IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2021, accepted September 15, 2021, date of publication September 23, 2021, date of current version October 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3115434

Identification of Free and WHO-Compliant Handwashing Moments Using Low Cost Wrist-Worn Wearables

JUAN M. SANTOS-GAGO, (Member, IEEE), MATEO RAMOS-MERINO, AND LUIS M. ÁLVAREZ-SABUCEDO

atlanTTic, School of Telecommunications Engineering, University of Vigo at Lagoas-Marcosende, 36310 Vigo, Spain

Corresponding author: Juan M. Santos-Gago (jsgago@det.uvigo.es)

This work was supported in part by the Spanish State Research Agency, and in part by the European Regional Development Fund (ERDF) through PALLAS (Plataforma de Servicios basada en Análisis Multimodal para Aprendizaje Autorregulado) Project under Grant TIN2016-80515-RAEI/EFRDEU.

ABSTRACT Hand washing is the simplest and most effective gesture, when correctly performed, for the prevention of many infections. For this reason, the World Health Organization (WHO) has defined a washing procedure that guarantees effective and safe cleaning. This organization recommends that States promote this activity and monitor it continuously. Based on this fact, this article presents a work oriented to study the feasibility of identifying the moments in which a person carried out a hand washing, determining its beginning and duration, as well as if these washings were compliant with the WHO guidelines. The identification of washing moments is made through the analysis, by means of Machine Learning techniques, of the data that can be collected from the inertial sensors of the smartwatch the person is wearing. This study was carried out with the participation of 15 volunteers. Data was not only collected in controlled settings but, also, more than 600 hours of sensor measurements come from free-live conditions. The results of the study showed that it is feasible to build a solid solution based on the use of low cost wearables for the identification of washing moments. The solution is very effective (with F1 over 95%) with user-dependent models. Also, with user-independent models, the identification of WHO washings is also very effective (with F1 above 85%), but more limited in the detection of free washings (F1 around 55%).

INDEX TERMS Data analysis, handwashing recognition, machine learning, smartwatch, wearable sensors.

I. INTRODUCTION

Water, soap, and a minute of time is all we need to wash our hands. Using our hands, we touch any kind of nearby objects and, then, ourselves. In particular, our mouth, nose, and eyes are gateways open to all types of germs. Many pathogenic microorganisms can survive for days on some surfaces. Hand washing is a very simple and inexpensive action that prevents the spread of many infections and saves countless lives.

The prevention of infections is a key element in strengthening national health systems. Hand hygiene is a fundamental pillar in order to prevent infections. For this reason, the World Health Organisation (WHO) devotes a great effort to promote handwashing in healthcare environments and a large amount of resources to articulate strategies aimed at raising awareness and educating citizens about this activity. Among the recommendations of the WHO in the context of epidemics such as COVID-19, even before using masks, the frequent hand washing is paramount.

It should be noted that it is important to wash hands both frequently and effectively. According to the guide published by the WHO [1], defective hand cleansing (e.g. use of an insufficient amount of soap and/or an insufficient duration of hand hygiene action) leads to poor hand decontamination. The above mentioned guide includes an action protocol consisting of a number of steps, as shown in Fig. 1: i) wet hands, ii) apply enough soap, iii) rub hands palm to palm, iv) rub right palm over left dorsum with interlaced fingers and vice versa, v) rub palm to palm with fingers interlaced, vi) rub with backs of fingers to opposing palms with fingers interlocked, vii) rub each thumb clasped in opposite hand using rotational movement, viii) rub tips of fingers in opposite palm in circular

The associate editor coordinating the review of this manuscript and approving it for publication was Salvatore Surdo.

FIGURE 1. World Health Organization (WHO) handwashing protocol.

motion, ix) rinse hands with water and x) dry thoroughly, xi) use towel to turn off faucet, and xii) your hands are now safe.

This guide from the WHO states that a strategy for promoting handwashing should include (primarily in the health context, but also in any other context, especially where there is a clear risk of disease transmission) the training/education on the importance of hand hygiene, the monitoring of hand cleansing practices, and the provision of remainders. Aware of this need, and in the framework of a research project focused on the definition of self-regulated learning services based on the use of wearable wrist devices, the authors decided to address a telematics solution that supports the automated monitoring of hand washing and the provision of personal recommendations related to hand hygiene. Such a solution could be a relevant contribution in various domains where hand hygiene is essential, ranging from the domain of logistics and industry to the public health sector. In particular, the solution could significantly contribute to push forward the United-Nations Sustainable Development Goal 3: ''Ensuring a healthy life and promoting well-being for all ages''.

This telematics solution is based on the analysis of data collectable from commonly used commercial wearable wrist devices. These devices, such as smartwatches or smartbands manufactured by companies such as Fitbit, Polar, Garmin, Apple, Samsung, Mobvoi, Fossil or Xiaomi, are becoming increasingly popular. Their popularity is mainly due to their usefulness as fitness trackers and sports monitoring tools [2], [3]. Nevertheless, their potential goes far beyond that, as shown by various proposals in different fields, either in the context of working environments [4], [5], in the domain of education [6], [7] or, above all, in the field of health [8]. Wrist wearables are portable devices, easy to use, practically transparent to the user, and are provided with an important set (depending on the device) of sensors, such as [9]: accelerometer, gyroscope, magnetometer, heart rate monitor, pedometer, barometer, altimeter, thermometer, light meter, oximeter, GSR, and even, sometimes, blood pressure or ECG meter. A recent study [10] showed that the vast majority of commercial wearable wrist devices have at least one inertial sensor, which provides accelerometry and gyroscope data. This data is, a priori, potentially enough to, through analysis, detect when a person, carrying a smartwatch or smartband, washes his/her hands.

This paper describes a work aimed to *study the possibility of identifying moments in which a person performs hand washing (and if these washings comply with the protocol defined by the WHO) by means of Machine Learning techniques, using the continuous flow of inertial measurement data collected from a common use commercial smartwatch*. It is convenient to underline the importance of discerning whether a wash is WHO-compliant, as the latter, to a certain extent, guarantees that the wash performed is effective and it adequately sanitizes the entire surface of the hands.

The following section (Section II) includes a brief description of the state of the art of automatic hand washing identification. Section III details the methodology used in the study conducted. Later, section IV shows the results obtained after the application of such methodology. Section V discusses these results and, finally, Section VI presents the conclusions and the new challenges that are currently being explored as a continuation of the work done.

II. STATE OF THE ART

Aware of the importance of the hand cleaning habit among the population, many companies and developers have made available apps for smartwatches that, in some way, allow to monitor and manage the daily hand washing. This is the case in the two main ecosystems of wearable wrist devices, i.e., Google and Apple, but also for other manufactures such as Samsung, Huawei, or Garmin.

In this line, as a similar approach to the proposal made in this work, it is worth highlighting the solution offered by Apple. This company has incorporated in the latest version of its OS for its smartwatch, WatchOS 7, functionalities related to the detection of hand washing in a generic manner [11]. The most noticeable functionality is the detection of the start of hand washing, which is spotted using the data from the inertial measurement sensor and the audio fetched by the microphone available in the Apple Watch. This microphone detects the sound produced by the water when a tap is opened. The user can configure the device so that, once a wash start is detected, a timer starts, and, when 20 seconds have elapsed, an alarm is triggered indicating that the wash process has been carried out for the appropriate minimum time.

This type of functionality also can be spotted, in a more rudimentary way, in apps from other ecosystems. In this sense, we can refer to some types of applications that are common in other environments such as Android. In this ecosystem, some apps seek to offer timers to determine the time of washing. Applications such as ''Hands Washing Timer'' [12] would fall into this category. Also, we can find applications oriented to a more childish target audience. It should be borne in mind that this target, young children, should be the main objective for training and education in

these good habits. Within this segment, apps like ''Ella's hand Washing Adventure'' [13] should be mentioned. Within this line, but considering a more mature public, we can find apps like ''SureWash Hand Hygiene'' [14]. This one aims to train users in the procedure recommended for hand washing by the WHO. However, no applications that automatically detect or monitor hand washing could be identified.

Other manufacturers, such as Samsung, have also dedicated efforts to this line of work [15]. However, it can be noted that, broadly speaking, the possibilities of using smartwatches for the detection, cataloguing, and evaluation of hand washing are not fully explored and its quality and performance is not yet sufficiently mature. At least, the authors understand that in a context like the current one, more open and sensitive solutions to the types and qualities of hand washing should be available to all users of these devices.

In the academic literature, a number of studies aimed at detecting handwashing activities based on the analysis of data collected from different types of wearable devices can be found. In that sense, a search was made for papers published in conferences and high-impact journals using specialized search engines (Scopus, Web of Science and Google Scholar). It should be noted that most of the works found could be included in the area of sensor-based human daily activity recognition [16], aimed at obtaining classifiers that, given the sensory data collected from an electronic device corresponding to a given daily activity (e.g. walking, eating or running), can discern the type of activity from a prefixed set of activity types.

As an example of this type of work, in [17] the authors study the potential use of wearable devices in one and both wrists to detect, using different classification techniques from the Machine Learning domain, hand washing from a set of activity types, including ''walking'', ''opening a jar containing candy'', ''opening and eating the candy'', ''tying shoes'', and ''applying bandages''. The authors achieve accuracies of approximately 90% and even higher in some of the experiments.

In [18], accelerometry, gyroscopy and audio data collected from a LG G Watch W100 smartwatch is used to identify hand washing and tooth brushing activities. This is done using a Naïve Bayes classifier trained on a dataset with a predefined set of activity types (not specified in the publication). Accuracy rate above 95% was obtained in detecting the considered tasks. Similarly to this work, in [19], a mechanism aimed to detect several early morning activities using audio data and accelerometry data is proposed. The activities considered in this case are ''teeth rushing'', ''hand washing'', ''shaving'', ''electric brushing'' and ''electric shaving''.

Among the concerns with the above-mentioned works, it must be pointed out that the validation of proposals is done using ''laboratory'' datasets, which are too synthetic and limited. This results in very good performance, but not necessarily generalizable to real contexts. To partially solve this issue, in [20] the authors propose the use of an Artificial Neural Network (ANN) of 3 hidden layers to classify data records captured from inertial sensors into HAND-WASHING activities or NULL activities (i.e., any other type of activity different from handwashing). They train their neural network using their own dataset, which includes accelerometry data (no data from gyroscope is used) from both, hand washings and other activities. Then, they test the effectiveness and robustness of their proposal using the WISDM dataset [21]. This dataset contains records of 18 types of activities (e.g. ''walking'', ''jogging'', ''sitting'', ''typing'', ''folding clothes'', ''eating pasta'', ''eating soup'', ''drinking from a cup'', etc.), some of which coincide with those considered in the dataset developed by the authors, but do not include hand washing activities. The own dataset was obtained from 16 participants who performed normative and non-normative handwashing and other activities such as ''wiping water from hands'', ''walking'', ''opening/closing doors'', ''using computers/phones'', ''eating'', and ''drinking''. It contains a total of 5 hours of data, for both the right and left hands. The neural network obtains an F1-score close to 80% in the activity classification. It should be noted that there is no difference between normative and non-normative hand washing and that the proposal of these authors does not allow to really estimate the moment or the duration of a hand washing, since it is specifically focused on the process of classification of series of data records that correspond to a concrete activity. Besides, the dataset used for training remains relatively artificial, since the hand washing and other data used was collected separately, in particular, the hand washing was not collected in free-living conditions.

