuramali, R. Rodrigo, Faster human activity recognition with SVM, In International Conference on Advances in ICT for Emerging Regions (2012) 197-203. doi:10.1109/ICTer.2012.6421415

[21] A. Stisen, H. Blunck, S. Bhattacharya, T.S. Prentow, M.B Kjærgaard, A. Dey, T. Sonne, M.M. Jensen, Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition, In Proceedings of the 13th ACM conference on embedded networked sensor systems (2015) pp. 127-140. doi:10.1145/2809695.2809718

[22] A. Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors, ACM Computing Surveys (CSUR) (2014) 46.3:33. doi:10.1145/2499621

[23] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) (1996) 58(1):267-88. doi:10.1111/j.2517-6161.1996.tb02080.x

[24] C.A. Ronao, S.B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks, Expert systems with applications (2016) 59:235-44. doi:10.1016/j.eswa.2016.04.032

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

[26] 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.

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

[28] J. Mentes, Oral Hydration in Older Adults: Greater awareness is needed in preventing, recognizing, and treating dehydration, The American Journal of Nursing (2006) 106(6):40-9. doi:10.1097/00000446-200606000-00023

[29] C.M. Sheehy, P.A. Perry, S.L. Cromwell, Dehydration: biological considerations, age-related changes, and risk factors in older adults, Biological research for nursing (1999) 1(1):30-7. doi:10.1177/109980049900100105

[30] C. Lecko, Improving hydration: an issue of safety, Nursing and Residential Care (2008) 10(3):149-50. doi:10.12968/nrec.2008.10.3.28593

[31] P. Pellicori, K. Kaur, A.L. Clark, Fluid Management in Patients with Chronic Heart Failure, Cardiac Failure Review (2015) 1(2):90-95. doi:10.15420/cfr.2015.1.2.90

[32] D. Ortega-Anderez, A. Lotfi, C. Langensiepen, K. Appiah, A multi-level refinement approach towards the classification of quotidian activities using accelerometer data, Journal of Ambient Intelligence and Humanized Computing (2019) 10(11):4319-30. doi:10.1007/s12652-018-1110-y

[33] 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

[34] 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 Systems with Applications (2020) 140:112888. doi:10.1016/j.eswa.2019.112888

CONFLICT OF INTEREST

There are no conflicts of interest that could have inappropriately influenced this research work.