Other works found in the literature present a different approach. Instead of trying to identify washes, they are oriented to detect the different steps of a WHO hand washing. [22] describes a platform that integrates an ad-hoc developed wearable with an inertial measurement unit that enables automated assessments of handwashing routines using a Hidden Markov Model-based analysis method. This platform is able to recognize the 12 steps of the WHO hand-washing procedure. The average accuracy is 92% with user-dependent models, and 85% for user-independent modeling.

Along this research line, other proposals address the problem of identifying the steps of washing using less common wearable devices. For example, [23] or [24] test the use of Thalmic Labs Myo gesture control armband devices. This device includes an inertial measurement unit and eight surface electromyography (EMG) sensors, which allow it to capture the muscular activity of the forearm. The first of the above-mentioned works uses only the data from the inertial sensor of the Myo device placed on the right forearm to detect, by means of a Support Vector Machines (SVM) classifier, hand washes in which any of the steps defined in the WHO protocol are not performed. The second of the above-mentioned works has a similar orientation, although in this case two Myo devices are used, one in each arm. K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) classifiers, as well as Hidden Markov Models (HMM) are explored to identify the steps and their sequencing in an

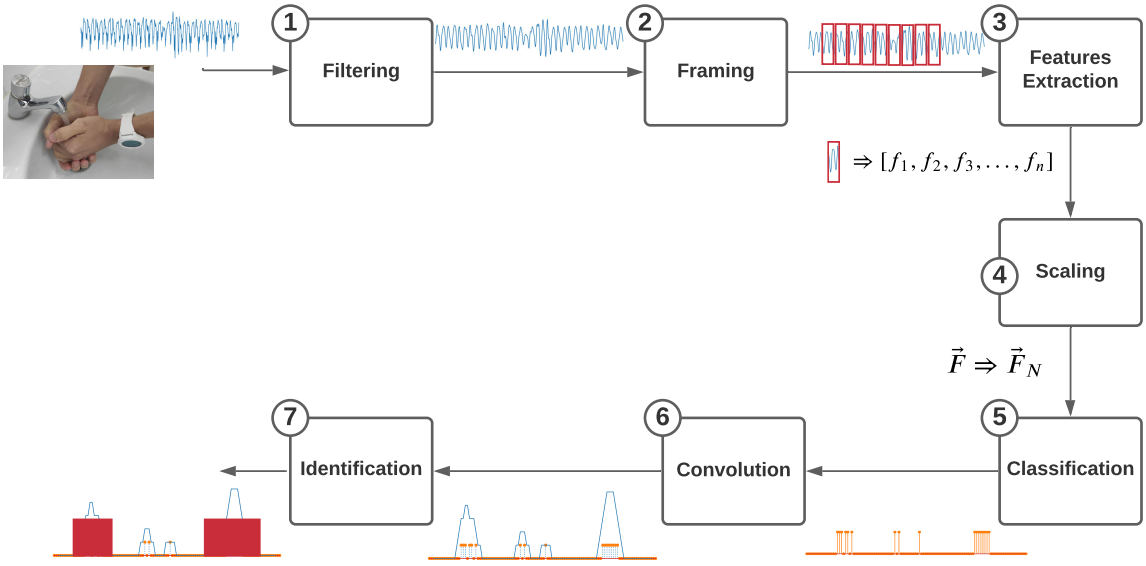

FIGURE 2. Data processing pipeline of the handwashing moments identification.

WHO hand washing. They achieve recognition rates of above 95% for individual gestures.

Another work exploring the use of alternative devices is [25]. Here, the authors designed a smart ring integrating an electrochemical fluid sensor that allows to detect hand washing events. It also allows the detection of a variety of hand washing agents, such as water, soaps, sanitizers, and antimicrobial agents. The main drawback with this type of proposal is linked to the problems of applicability in realistic contexts, since this type of device is uncommon, especially considering the type of sensor used.

Different approaches to detecting and monitoring hand washing have been found in the review. However, none of them meet the needs and premises set out in the previous section. Thus, we did not find solutions that allow detecting moments in which hand washing occurs, identifying when and for how long it lasts, and distinguishing between free washing and washing according to WHO indications. From the point of view of the authors, these solutions should be implementable using common, popular, and user-transparent devices. Solutions that are invasive of users' privacy should also be avoided, i.e., it seems unreasonable, from an ethical perspective, to use microphones to get the data required for the proposed detection and classification task. It should be possible to carry out this task using any existing commercial smartwatch without using any more sensors than strictly necessary. And finally, from the point of view of the authors, the validation of the proposals should be based on data collected in free-living conditions.

III. METHODOLOGY

In the present proposal, the detection of hand washing is based on the smart analysis of the data collected from the wrist-worn wearable carried by the user. In particular, it was decided to use only the data from the inertial sensor within the wearable, i.e., measures provided by an accelerometer and a gyroscope. The use of data from other types of sensors, such as hygrometers or microphones was not considered because, although they might facilitate the construction of a more accurate detector, the existence of this type of sensors in popular wearables is limited and its usage may pose a threat to the privacy of the users. Therefore, the final technological solution can be implemented in a wide variety of low-cost and popular smartwatches.

To carry out the identification of moments in which a person has performed a hand washing, and to check if this has been according to the protocol defined by the WHO, the 3-axis accelerometer and 3-axis gyroscope data collected from the smartwatch (6 signals in total) is processed according to the 7 stages shown in Fig. 2. Basically, in the data processing pipeline, a pre-filtering and framing of the signals (stages 1 and 2 of the pipeline) is carried out to extract a set of normalized features (stages 3 and 4) that can be used by a classifier (stage 5) to determine whether small time frames correspond to fragments of wash moments, either free (i.e., not adhering to any standard) or WHO compliant. Based on the output of the classifier, time periods (stages 6 and 7 of the pipeline) of between 10 and 100 seconds are identified that actually can correspond to a hand washing activity. The 7 specific stages are the following:

1) **Filtering**. In order to remove noise from the 6 inertial measurement signals collected from the smartwatch, a pre-filtering stage is implemented. Since we cannot determine a priori which is the most suitable type of filtering for the problem at hand, we have decided to try different 5-order digital Butterworth Low Pass (LP) filters with different critical frequencies.

2) **Framing**. To extract the features to be used in the classification algorithm, the data corresponding to small time frames of the 6 initial inertial measurement signals are considered. A priori, it is not possible to determine the ideal frame size, so it has been tested with frames between 1 and 8 seconds (step 1 sec). The classification algorithm must estimate if a certain frame corresponds to a fragment of a hand washing process.
3) **Features Extraction**. From the data of the signals that fall within a time frame, the feature extraction is carried out, both in the time and frequency domains (c.f. Table 1). From the time domain a total of 39 commonly-used features are calculated including mean, magnitude of mean, variance, correlation, covariance, interquartile range and zero-crossings rate. From the frequency domain a total of 21 commonly-used features are calculated including spectral energy, magnitude of spectral energy, FFT peak value and FFT peak frequency.

TABLE 1. Features extracted from frames in the time and the frequency domains.

Domain	Feature	#
Time	Mean	6
	Magnitude of mean	3
	Variance	6
	Correlation	6
	Covariance	6
	Interquartile range	6
	Zero-crossings rate	6
Frequency	Spectral energy	6
	Magnitude of spectral energy	3
	FFT peak value	6
	FFT peak frequency	6

4) **Scaling**. This process allows homogenizing the range of the feature values, which facilitates the process of optimizing the subsequent classification algorithm and reduces overfitting. The following scaling methods have been tested:
   - *Min-Max Scaling*: values are shifted and rescaled so that they end up in a range between -1 and 1.
   - *Robust Scaling*: subtracts the median and then dividing by the interquartile range (75% value $-$25% value). It is more resilient to the effect of outliers than the previous approach.
   - *Standardization* (or *Standard Scaling*): subtracts the mean and then dividing by the standard deviation, resulting in a zero-mean and unit-variance dataset.
   - *L1-norm*: each feature is regularized by applying the l1 (Manhattan) normalization.
   - *L2-norm*: each feature is regularized by applying the l2 (Euclidean) normalization.
   - *No scaling*. No data scaling is performed.
5) **Classification**. A classification algorithm based on Machine Learning techniques estimates whether a time frame, according to the features extracted from it, corresponds to a fragment from a normative hand washing, from a free hand washing or it does not correspond to a hand washing. It has been tested with several configurations of different classification algorithms in the field of Machine Learning, in particular, variants of Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Artificial Neural Network (ANN).
6) **Convolution**. A discrete one-dimensional signal with values 0-1 (not washed and washed, correspondingly) is composed from the outcome of the previously described classifier. The signal composed this way is convolved with a square signal. The result is a new signal with high values in the time periods in which frames corresponding to washing are concentrated.
7) **Identification of washing moments**. Using the output signal from stage 6, areas where a certain threshold is exceeded are labeled as potential periods of hand washing. Areas lasting less than 10 seconds and areas lasting more than 100 seconds are discarded. The discard of these areas is based on the observation of the data obtained (cf. section IV-B): the probability that an actual washing lasts less than 10 seconds is very small. The same applies to those that last more than 100 seconds. Also, it seems reasonable to consider that a washing lasting less than 10 seconds is an incorrect washing and, therefore, should not be considered as such.

In order to determine the most effective types of pre-filtering and scaling, identify the most suitable frame size and, above all, train the classifiers, a strategy was designed to capture properly labeled training data. This strategy involves inviting volunteers who must wear a wrist wearable device for a period of time that continuously collects inertial measurement data. An application developed ad-hoc to mark the beginning and end of a specific activity, mainly a hand washing, was used.

The software developments and the first tests were carried out with a Polar M600 device, a smartwatch with Wear OS operating system, a version of Google's Android operating system for wearables. However, for capturing data from experiments on the volunteers, TicWatch S2 devices were chosen. The TicWatch S2 is a smartwatch with similar characteristics to the first one, but at a lower cost. As mentioned, an application was developed and installed on the smartwatch to collect the mentioned data. This application collects the data from the sensors on the device and sends it to an analysis server either directly via Wifi or using a cell phone as an intermediary agent. In this case, the smartwatch sends the data via Bluetooth and the cell phone is responsible for sending the data to the server. This app was developed as the client of an environment created by the authors in a previous project, oriented to capture data to estimate different

parameters related to the user's sleep [26] or stress [27], [28]. Through the functionalities offered by this infrastructure, it is possible to configure the sensors that are activated and used in the data collection (in our case 3 accelerometer signals and 3 gyro signals, as indicated above) and the sampling frequency of each one of them. In our case we chose to use the `SENSOR_DELAY_GAME` range offered by the Wear OS API, which corresponds to a sample rate of approximately 50 Hz in the smartwatch used.

Fig. 3 shows an image of the app used by a volunteer during a test. The volunteer keeps this app running and, when he/she is going to perform a hand washing, he/she must press the green button for 2 seconds, if it is a free wash, or the red button in case of a hand washing according to the WHO protocol. When the washing is finished, it must be indicated by pressing the corresponding button. The inertial measurement data is continuously stored in the device, as well as the start, end and washing type marks. The app includes a button to compress and send the gathered data to the analysis server.

FIGURE 3. Data capture app running on TicWatch S2.

The strategy of training data capture, similar to the one used in [29], is articulated around 3 different test scenarios:

1) **Laboratory Experiments**. This scenario corresponds to a collection of data generated by carrying out a series of predefined activities. Specifically, in this scenario, the volunteer subject is required to perform the following 9 activities: 1) handwriting for 2 minutes, 2) free hand washing for 1 minute, 3) WHO compliant hand washing for 1 minute, 4) playing with a Rubik's cube for 2 minutes, 5) second free hand washing for 1 minute, 6) second WHO hand washing for 1 minute, 7) cleaning a glass and plate for 2 minutes, 8) walking around for 1 minute, and 9) typing on a keyboard for 1 minute. This set of activities, and these specific durations, have been chosen to provide a balanced set of data that includes handwashing, activities that are gesture-analogous to washing, and activities that should not be confused with washing at all.
2) **7-days Experiments**. In this scenario, the volunteer subject must carry the smartwatch for one week and collect about 50 hours of data from the sensors. The subject must keep his daily routine and use the app as described above to indicate the beginning and the end of the hand washes he/she performs, as well as their type, i.e., whether it is a free wash or a WHO wash. The subject is expected to record about 30 hand washes of each type.
3) **30-days Experiments**. This is a similar scenario to the previous one, although in this case the subject uses the wrist wearable for one month and is expected to collect data for approximately 200 hours, as well as record at least 100 washes of each type.

With this strategy, the data collected covers two main contexts. On the one hand, controlled environments are considered where very specific activities, several of which involve performing movements similar to hand washing, are included. And, also, real free-live condition environments are involved as they are the best possible scenarios for evaluating the actual performance of the washing detection algorithm.

When establishing the parameters and hyperparameters involved in the data processing pipeline that achieve the best performance, the goal is to maximize the values of the F-measure at the output of the classifier (stage 5) and/or at the output of the moment identifier (stage 7). The best value of F1 (harmonic mean of precision and recall) will be sought, and, in particular, the best value of the macro F1 [30], [31], so that the same importance is given to each of the classes considered (NO-washing/FREE-washing/WHO-washing). Macro F1 will be low for models that only perform well on the common classes while performing poorly on the rare classes.

In order to evaluate the performance of the hand washing moment identification algorithm more thoroughly, beyond the detection of the existence of a wash moment, a new metric, called *overlapping*, is proposed. This metric allows evaluating the success in predicting the start and duration of a hand washing moment. For a correctly predicted hand washing, the identification algorithm offers an approximation of its start and end moments, called in Fig. 4 as ti_s and ti_e respectively, which constitutes the *ti* interval. The correspondent real hand washing may have another different pair of start and end time instants, denominated tr_s and tr_e, conforming the *tr* interval.

Under these conditions, firstly, the absolute overlapping (*ao*) is defined using the equation 1. Then, since the length of a hand washing is variable, the relative overlapping is calculated taking into account the length of the time intervals *ti* and *tr* (equations 2 and 3). Finally, a conservative strategy is applied and the lower of the relative overlaps is taken as the overlapping value (equation 4). It can be noted that the value of overlapping is 1.0 when the identification algorithm

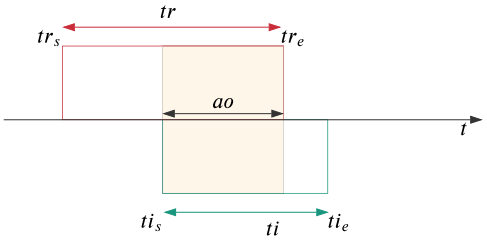

FIGURE 4. Overlapping of a real and predicted handwashing moment.

returns a perfect result (i.e., $tr_s = ti_s$ and $tr_e = ti_e$).

$$ao = min(tr_e, ti_e) - max(tr_s, ti_s) \quad (1)$$

$$ro_i = \frac{ao}{ti} \quad (2)$$

$$ro_r = \frac{ao}{tr} \quad (3)$$

$$overlapping = min(ro_i, ro_r) \quad (4)$$

IV. RESULTS

Fifteen volunteers (40% women) took part in the capture of the training data. The subjects, all of them of legal age, were informed about the tests to be performed and signed the corresponding consent form.

The data collected in each experiment was used both, to train the classifier (stage 5 in the data processing pipeline represented in Fig. 2), and to obtain the hyperparameters that give the best result either at the classifier output, or at the output of the wash moment identifier, depending on the experiment considered. The hyperparameters concern to: i) the configuration parameters of each classification algorithm considered, ii) the different alternatives considered in stages for the signal conditioning and framing (stages 1 to 4 of the data processing pipeline), and iii) the possible options for the washing moment identification stage (stages 6 and 7). Table 2 presents the specific hyperparameters considered for each of the stages.

The data processing consisting of the stages represented in Fig. 2 was implemented through 4 parameterizable pipelines, one for each type of classifier considered. For training the classifier and obtaining optimal hyperparameters, 5-fold cross validation was used in all cases. In this way, overfitting problems are greatly reduced and the generalizability of the results obtained is increased as much as possible.

The following sections describe the particularities and analysis that were carried out in each of the 3 test scenarios considered.

A. LAB EXPERIMENTS

All 15 volunteer subjects participated in the tests performed in the laboratory. As described above, the subjects performed 9 activities, including 2 WHO hand washes and 2 free hand washes. From the inertial measurement signals obtained from each subject, time windows of between 40 and 120 seconds were maintained, corresponding to the periods in which each of the 9 activities was carried out. From these signal windows,

TABLE 2. Hyperparameters.

Stage	Hyperparameter alternatives
Filtering	Butterworth LP (f_c = [5,10,20] Hz) None
Scaling	Min-Max scaling Robust Scaling Standardization L1-norm L2-norm None
Framing	frameSize = {1, 2, 3, 4, 6}
Classification (SVM)	'C': [0.1, 1, 10] 'gamma': ['scale', 'auto'] 'kernel': ['rbf', 'poly', 'sigmoid']
Classification (RF)	'n_estimators': [30, 100, 200] 'max_features': ['auto',None] 'bootstrap': [False, True (with max_samples = [0.33, 0.66])]
Classification (LR)	'C': [0.1, 1, 10], 'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
Classification (ANN)	'hidden_layer_sizes': [(25,25), (25,25,25), (50,50), (50,50,50)] 'activation': ['relu', 'tanh', 'logistic']
Convolution	windowSize = frameSize * [2,3,4,5,6,7,8]

the initial 3 seconds and the final 3 seconds were cut to avoid potential noise at the beginning and end of the activity (Fig. 5 shows the signal windows for one volunteer). The remaining portions of the signals were used to train the classifier and to obtain the optimal hyperparameters. In particular, 2 groups of analysis were carried out, which are described in the following subsections.

1) ALL VERSUS ALL

In this analysis, all the signal windows from each of the participants are used to train the classifiers and to obtain the hyperparameters that maximize the final performance.

Table 3 shows the performance scores achieved in the frame classification process (stage 5 in the full data pipeline) for each of the classifier types considered and the hyperparameters that allow that value to be reached. In this case, they are considered binary classifiers, i.e., they label a frame as 1, when it corresponds to a fragment of a hand washing activity (regardless of whether it is a WHO or a FREE wash), or 0, otherwise.

As the reader may note, using an RF classifier, an F1 of 92.1% is obtained (with an average precision and recall also of 92%). This score is obtained by using a low-pass filter with a cut-off frequency of 5 Hz, a Min-Max scaling, and considering a frame size of 6 seconds. The internal configuration hyperparameters that maximize this classifier are "bootstrap" = false, "max_features" = "auto" and "n_estimators" = 100. With the SVM classifier, almost

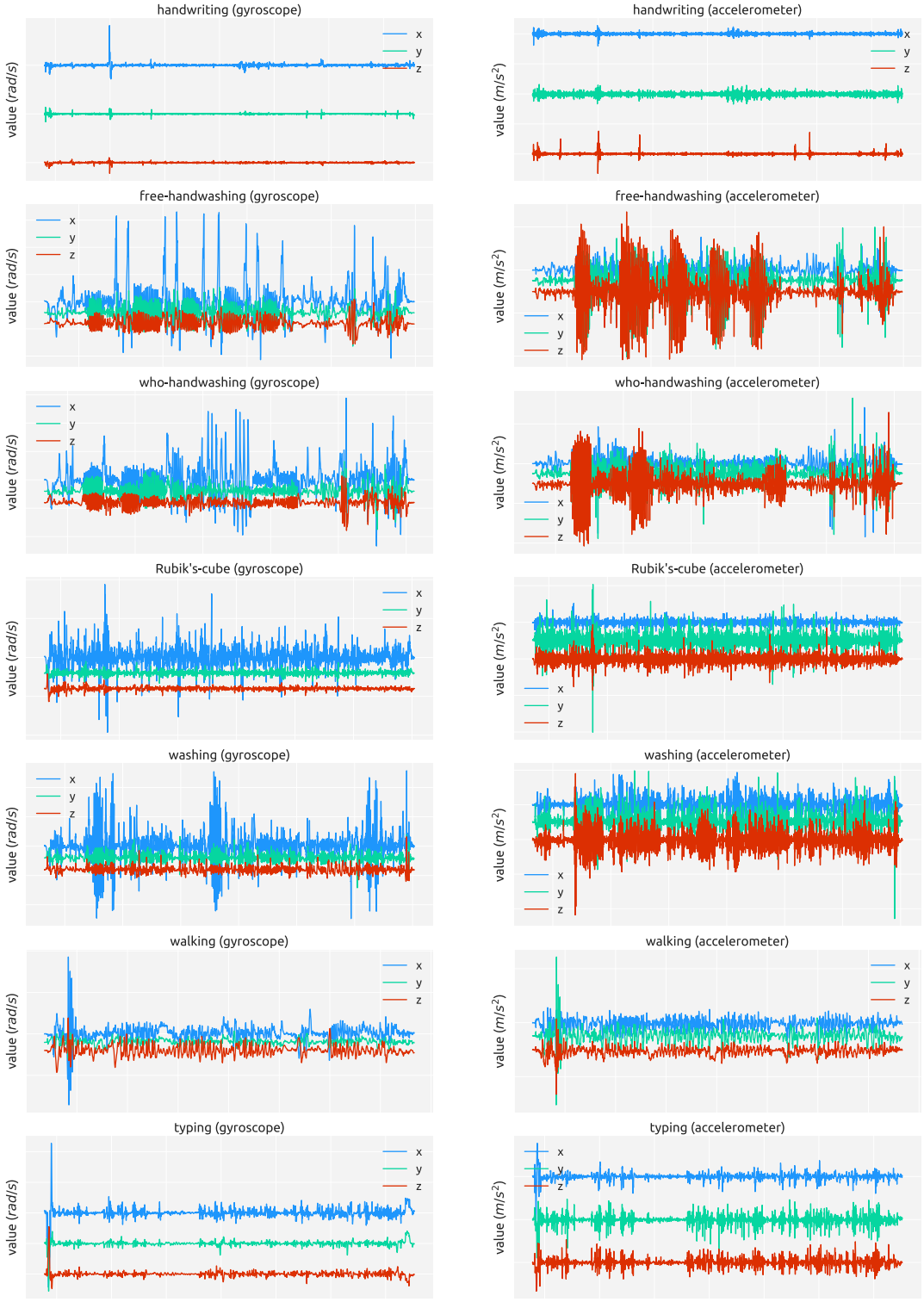

FIGURE 5. **Example of signal windows of the activities performed in the laboratory experiments.**

identical scores are obtained (in this case using a low-pass filter with a 10 Hz cutoff frequency). The performance obtained with an ANN classifier (2 layers with 50 neurons per layer) was also virtually the same as in the previous two ones (with an F1 of 91.5%) and using an LR classifier, the F1 is slightly reduced to 88.7%.

On Table 3 it can be seen that, except for the SVM classifier, the filtering providing the best results is a 5 Hz

TABLE 3. Best scores for the "all vs all" analysis, considering 2 frame classes (washing/no-washing), in the Lab Experiments scenario.

	LR	RF	SVM	ANN
global scores	F1: 0.887 (0.037) P: 0.898 (0.016) R: 0.889 (0.042)	F1: 0.921 (0.034) P: 0.923 (0.027) R: 0.922 (0.037)	F1: 0.920 (0.036) P: 0.928 (0.025) R: 0.921 (0.039)	F1: 0.915 (0.023) P: 0.917 (0.018) R: 0.916 (0.022)
filtering	lowpass 5Hz	lowpass 5Hz	lowpass 10Hz	lowpass 5Hz
scaling	minmax	minmax	minmax	minmax
framing	6 secs	6 secs	6 secs	6 secs
classifier hyperparameters	"C": 1 "solver": "liblinear"	"bootstrap": false "max_features": "auto" "n_estimators": 100	"C": 10.0 "gamma": "scale" "kernel": "rbf"	"activation": "relu", "hidden_layer_sizes": [50, 50] "solver": "adam"

low-pass filter. However, a detailed analysis of this hyperparameter, both, in these laboratory experiments and in subsequent experiments in free-living conditions, showed that the performance of the classifiers is practically independent of the type of filtering applied. As an example, Fig. 6 represents the best value obtained of F1 in each type of classifier for the 5 low pass filter types considered ($f_c = 5, f_c = 10, f_c = 15$, $f_c = 20$ and none). This figure shows that, for a given classifier, the best performances are practically identical regardless of the filter used. Therefore, except where otherwise stated, all the analysis below correspond to a null filtering. The latter was chosen to be the default filtering, as it reduces the computational cost of data processing. The sampling of the signal at 50 Hz itself seems to be a sufficiently selective filtering for the problem being addressed.

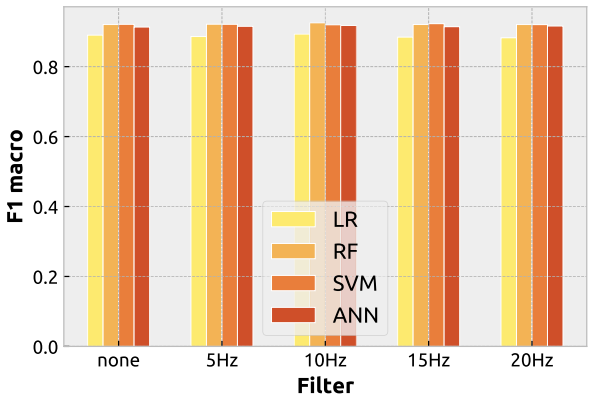

FIGURE 6. F1 vs LP filter.

Regarding the scaling of features before entering the classifier (stage 4 in the full data pipeline), Min-Max provided the best results in all cases. Therefore, it was decided that this would be the default scaling in subsequent analysis since, in addition to get the best performance, it is the easiest one to implement in a final system, as it does not involve the calculation of averages or other statistical values.

Regarding the frame size, the best performance is obtained with a size of 6 seconds in all cases. Although this has been the maximum size considered in this analysis, further observations (as seen in subsection IV-B2) have shown that, indeed, 6 seconds is, overall, the size that allows better performance, although there is some variability depending on the analysis considered.

When distinguishing between WHO and free washing, the best F1 is achieved using the SVM algorithm, as shown in Table 4. In this case, the F1 is 77.8% (with a precision of 79.4% and a recall of 77.9%). That is very close to the value obtained with the RF algorithm and also close to the ANN result. The best performance is obtained with almost the same hyperparameters as in the binary case (washing/non-washing). It is worth mentioning, as a curiosity, that the best scaling in LR is achieved with a standardization, although the differences with a Min-Max scaling are negligible.

Regarding the identification of washing moments (stage 7, final, in the full data pipeline), when just considering washing/non-washing, an F1 of 98% is obtained (precision: 100% and recall: 96%) and an overlapping of 0.93 (SD: 0.10). This value is obtained for the RF algorithm and with a convolution signal size of 18 seconds. When distinguishing between free and WHO washing, the F1 obtained is reduced to 93% (precision: 100% and recall 86%) with an overlapping of 0.70 (SD: 0.17) for the FREE wash. In the case of a WHO wash, it is achieved an F1 value of 98% (precision: 100% and recall: 96%) with an overlapping of 0.87 (SD: 0.09), using a convolution signal of 18 seconds.

2) ALMOST ALL VERSUS ONE

On the second analysis carried out on the grounds of the laboratory experiments, the data corresponding to 14 of the 15 volunteers is used to obtain the hyperparameters of a complete pipeline. The performance of the pipeline obtained is evaluated using the samples of the remaining subject. This is repeated 15 times, so that in each iteration the evaluated user is a different one. It should be noted that the obtained pipeline is tested using samples that, on one hand, are "fresh"

TABLE 4. Best scores for the "all vs all" analysis, considering 3 frame classes (NO-washing/FREE-washing/WHO-washing), in the Lab Experiments scenario.

	LR	RF	SVM	ANN
global scores	F1: 0.722 (0.058) P: 0.730 (0.035) R: 0.730 (0.070)	F1: 0.765 (0.038) P: 0.780 (0.028) R: 0.766 (0.045)	F1: 0.778 (0.054) P: 0.794 (0.029) R: 0.779 (0.064)	F1: 0.749 (0.043) P: 0.755 (0.040) R: 0.751 (0.046)
filtering	lowpass 5Hz	lowpass 15Hz	lowpass 5Hz	lowpass 5Hz
scaling	standarization	minmax	minmax	minmax
framing	6 secs	6 secs	6 secs	6 secs
classifier hyperparameters	"C": 0.1 "solver": "newton-cg"	"bootstrap": true "max_features": "auto" "max_samples": 0.6 "n_estimators": 100	"C": 10.0 "gamma": "scale" "kernel": "rbf"	"activation": "relu" "hidden_layer_sizes": [50, 50] "solver": "adam"

TABLE 5. Best scores for the "Almost all vs one" analysis, considering 2 frame classes (washing/no-washing), in the Lab Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.897 (0.071) P: 0.908 (0.053) R: 0.900 (0.070)	F1: 0.914 (0.0564) P: 0.921 (0.048) R: 0.915 (0.057)	F1: 0.917 (0.057) P: 0.929 (0.041) R: 0.918 (0.055)	F1: 0.906 (0.057) P: 0.916 (0.044) R: 0.907 (0.056)
training scores	F1: 0.891 (0.032) P: 0.893 (0.030) R: 0.893 (0.034)	F1: 0.911 (0.020) P: 0.912 (0.020) R: 0.913 (0.021)	F1: 0.911 (0.029) P: 0.913 (0.029) R: 0.912 (0.029)	F1: 0.903 (0.029) P: 0.906 (0.027) R: 0.905 (0.029)
framing	6 secs	6 secs	6 secs	6 secs
classifier hyperparameters	'C': 1 'solver': 'liblinear'	'bootstrap': False 'max_features': 'auto' 'n_estimators': 100	'C': 10.0 'gamma': 'scale' 'kernel': 'rbf'	'activation': 'relu' 'hidden_layer_sizes': [50, 50] 'solver': 'adam'

(i.e., not used yet on the training process) and, on the other hand, correspond to a new subject, possibly with a "different" handwashing behaviour. This analysis allows us to assess the potential of a solution for identifying washing moments with subjects who were not involved in the training process. For the sake of simplicity (since the features would be virtually the same, as seen in the previous analysis) and to reduce analysis times, the initial filtering process was dropped and scaling was set to Min-Max.

As shown in Table 5, when considering binary classification, the classification algorithm with the best performance was SVM (F1: 91.7%, precision: 92.9%, and recall: 91.8%), although with performance very similar to RF (F1: 91.4%, precision: 92.1%, and recall: 91.5%) and also ANN (F1: 90.6%, precision: 91.6%, and recall: 90.7%). In the case of LR, performance falls by only 2.2% with respect to SVM. As can be seen in the table, the performance obtained using only the training data (corresponding to 14 subjects) is practically the same as the one obtained including the validation data (i.e., using the remaining subject). So it can be stated that the obtained solution is generalizable.

Using the pipeline corresponding to the SVM classifier, the F1 score obtained was 94% (precision: 98.2% and recall: 91.6%) for the identification of washing moments, with an overlapping of 0.868 (SD: 0.117). These values are obtained with a convolution signal of 18 seconds.

When differentiating between types of hand washing (cf. Table 6), performance lowers considerably, although it remains relatively high overall. The best pipeline is achieved by using an SVM classifier, which, in this case, is significantly ahead of the other three classifiers. Using SVM, the overall F1 score achieved by evaluating the samples of the external subjects is 80.3% (precision: 81.7%, recall: 81%). This is a remarkable output, but it should be noted that in the classification of frames corresponding to fragments of hands-free washing activities, the F1 is reduced to 53.9% (precision: 55.2%, recall: 56.8%). This compares with the F1 of 77.2% (precision: 55.2%, recall: 56.8%) corresponding to the classification of WHO washing activities

TABLE 6. Best scores for the "Almost all vs one" analysis, considering 3 frame classes (NO-washing/FREE-washing/WHO-washing), in the Lab Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.690 (0.106) P: 0.730 (0.087) R: 0.706 (0.107)	F1: 0.737 (0.099) P: 0.746 (0.098) R: 0.745 (0.100)	F1: 0.803 (0.065) P: 0.817 (0.068) R: 0.810 (0.060),	F1: 0.717(0.110) P: 0.730 (0.108) R: 0.730 (0.107)
testing NO hand-washing scores	F1: 0.904 (0.061) P: 0.895 (0.099) R: 0.924 (0.071)	F1: 0.923 (0.044) P: 0.923 (0.083) R: 0.931 (0.063)	F1: 0.922 (0.055) P: 0.920 (0.068) R: 0.932 (0.085)	F1: 0.922 (0.058) P: 0.914 (0.077) R: 0.936 (0.072)
testing WHO hand-washing scores	F1: 0.654 (0.124) P: 0.715 (0.123) R: 0.643 (0.181)	F1: 0.699 (0.089) P: 0.705 (0.112) R: 0.709 (0.116)	F1: 0.722 (0.098) P: 0.758 (0.129) R: 0.714 (0.142)	F1: 0.703 (0.115) P: 0.740 (0.106) R: 0.692 (0.160)
testing FREE hand-washing scores	F1: 0.511 (0.225) P: 0.580 (0.230) R: 0.552 (0.307)	F1: 0.590 (0.218) P: 0.609 (0.211) R: 0.594 (0.260)	F1: 0.539 (0.199) P: 0.552 (0.206) R: 0.568 (0.264)	F1: 0.528 (0.230) P: 0.535 (0.218) R: 0.561 (0.289)
training scores	F1: 0.707 (0.039) P: 0.722 (0.037) R: 0.702 (0.042)	F1: 0.752 (0.041) P: 0.760 (0.036) R: 0.751 (0.043)	F1: 0.745 (0.040) P: 0.754 (0.042) R: 0.745 (0.040)	F1: 0.730 (0.051) P: 0.744 (0.047) R: 0.729 (0.051)
framing	6 secs	6 secs	6 secs	6 secs
classifier hyperparameters	'C': 1 'solver': 'liblinear'	'bootstrap': False 'max_features': 'auto' 'n_estimators': 100	'C': 10.0 'gamma': 'scale' 'kernel': 'rbf'	'activation': 'relu' 'hidden_layer_sizes': [50, 50] 'solver': 'adam'

frames. This should not be considered strange at all, since hands-free washing presents much more variability, especially if different people are considered.

With a pipeline based on a SVM classifier and an 18-second square convolution signal (three times the frame size), an identification of WHO washing moments is achieved with an average value for F1 of 92% (precision: 87.8%, recall: 96.6%) and an average overlapping of 0.837. In the identification of free washing moments, an F1 of 69.8% (precision: 66.6%, recall: 73.3%) and an overlapping of 0.50 are obtained. As expected, the performance of free washing identification is significantly lower than the performance of WHO compliant hand washing.

B. 7-DAYS EXPERIMENTS

The one-week test scenario involved 6 subjects (50% women), who regularly wore the smartwatch on a day-to-day basis (except for their nightly rest), while performing their ordinary tasks. Participants were required to use the app installed on the clock to indicate the beginning and the end of each hand washing (as well as the type). As sometimes volunteers forgot to press the button to signalize the end of the hand washing, it was decided to eliminate data periods labeled as hand washing that lasted longer than 2 minutes.

In total, 382 hours of data were collected in free-live conditions, including 365 washes, of which 186 are free hand washes (an average of 31 per user) and 179 are WHO compliant washes (an average of 29.83 per user). The distribution of hand washing duration for each of the participants is shown in Fig. 7. The average duration of a free hand washing is 36.78 seconds (SD: 7.37) and the average duration of WHO hand washes is 61.01 seconds (SD: 9.88). It can be noted, however, that there are significant differences among the subjects, both in free and normative handwashing.

As the amount of data corresponding to non-washing periods is larger than the corresponding to washing periods, it was necessary to carry out a downsampling process, in order to obtain a balanced training sample set and, thus, avoid classification bias. After considering several possibilities of downsampling, it was decided to generate a set of training samples made up of the time frames containing washes of both types and with an equal number of time frames without washes. The latter were obtained randomly from the data corresponding to periods without washes. Its duration is the average duration of free and WHO compliant washes. Five datasets were constructed in this way and the analysis was carried out with each of them individually. The following subsections describe the results for the 3 analysis processes carried out.

1) ALL VERSUS ALL

In this analysis, data from all participants is used to train the classifiers, to obtain the optimal hyperparameters and to

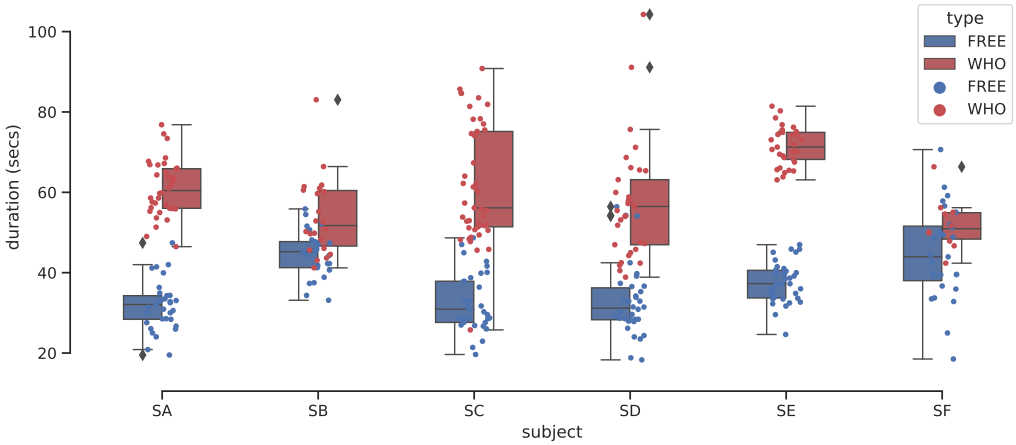

FIGURE 7. Distribution of the duration of hand washing for each of the participants on the "7-days Experiments" scenario.

TABLE 7. Best scores for the "all vs all" analysis, considering 2 frame classes (washing/no-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
global scores	F1: 0.954 (0.017) P: 0.956 (0.014) R: 0.953 (0.018)	F1: 0.960 (0.024) P: 0.961 (0.021) R: 0.959 (0.025)	F1: 0.965 (0.016) P: 0.966 (0.014) R: 0.964 (0.018)	F1: 0.960 (0.012) P: 0.966 (0.011) R: 0.959 (0.013)
classifier hyperparameters	"C": 10 "solver": "newton-cg"	"bootstrap": false "max_features": "auto" "n_estimators": 200	"C": 10.0 "gamma": "scale" "kernel": "rbf"	"activation": "tanh" "hidden_layer_sizes": [50, 50] "solver": "adam"

TABLE 8. Best scores for the "all vs all" analysis, considering 3 frame classes (NO-washing/FREE-washing/WHO-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
global scores	F1: 0.799 (0.026) P: 0.807 (0.024) R: 0.794 (0.027)	F1: 0.839 (0.036) P: 0.849 (0.033) R: 0.831 (0.039)	F1: 0.830 (0.031) P: 0.835 (0.028) R: 0.829 (0.035)	F1: 0.828 (0.033) P: 0.830 (0.028) R: 0.828 (0.037)
classifier hyperparameters	'C': 10 'solver': 'newton-cg'	'bootstrap': False 'max_features': 'auto' 'n_estimators': 100	'C': 10.0 'gamma': 'scale' 'kernel': 'rbf'	'activation': 'relu' 'hidden_layer_sizes': [50, 50] 'solver': 'lbfgs'

evaluate the performance. Initially, the following hyperparameters were selected: no filtering, Min-Max scaling, and 6-second frame time.

On Table 7, the results obtained for each type of classifier, considering a binary classification (washing/non-washing), are shown. As can be seen, the difference in the scores obtained is minimal. The pipeline based on the SVM classifier with ''rbf kernel'', as in most of the previous analysis carried out, offers a slight advantage over the alternatives, reaching an F1 of 96.5% (precision: 96.6% and recall: 96.4%). Using the pipeline with this classifier, and taking an 18-second square convolution signal, a wash identification solution is obtained with an F1 of 97% (precision: 99% and recall: 95%) and an overlapping of 0.937 (SD: 0.135).

If we consider the two types of washes, the classifier with the best behavior is the one based on the RF algorithm (cf. Table 8), although with scores very close to SVM and ANN, and even to LR. The F1 score achieved with RF using an rbf kernel is 83.9% (precision: 84.9% and recall: 83.1%). A data pipeline based on this classifier allows to build a solution for identifying free wash moments with an F1 of 72% (precision: 95% and recall: 58%) and an overlapping of 0.764 (SD: 0.216). In the identification of WHO washing moments,

TABLE 9. Best scores for the "almost all vs one" analysis, considering 2 frame classes (washing/no-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.932 (0.042) P: 0.941 (0.029) R: 0.930 (0.046)	F1: 0.944 (0.040) P: 0.950 (0.031) R: 0.942 (0.044)	F1: 0.943 (0.040) P: 0.951 (0.028) R: 0.940 (0.045)	F1: 0.937 (0.037) P: 0.942 (0.030) R: 0.938 (0.036)
training scores	F1: 0.956 (0.011) P: 0.957 (0.011) R: 0.955 (0.012)	F1: 0.962 (0.013) P: 0.963 (0.012) R: 0.961 (0.014)	F1: 0.963 (0.011) P: 0.965 (0.009) R: 0.962 (0.012)	F1: 0.955 (0.010) P: 0.955 (0.011) R: 0.955 (0.013)
classifier hyperparameters	'C': 10 'solver'='newton-cg'	'bootstrap': False 'max_features': 'auto' 'n_estimators': 200	'C': 10.0 'gamma': 'scale' 'kernel': 'rbf'	'activation': 'relu' 'hidden_layer_sizes': [50, 50] 'solver': 'lbfgs'

it achieves an F1 of 95% (precision: 92% and recall: 98%), and the overlapping is 0.824 (SD: 0.09).

2) ALMOST ALL VERSUS ONE

In order to study the potential of building a user-independent system for identifying washing moments, an analysis similar to the one described in section IV-A2 was carried out. A pipeline was obtained using data from 5 subjects and their performance was evaluated with the remaining subject. This was repeated 6 times, varying in each iteration the validation subject.

In Table 9 the reader can see the scores obtained when no difference is made between wash types for the 6-second frame classification process. The results are almost identical for the RF-based classifier (with 200 estimators), SVM (with rbf kernel), and also very close to the ANN-based classifier (with 2 layers of 50 neurons per layer and relu activation function). The full data pipeline with an RF classifier allows to obtain a solution for the identification of washing moments with an F1 of 91.4% and an overlapping of 0.889 (SD: 0.151).

When the separation is made between WHO washes and free washes, the performance is reduced, especially in the classification of fragments corresponding to free washing, as could be expected. In Table 10 the classification scores obtained for the different algorithms considered are presented. As the reader may note, a similar performance is achieved in the three classification approaches. Nevertheless, it can be noted that the use of the LR algorithm results in a slightly worse performance. With RF an overall F1 of 76.3% is obtained, although the classification of free washing frames only obtains an F1 of 57.9% with a recall noticeable low that does not reach more than 55.6% (and also with a noticeable standard deviation).

In this analysis, special attention is paid to how the length of the frame affects the performance at the output of the classifier (stage 5 in the data processing pipeline represented in Fig. 2) for the different algorithms considered. Figs. 8, 9, 10, and 11 show, respectively for LR, RF, SVM, and ANN, the variations of F1, precision and recall as a function of the frame size. Both, the global scores and the scores for the classification of free washing and normative washing frames are represented. As can be seen, the best performance is obtained for a size of 6 seconds in the case

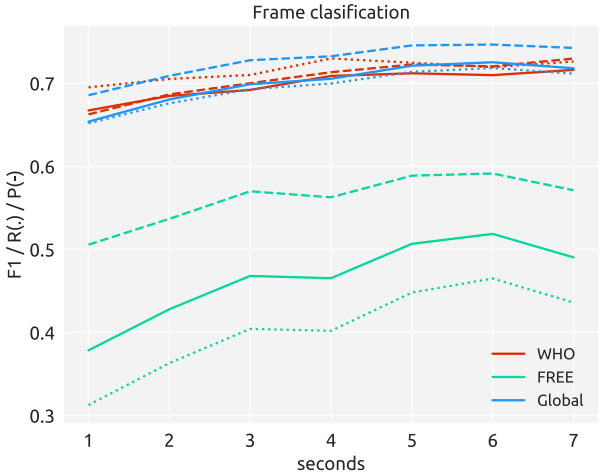

FIGURE 8. Performance vs frame size in the LR classification stage.

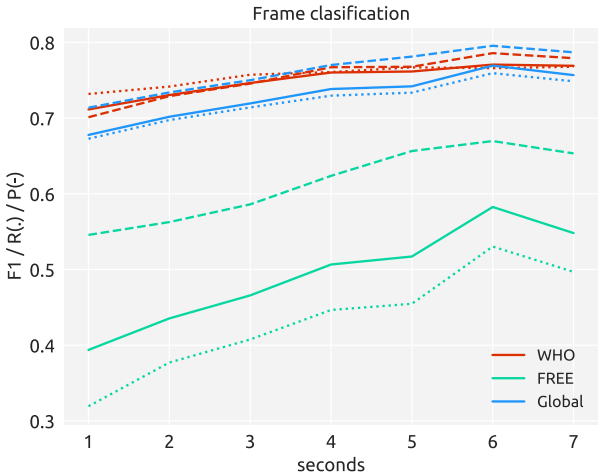

FIGURE 9. Performance vs frame size in the RF classification stage.

TABLE 10. Best scores for the "almost all vs one" analysis, considering 3 frame classes (NO-washing/FREE-washing/WHO-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.725 (0.037) P: 0.746 (0.032) R: 0.718 (0.043)	F1: 0.769 (0.061) P: 0.795 (0.044) R: 0.759 (0.069)	F1: 0.763 (0.070) P: 0.785 (0.057) R: 0.755 (0.077)	P: 0.757 (0.044) R: 0.733 (0.057) F1: 0.739 (0.052)
testing NO hand-washing scores	F1: 0.947 (0.032) P: 0.927 (0.065) R: 0.970 (0.008)	F1: 0.954 (0.033) P: 0.931 (0.062) R: 0.981 (0.006)	F1: 0.953 (0.032) P: 0.930 (0.065) R: 0.979 (0.007)	F1: 0.947 (0.031) P: 0.931 (0.061) R: 0.966 (0.013)
testing WHO hand-washing scores	F1: 0.709 (0.064) P: 0.720 (0.058) R: 0.719 (0.136)	F1: 0.770 (0.054) P: 0.785 (0.049) R: 0.765 (0.099)	F1: 0.756 (0.065) P: 0.801 (0.081) R: 0.731 (0.108)	F1: 0.733 (0.066) P: 0.745 (0.063) R: 0.736 (0.122)
testing FREE hand-washing scores	F1: 0.518 (0.064) P: 0.591 (0.079) R: 0.464 (0.062)	F1: 0.582 (0.119) P: 0.669 (0.073) R: 0.530 (0.169)	F1: 0.579 (0.141) P: 0.623 (0.081) R: 0.556 (0.203)	F1: 0.537 (0.104) P: 0.596 (0.079) R: 0.497 (0.128)
training scores	F1: 0.794 (0.015) P: 0.804 (0.012) R: 0.788 (0.018)	F1: 0.832 (0.020) P: 0.846 (0.015) R: 0.824 (0.024)	F1: 0.833 (0.018) P: 0.840 (0.017) R: 0.830 (0.026)	F1: 0.817 (0.020) P: 0.821 (0.018) R: 0.816 (0.022)
classifier hyperparameters	'C': 10 'solver': 'newton-cg'	'bootstrap': False 'max_features': 'auto' 'n_estimators': 200	'C': 10.0 'gamma': 'scale' 'kernel': 'rbf'	'activation': 'relu' 'hidden_layer_sizes': [50, 50] 'solver': 'lbfgs'

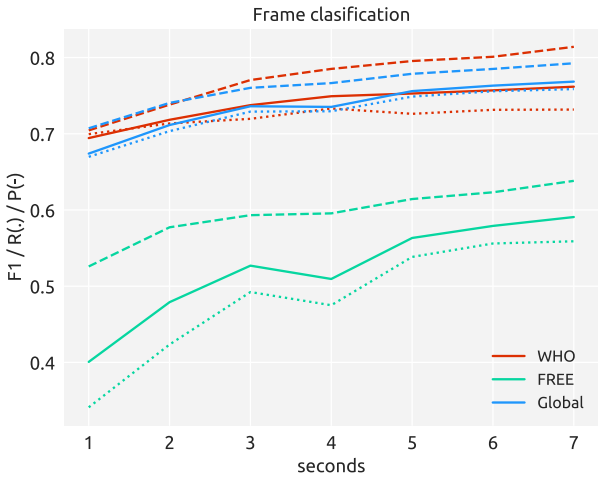

FIGURE 10. Performance vs frame size in the SVM classification stage.

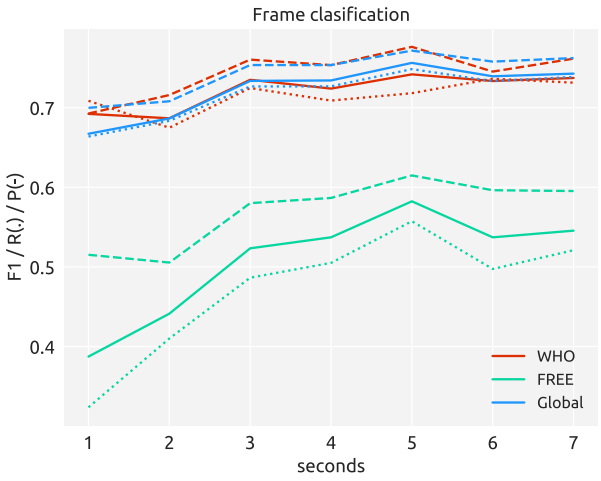

FIGURE 11. Performance vs frame size in the ANN classification stage.

of LR and RF classifiers. For an ANN classifier, the ideal frame size is 5 seconds. In the case of SVM, the best scores are obtained around 7 seconds, although with a very small difference with respect to 6 seconds (and even 5 seconds). Therefore, we consider that, globally, the ideal frame size for the classification process is 5 or 6 seconds.

Considering the output of the full data pipeline, a solution for WHO hand washing moment identification with a global F1 of approximately 90% is achieved using both, RF and SVM. However, in hand washing identification the F1 is reduced to 46%, obtaining a precision of 90% but a recall of only 30.4% and an overlapping of 0.663.

Figs. 12 and 13 show, for the SVM-based pipeline, how the frame size impacts on the performance regarding moment identification, both in F1 and overlapping, respectively. As can be observed, a frame size of 6 seconds is the most appropriate option for the implementation of an optimal final solution. Although the decision of using 7 seconds

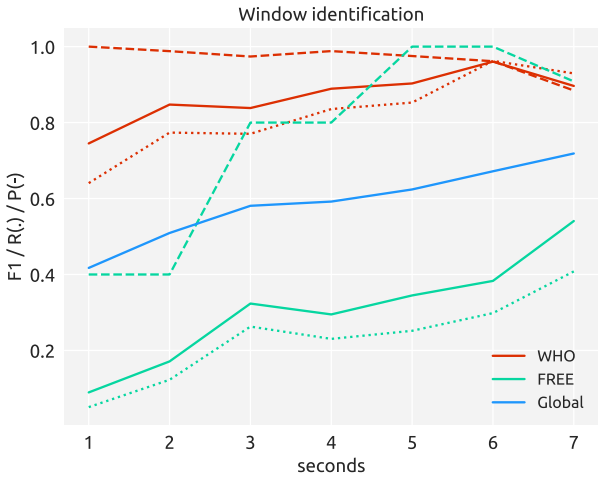

FIGURE 12. Performance vs frame time in the moment identification stage (SVM classification).

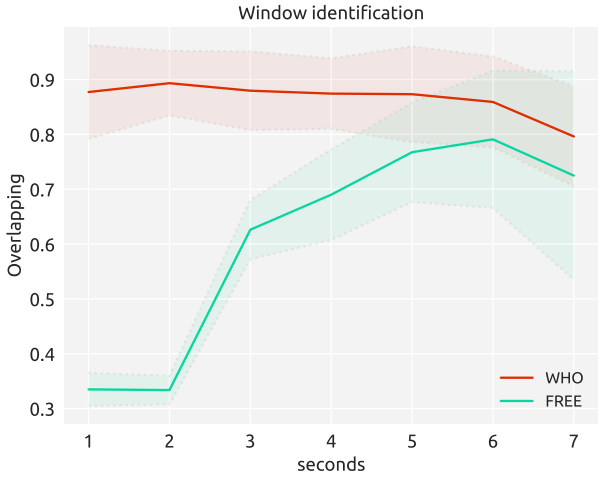

FIGURE 13. Overlapping vs frame time in the moment identification stage (SVM classification).

improves the recall of free wash identification, this is at the cost of a notable decrease in precision, as well as in the WHO wash identification scores and in overlappings (as shown in Fig. 12).

3) LAB VERSUS ALL

The last analysis made with the data collected in the 7-days experiments uses the data from the laboratory experiments to train the classifiers. In particular, a full data pipeline was obtained from all the data collected in the laboratory and its performance was evaluated in the 7-day free-live conditions datasets. Tables 11 and 12 show the results obtained, considering binary classification and classification differentiating free washes and WHO washes.

As can be noted, high scores are obtained in all four types of algorithms for binary classification. Being the highest F1, of 90%, in the case of RF. In all four cases, the training and testing scores are similar.

For NO-washing/FREE-washing/WHO-washing classification, an overall F1 of 73.2% is obtained with RF, although the F1 for free washing frames drops to 55.9%. In any case, the similarities between training and testing performances are maintained.

A solution for the identification of washing moments can reach, for the case of binary classification, an F1 of 89.9% (precision: 88.8% and recall: 91%) with an overlapping of 0.771 (SD: 0.272). A solution for the identification of washing moments that considers free washing and WHO washing can reach, in the first case, an F1 of 55.3% (precision: 86.7% and recall: 40.6%) with an overlapping of 0.75 (SD: 0.195). And, in the second case, it is possible to obtain an F1 of 87.7% (precision: 81.6% and recall: 94.8%) with an overlapping of 0.817 (SD: 0.097).

C. 30-DAYS EXPERIMENT

In the 30-days tests scenario in free-life conditions participated 2 volunteer subjects, both men. One of the subjects collected 192.5 hours of data, including 128 free washes with a duration of between 18.83 and 80.44 seconds (mean: 33.35, SD: 7.97) and 117 WHO washes with a duration of between 18.22 and 78.12 seconds (mean: 58.3 and SD: 10.20). The second subject collected 174.76 hours of data. He performed 129 free washes, taking between 23.39 and 81.67 seconds (mean: 40.30 and SD: 9.67) and 123 WHO washes, taking between 39.60 and 84.73 seconds (mean: 54.71 and SD: 8.65).

The data collected in these experiments was used to evaluate the performance of pipelines obtained in the previous scenarios, both the laboratory and the 7-days in free-life conditions experiments. It is worth mentioning that the subjects also participated in the previous scenarios.

1) LAB VERSUS ALL

By training the classifiers and obtaining the ideal hyperparameters from the data collected in the laboratory scenario, the performances shown in Table 13 are achieved when the optimal pipelines obtained are applied to the data captured by the 2 subjects participating in the 30-days experiments. The pipeline with the best performance in terms of frame classification was the RF-based algorithm with F1, precision and recall higher than 96%. Using LR and SVM, identical results are obtained, which are very close to RF. ANN in this case has the worst performance, although the differences among the four algorithms are small.

The four pipelines obtain almost identical washing moment identification performance, with F1, precision and recall over 97% and an overlapping between 0.90 and 0.95.

Table 14 shows the results when the classification distinguishes between free washes and normative washes. In this case, the pipeline with the best performance is the one based on SVM (although with small deviations from RF and LR),

TABLE 11. Best scores for the "lab vs one" analysis, considering 2 frame classes (washing/no-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.863 P: 0.863 R: 0.862	F1: 0.909 P: 0.910 R: 0.908	F1: 0.863 P: 0.863 R: 0.863	F1: 0.869 P: 0.872 R: 0.866
training scores	F1: 0.884 (0.066) P: 0.900 (0.054) R: 0.887 (0.063)	F1: 0.920 (0.041) P: 0.920 (0.037) R: 0.912 (0.040)	F1: 0.900 (0.052) P: 0.913 (0.041) R: 0.903 (0.049)	F1: 0.853 (0.059) P: 0.865 (0.057) R: 0.853 (0.057)
classifier hyperparameters	'C': 10 'solver'='newton-cg'	'bootstrap': False 'max_features': 'auto' 'n_estimators': 200	'C': 10.0 'gamma': 'scale' 'kernel': 'rbf'	'activation': 'relu' 'hidden_layer_sizes': [50, 50] 'solver': 'lbfgs'

TABLE 12. Best scores for the "lab vs one" analysis, considering 3 frame classes (NO-washing/FREE-washing/WHO-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.667 P: 0.671 R: 0.664	F1: 0.732 P: 0.746 R: 0.722	F1: 0.698 P: 0.70 R: 0.695	F1: 0.624 P: 0.626 R: 0.631
testing NO hand-washing scores	F1: 0.887 P: 0.880 R: 0.893	F1: 0.928 P: 0.900 R: 0.958	F1: 0.885 P: 0.868 R: 0.903	F1: 0.839 P: 0.850 R: 0.829
testing WHO hand-washing scores	F1: 0.622 P: 0.615 R: 0.630	F1: 0.710 P: 0.722 R: 0.698	F1: 0.642 P: 0.676 R: 0.611	F1: 0.599 P: 0.652 R: 0.554
testing FREE hand-washing scores	F1: 0.493 P: 0.519 R: 0.469	F1: 0.559 P: 0.617 R: 0.511	F1: 0.565 P: 0.559 R: 0.57	F1: 0.433 P: 0.376 R: 0.511
training scores	F1: 0.691 (0.036) P: 0.711 (0.051) R: 0.694 (0.039)	F1: 0.704 (0.049) P: 0.741 (0.077) R: 0.706 (0.043)	F1: 0.704 (0.084) P: 0.720 (0.093) R: 0.703 (0.082)	F1: 0.691 (0.044) P: 0.702 (0.048) R: 0.699 (0.048)
classifier hyperparameters	'C': 10 'solver': 'newton-cg'	'bootstrap': False 'max_features': 'auto' 'n_estimators': 200	'C': 10.0 'gamma': 'scale' 'kernel': 'rbf'	'activation': 'relu' 'hidden_layer_sizes': [50, 50] 'solver': 'lbfgs'

TABLE 13. Best scores for the "lab vs all" analysis, considering 2 frame classes (washing/no-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.955 P: 0.954 R: 0.958	F1: 0.964 P: 0.963 R: 0.966	F1: 0.955 P: 0.954 R: 0.958	F1: 0.932 P: 0.931 R: 0.934

with an overall F1 of 83.6% and a remarkable F1 of 74.4% (precision: 70.5% and recall: 78.6%) for the classification of frames corresponding to free hand washes.

Using the pipeline with SVM, a solution for the identification of washing moments is obtained with an F1 of 96.1% and an overlapping of 0.819 (SD: 0.099) for WHO washing. For free washes, an F1 of 83.6% (precision: 88.1% and recall: 79.6%) and an overlapping of 0.824 (SD: 0.202) are achieved.

2) 7-DAYS VERSUS ALL

In the last analysis conducted in this study, the performance of the pipelines previously obtained from the 7-days experiments was evaluated using the data of the two subjects participating in the 30-days experiments. Table 15 shows the

TABLE 14. Best scores for the "lab vs all" analysis, considering 3 frame classes (NO-washing/FREE-washing/WHO-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.820 P: 0.815 R: 0.826	F1: 0.831 P: 0.830 R: 0.833	F1: 0.836 P: 0.830 R: 0.843	F1: 0.764 P: 0.756 R: 0.778
testing NO hand-washing scores	F1: 0.964 P: 0.988 R: 0.942	F1: 0.971 P: 0.981 R: 0.962	F1: 0.955 P: 0.970 R: 0.941	F1: 0.908 P: 0.962 R: 0.859
testing WHO hand-washing scores	F1: 0.788 P: 0.760 R: 0.817	F1: 0.807 P: 0.789 R: 0.826	F1: 0.808 P: 0.814 R: 0.802	F1: 0.762 P: 0.715 R: 0.815
testing FREE hand-washing scores	F1: 0.708 P: 0.698 R: 0.719	F1: 0.716 P: 0.720 R: 0.712	F1: 0.744 P: 0.705 R: 0.786	F1: 0.624 P: 0.591 R: 0.660

TABLE 15. Best scores for the "7-days vs all" analysis, considering 2 frame classes (washing/NO-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.981 P: 0.979 R: 0.982	F1: 0.992 P: 0.992 R: 0.993	F1: 0.990 P: 0.990 R: 0.991	F1: 0.987 P: 0.986 R: 0.988

TABLE 16. Best scores for the "7-days vs all" analysis, considering 3 frame classes (NO-washing/FREE-washing/WHO-washing), in the 7-days Experiments scenario.

	LR	RF	SVM	ANN
testing scores	F1: 0.871 P: 0.874 R: 0.871	F1: 0.962 P: 0.959 R: 0.966	F1: 0.926 P: 0.920 R: 0.934	F1: 0.915 P: 0.910 R: 0.920
testing NO hand-washing scores	F1: 0.985 P: 0.999 R: 0.972	F1: 0.994 P: 0.998 R: 0.991	F1: 0.992 P: 1.0, R: 0.985	F1: 0.986 P: 0.996 R: 0.976
testing WHO hand-washing scores	F1: 0.857 P: 0.818 R: 0.899	F1: 0.957 P: 0.963 R: 0.950	F1: 0.913 P: 0.931 R: 0.896	F1: 0.901 P: 0.895 R: 0.908
testing FREE hand-washing scores	F1: 0.772 P: 0.805 R: 0.741	F1: 0.936 P: 0.916 R: 0.958	F1: 0.873 P: 0.830 R: 0.922	F1: 0.857 P: 0.840 R: 0.875

results applied to binary classification. As the reader may note, outstanding classification results are achieved. As a matter of fact, F1 reached values higher than 98% in all 4 cases. With this successful classification rate, the obtained solution for moment identification reaches an F1 score close to 99% and an overlapping of 0.98.

Table 16 shows the results when distinguishing between free washing and WHO washing. The scores are also very high for RF, SVM and ANN. With the RF algorithm, a global classification F1 of 96.2% is obtained (precision: 95.9% and recall: 96.6%). In this case, even the classification of frames corresponding to free washes obtains very high scores, with an F1 of 93.6% (precision: 91.6% and recall: 95.8%).

Using an RF-based pipeline, a WHO hand washing identification solution with F1 of 98.8% (precision: 99.2% and recall: 98.4%) and an overlapping of 0.808 (SD: 0.064) is achieved. Besides that, in free hand washing identification, a remarkable F1 of 97.6% (precision: 100% and recall: 95.3%) is obtained, with an overlapping of 0.965 (SD: 0.097).

V. DISCUSSION

The various analyses carried out with the data collected in the different types of experiments have shown the feasibility of developing a handwashing identification solution with acceptable performance. This solution is aimed at detecting handwashing moments from the continuous flow of inertial measurement data collected from a popular commercial smartwatch, mainly when the handwashing follows the guidelines recommended by the WHO. The reader should note that the purpose of this work was not to explore the possibility of identifying the actual execution of the 11 individual steps that make up the WHO protocol (c.f. Fig. 1), but to identify the proximity of a certain handwashing to such protocol in a holistic manner.

In line with the existing literature in the domain, it has been proven that using data collected in a controlled manner, it is possible to obtain a user-dependent classifier capable of detecting 6-second frames corresponding to hand washing with an F1 of over 90%. When free washes are classified separately from WHO washes, the effectiveness of the classification is reduced to an F1 of 77.8%. In any case, based on this classification, it is possible to build a solution that allows identifying washing moments with an F1 higher than 90% and an overlapping around 0.90. When building a user-independent solution, that is, obtaining the data pipeline with data not coming from the subject being evaluated, the identification of WHO washing moments maintains a very high performance, with an F1 higher than 90%. However, the identification of free washes diminishes significantly, lowering the F1 to approximately 70% and an overlapping of 0.50, which indicates that the prediction of the instant of beginning of the wash and its duration is not very accurate.

Using data collected in free-live conditions, a user-dependent solution, when distinguishing the types of washing, can reach, for the identification of WHO washing moments, an F1 of 95% (overlapping: 0.824), and for the identification of free wash moments an F1 of 72% (overlapping: 0.764). A user-independent approach achieves a solution for the identification of WHO hand washing moments with an F1 of 90%,

but in the identification of free hand washing moments the F1 falls to 46%, suffering in particular the recall (i.e. although the detected washings are correct in their great majority, an important proportion of free washings is not detected). This substantial decrease in performance when it comes to identifying free washes may seem to be quite limiting in some contexts for the obtained solution. However, this result is logical and, to a certain extent, foreseeable, since each person washes his hands in a very different way when he/she does not follow pre-established guidelines. Thus, a predictive model obtained with a group of outsiders is not going to provide good results in this task, except, perhaps, if this group is large enough, and, to some extent, can recreate a valid model for the variety of behaviors that exist among people. This is an issue that can be analyzed in the future, with a much greater number of participants. The fact that a user-dependent model provides good results in identifying free washing indicates that the same person tends to wash his or her hands always in a similar way, although quite differently from the WHO-recommended washing protocol.

Applying a pipeline obtained with the data from the laboratory experiments in the identification of washing moments of the 6 participants in the 7-days scenario, an F1 higher than 85% is obtained for WHO washes and an F1 higher than 55% for free washes (overlapping higher than 0.75 in both cases). This same pipeline applied to the 2 subjects who participated in the 30-days scenario reaches high scores, with an F1 of 96.1% and 83.6%, respectively for WHO and free washes, and an overlapping above 0.80. These results show that it is feasible to build a high-performance handwashing identification solution based on data collected only in controlled environments. It should be noted, however, that this analysis results in a user-dependent predictive model.

A final analysis tested a pipeline obtained from the 7-days experiments with the subjects who participated in the 30-days scenario. In this case, the results are really good, obtaining F1s and overlappings greater than 95% and 0.80, respectively, for both free and WHO washes.

In short, the results obtained allow us to conclude that, with sufficient data, it is possible to obtain a very effective user-dependent solution for the identification of both types of washing, both free and according to the WHO protocol. A user-independent model, on the other hand, results in an effective solution for washes that follow the recommendations provided by the WHO (and therefore have an underlying pattern) but with poorer results in the identification of free washes. It is worth mentioning that the pipelines obtained have always tried to maximize the global F1, and not independently for WHO and free washes. An approach that seeks to maximize specific F1 for free washes is likely to perform better in this task. This approach would result in two pipelines, one for each type of wash, thus increasing the computational complexity of the final solution.

In the analysis processes carried out, the 4 classification algorithms that usually provide better results in the field of activity recognition were tested. It has been proven that very similar performances are obtained with RF, SVN, and ANN. When the LR algorithm is applied, the performance is usually somewhat lower, although the differences remain, in most cases, not very significant. Depending on the type of analysis carried out, in some cases the best models are obtained with SVN and in others with RF, always closely followed by ANN. The algorithm that allows to obtain better performances in more analysis has been SVM, although this must be considered in a relative way, because the differences with RF, and even with ANN, are small, so a finer tuning of the hyperparameters can turn the table.

Regarding to the prediction algorithms used in our study, the reader may note the lack of an approach closer to the area of Deep Learning (DL), a hot topic in the domain. Indeed, the present study has used algorithms based on neural networks, the base model for DL, although the networks used were structurally simple. Tests have not been carried out with deep neural networks since obtaining the optimal parameters, i.e., training, in these complex networks requires huge amounts of training data. This would require the participation in the experiments of a larger number of volunteer subjects for longer periods. Currently, there are no, to the best knowledge of the authors, pre-trained deep neural networks that could serve as a starting point to obtain, in a reasonable way, appropriate DL predictors for our purpose.

The main limitation of the study carried out has been, in fact, the number of volunteer subjects in the experiments to obtain training data. Certainly, a larger number of labelled samples would allow us to obtain more general conclusions (and, to a certain extent, more optimized prediction models). However, the number of training data obtained (more than 600 hours of samples, which significantly exceeds those used in similar studies found in the literature) has allowed us to carry out multiple analytical processes, from different perspectives, trying to highlight possible weaknesses and limiting biases in the conclusions reached. From the point of view of the authors, the main problem derived from not being able to count on a larger number of volunteers has been the impossibility of obtaining effective user-independent models for the identification of freehand washing, as above mentioned.

VI. CONCLUSION

In this paper we have presented a study on the feasibility of building a solution that, using data collected from a wearable wrist device, can identify the moments when a person performs a hand washing and if this washing is compliant with the WHO protocol. This study is an advance on existing proposals in the academic literature in three fundamental aspects: first, it explicitly distinguishes between free washing and washing in accordance with the protocol defined by the WHO; second, it makes use of datasets obtained in free-live conditions, specifically more than 600 hours of data have been collected; and third, the study is not limited to

classifying activities as corresponding to washing, but allows identifying when and for how long this activity has taken place.

The study has shown that it is possible to build a user-dependent solution that is very effective in detecting any type of wash, and a user-independent solution that is very effective in detecting WHO washes, but more limited in detecting free washes. This solution uses only data from the accelerometer and the gyroscope data, which facilitates its implementation, since the vast majority of commercial devices of common use and low cost include these sensors, and maximizes user privacy, since it does not use sensors such as the microphone.

As a future line, it is proposed to build a telematics platform to provide recommendations adapted to the behavior of the individual and, above all, to promote self-reflection and awareness of the importance of hand washing. This will involve obtaining predictive models complementary to those already obtained to, once a WHO handwashing has been detected, identify the individual steps that make up the WHO protocol and, based on the correct/incorrect realization of these steps, give the proper recommendations. To obtain these prediction models, the use of DL techniques is considered, as these techniques could be more feasible since the collection of training data corresponding to non-handwashing is not required. Also, the use of optimization techniques such as adaptive sliding framing [32] is contemplated.

REFERENCES

[1] *WHO Guidelines on Hand Hygiene in Health Care*, World Health Org., Geneva, Switzerland, 2009.

[2] J. M. Santos-Gago, M. Ramos-Merino, S. Vallarades-Rodriguez, L. M. Álvarez-Sabucedo, M. J. Fernández-Iglesias, and J. L. García-Soidán, "Innovative use of wrist-worn wearable devices in the sports domain: A systematic review," *Electronics*, vol. 8, no. 11, p. 1257, Nov. 2019.

[3] M. Kos and I. Kramberger, "A wearable device and system for movement and biometric data acquisition for sports applications," *IEEE Access*, vol. 5, pp. 6411–6420, 2017.

[4] J. Khakurel, H. Melkas, and J. Porras, "Tapping into the wearable device revolution in the work environment: A systematic review," *Inf. Technol. People*, vol. 31, no. 3, pp. 791–818, May 2018.

[5] K. Maltseva, "Wearables in the workplace: The brave new world of employee engagement," *Bus. Horizons*, vol. 63, no. 4, pp. 493–505, Jul. 2020.

[6] G. Koutromanos and G. Kazakou, "The use of smart wearables in primary and secondary education: A systematic review," *Themes eLearn.*, vol. 13, pp. 33–53, Sep. 2020.

[7] F. de Arriba-Pérez, M. Caeiro-Rodríguez, and J. M. Santos-Gago, "Towards the use of commercial wrist wearables in education," in *Proc. 4th Experiment@International Conf. (exp.at)*, Jun. 2017, pp. 323–328.

[8] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution," *Personalized Med.*, vol. 15, no. 5, pp. 429–448, Sep. 2018.

[9] A. Kamišalić, I. Fister, M. Turkanović, and S. Karakatič, "Sensors and functionalities of non-invasive wrist-wearable devices: A review," *Sensors*, vol. 18, no. 6, p. 1714, May 2018.

[10] F. de Arriba-Pérez, M. Caeiro-Rodríguez, and J. Santos-Gago, "Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios," *Sensors*, vol. 16, no. 9, p. 1538, Sep. 2016.

[11] Apple. (2020). *Apple Watch User Guide*. Accessed: Dec. 5, 2020. [Online]. Available: https://support.apple.com/guide/watch/set-up-handwashing-apdc9b9f04a8/watchos

[12] Nani Innovations. *Hands Washing Timer*. Accessed: Dec. 5, 2020. [Online]. Available: https://play.google.com/store/apps/-details?id=com.tmanswap.handwashtimer

[13] J. Logenius. *Ella's Hand Washing Adventure*. Accessed: Dec. 5, 2020. [Online]. Available: https://play.google.com/store/apps/-details?id=air.com.sca.EllaHandWash

[14] SureWash. *SureWash Hand Hygiene*. Accessed: Dec. 5, 2020. [Online]. Available: https://play.google.com/store/apps/details?id=com.surewash.surewash

[15] Samsung. *Hand Wash*. Accessed: Dec. 5, 2020. [Online]. Available: https://galaxystore.samsung.com/geardetail/-com.samsung.washcare

[16] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla, "Sensor-based datasets for human activity recognition—A systematic review of literature," *IEEE Access*, vol. 6, pp. 59192–59210, 2018.

[17] V. Galluzzi, T. Herman, and P. Polgreen, "Hand hygiene duration and technique recognition using wrist-worn sensors," in *Proc. 14th Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2015, pp. 106–117.

[18] L. N. S. Wijayasingha and B. Lo, "A wearable sensing framework for improving personal and oral hygiene for people with developmental disabilities," in *Proc. IEEE Wireless Health (WH)*, Oct. 2016, pp. 1–7.

[19] J. Cherian, V. Rajanna, D. Goldberg, and T. Hammond, "Did you remember to brush? A noninvasive wearable approach to recognizing brushing teeth for elderly care," in *Proc. 11th EAI Int. Conf. Pervas. Comput. Technol. Healthcare (PervasiveHealth)*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 48–57, doi: 10.1145/3154862.3154866.

[20] M. A. Sayeed Mondol and J. A. Stankovic, "HAWAD: Hand washing detection using wrist wearable inertial sensors," in *Proc. 16th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, May 2020, pp. 11–18.

[21] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and smartwatch-based biometrics using activities of daily living," *IEEE Access*, vol. 7, pp. 133190–133202, 2019.

[22] H. Li, S. Chawla, R. Li, S. Jain, G. D. Abowd, T. Starner, C. Zhang, and T. Plötz, "Wristwash: Towards automatic handwashing assessment using a wrist-worn device," in *Proc. ACM Int. Symp. Wearable Comput.*, Oct. 2018, pp. 132–139.

[23] A. Banerjee, V. N. S. A. Amperyani, and S. K. S. Gupta, "Hand hygiene compliance checking system with explainable feedback," in *Proc. 6th ACM Workshop Wearable Syst. Appl. (WearSys)*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 34–36, doi: 10.1145/3396870.3400015.

[24] E. Kutafina, D. Laukamp, R. Bettermann, U. Schroeder, and S. Jonas, "Wearable sensors for eLearning of manual tasks: Using forearm EMG in hand hygiene training," *Sensors*, vol. 16, no. 8, p. 1221, Aug. 2016.

[25] X. Zhang, K. Kadimisetty, K. Yin, C. Ruiz, M. G. Mauk, and C. Liu, "Smart ring: A wearable device for hand hygiene compliance monitoring at the point-of-need," *Microsyst. Technol.*, vol. 25, no. 8, pp. 3105–3110, Aug. 2019, doi: 10.1007/s00542-018-4268-5.

[26] F. de Arriba-Pérez, M. Caeiro-Rodríguez, and J. M. Santos-Gago, "How do you sleep? Using off the shelf wrist wearables to estimate sleep quality, sleepiness level, chronotype and sleep regularity indicators," *J. Ambient Intell. Humanized Comput.*, vol. 9, no. 4, pp. 897–917, 2018.

[27] F. de Arriba-Pérez, J. M. Santos-Gago, M. Caeiro-Rodríguez, and M. Ramos-Merino, "Study of stress detection and proposal of stress-related features using commercial-off-the-shelf wrist wearables," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 12, pp. 4925–4945, Dec. 2019.

[28] F. de Arriba Pérez, J. M. Santos-Gago, M. Caeiro-Rodríguez, and M. J. F. Iglesias, "Evaluation of commercial-off-the-shelf wrist wearables to estimate stress on students," *J. Visualized Exp.*, vol. 136, Jun. 2018, Art. no. 57590.

[29] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. (UbiComp)*, 2015, pp. 1029–1040.

[30] A. Santos, A. Canuto, and A. F. Neto, "A comparative analysis of classification methods to multi-label tasks in different application domains," *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.*, vol. 3, pp. 218–227, Jan. 2011.

[31] J. Opitz and S. Burst, "Macro f1 and macro f1," 2019, *arXiv:1911.03347*. [Online]. Available: http://arxiv.org/abs/1911.03347

[32] C. Ma, W. Li, J. Cao, J. Du, Q. Li, and R. Gravina, "Adaptive sliding window based activity recognition for assisted livings," *Inf. Fusion*, vol. 53, pp. 55–65, Jan. 2020.

JUAN M. SANTOS-GAGO (Member, IEEE) received the M.S. degree in telecommunications engineering, in 1998, and the Ph.D. degree (Hons.) in telematics engineering from the University of Vigo, Spain, in 2008.

Since 2010, he has been an Associate Professor with the Department of Telematics Engineering, University of Vigo, where he has lectured courses on computer architectures, programming, and artificial intelligence. He has participated in several national and international research and development projects, in seven of them as a principal investigator. He is the author of approximately 200 publications, including articles in impact journals and communications in international peer-reviewed conferences. His main research interests include the use of semantic technologies and data analytics techniques to support the construction of advanced services in the eLearning and the eHealth domains.

MATEO RAMOS-MERINO received the M.S. degree in telecommunications engineering, in 2014, and the Ph.D. degree (Hons.) in information and communication technologies from the University of Vigo, Spain, in 2020.

Since 2014, he has been a Research Assistant with the atlanTTic, a research center for telecommunication technologies, University of Vigo. He has lectured courses on computer science and multimedia services. His doctoral dissertation dealt with process mining and traceability techniques in non-exhaustive monitoring environments. His research interests include the development of smart applications in the eHealth domain and the application of wearable devices to support personalized services.

LUIS M. ÁLVAREZ-SABUCEDO received the Ph.D. degree (Hons.) from the Department of Telematics Engineering, University of Vigo, Spain, in 2008.

Since 2010, he has been working as an Associate Professor with the Department of Telematics Engineering, University of Vigo. Within this framework, he develops his teaching and research activity. As a Researcher, he has carried out investigations whose results have been disseminated in international journals and conferences. This active participation is articulated in the context of projects, both publicly and privately founded, as a collaborator and a main researcher. His research interests include applied semantics and web technologies in the areas of eGovernment and eHealth.