APPLIED RESEARCH

Unstructured Handwashing Recognition Using Smartwatch to Reduce Contact Transmission of Pathogens

EMANUELE LATTANZI, LORENZO CALISTI, AND VALERIO FRESCHI
Department of Pure and Applied Sciences, University of Urbino, 61029 Urbino, Italy
Corresponding author: Emanuele Lattanzi (emanuele.lattanzi@uniurb.it)

ABSTRACT Current guidelines from the World Health Organization indicate that the SARS-CoV-2 coronavirus, which results in the novel coronavirus disease (COVID-19), is transmitted through respiratory droplets or by contact. Contact transmission occurs when contaminated hands touch the mucous membrane of the mouth, nose, or eyes so hands hygiene is extremely important to prevent the spread of the SARS-CoV-2 as well as of other pathogens. The vast proliferation of wearable devices, such as smartwatches, containing acceleration, rotation, magnetic field sensors, etc., together with the modern technologies of artificial intelligence, such as machine learning and more recently deep-learning, allow the development of accurate applications for recognition and classification of human activities such as: walking, climbing stairs, running, clapping, sitting, sleeping, etc. In this work, we evaluate the feasibility of a machine learning based system which, starting from inertial signals collected from wearable devices such as current smartwatches, recognizes when a subject is washing or rubbing its hands. Preliminary results, obtained over two different datasets, show a classification accuracy of about 95% and of about 94% for respectively deep and standard learning techniques.

INDEX TERMS COVID-19 prevention, handwashing recognition, machine learning, wearable sensors.

I. INTRODUCTION

The World Health Organization (WHO) indicates that hands hygiene is extremely important to prevent the transmission of bacteria and viruses by avoiding its transfer from contaminated surfaces to the mucous membrane of the mouth, nose, or eyes. Also with regard to COVID-19 disease, it is estimated that a non-negligible part of infections occurs due to contact, through our hands, with contaminated surfaces [1]. For these reasons, one of the most important measures which any person can put in place to prevent the transmission of harmful germs is to take care of hands hygiene. To ensure proper hands hygiene, WHO suggests that one should follow either handrub, using an alcohol-based formulation, or handwash with soap and water. The two suggested procedures, described in Figure 1, entail different steps with different duration.

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

In particular, water and soap handwashing comprises eleven steps and should last between 40 to 60 seconds while handrubbing only includes 8 steps with a duration of about 20 and 30 seconds. WHO also suggests that alcohol-based handrub should be used for routine decontamination of hands, while handwash with soap and water is recommended when hands are visibly soiled.

Despite the proven effectiveness of these two procedures, most ordinary people ignore or simply do not follow them due to their non-trivial implementation. This results in a significant number of persons that limits themselves to washing/rubbing their hands as they have always been accustomed to.

Wearable devices, such as modern smartwatches, are equipped with several sensors capable of continuously measuring characteristic parameters of our body movement. For instance, Wang *et al.* in 2020 have measured the accuracy of some wearable bracelets, equipped with accelerometers,

(a) hand washing

(b) hand rubbing

FIGURE 1. WHO suggested procedure to accomplish handwashing (a) and handrubbing (b).

gyroscopes, and electrodes for surface electromyography (sEMG), in identifying and monitoring the handwashing/handrubbing procedures suggested by WHO reaching an accuracy of over 96% in recognizing the hands activity [2]. Before them, several authors have demonstrated the effectiveness of wearable devices in the classification and identification of general human activities such as running, walking, climbing/descending stairs, jumping, and sitting [3]–[8].

Moreover, the availability of very large data sets in human activity recognition and in medicine in general, together with the recent developments in deep learning, have led to impressive results in achieving human (or even superhuman) performance. Furthermore, current machine learning techniques have given a further boost to these studies by significantly increasing their classification accuracy which, for general macro-activities, now reaches values up to 99% [9]–[12]. These extraordinary performances together with the availability of a large amount of sensitive data have led to the birth of new ethical issues that cannot be neglected [13].

When no WHO procedures are followed, the activity of washing one's hands is executed in a very personal way involving some totally arbitrary steps so that it can be defined as completely "unstructured". Moreover, unlike activities that involve the whole body, such as walking, running, climbing stairs, etc., handwashing involves micro-movements of highly specialized segments of the body (such as fingers), the activity of which is not taken for granted that it can be recognized starting from signal collected, for example, on the subject's wrist by means of a common smartwatch. This high level of signals indirection together with the strong inter-subject variability which characterizes the unstructured handwashing can prevent the traditional machine learning tools from accurately recognizing it. For this reason, in this

work, we focus on the recognition of unstructured handwashing/handrubbing with the aim to investigate the classification accuracy of an automatic smartwatch-based system capable to monitor hands hygiene in the greater part of common people.

In particular, in this article, we introduce an extensive experimental study aimed at evaluating the ability of an automatic machine learning based system to distinguish the handwashing and handrubbing gestures from the rest of the activities that each person performs every day without the use of invasive instruments but relying only on commonly used wearable devices such as commercial smartwatches. Experimental results conducted on two different human activity recognition datasets across different machine learning models provide evidence of the effectiveness of this tool, which could potentially enable automatic and continuous indirect monitoring of hands hygiene of users, in an attempt to reduce the diffusion of COVID-19 and other diseases due to pathogens transmissible through direct contact.

The paper is organized according to the following structure: in Section II we describe state-of-the-art approaches related to our work, according to the scientific literature; in Section III we report a background description of the proposed machine learning tools; in Section IV we illustrate the proposed method and the related design choices; in Section V we present the results of the experimental evaluation; in Section VI we report some conclusive remarks.

II. RELATED WORK

At present there are no scientifically validated devices or applications that are able to recognize the activity of washing/rubbing hands by means of wearable tools. A commercially available system called *SureWash*, produced by

GLANTA Ltd, is able to detect the hands' movements of the hospital staff, through the use of video cameras, in order to provide information about the correct execution of the procedure defined by WHO [14]. Detecting the handwashing activity by means of video data has been investigated by several authors, for instance, Zhong *et al.* in 2020 presented a multi-camera system that uses video analytics to recognize seven specific actions within the hand-hygiene period [15]. More recently, Yue *et al.* use a pre-trained machine learning tool for gesture recognition called YOLO v3 to successfully identify and detect the seven steps hand-washing method pre-scribed by several hospitals [16]. However, one of the main problems facing camera-based systems is privacy, as such systems inevitably require the installation of cameras in several rooms, and, second, it is non-ubiquitous (it is not possible to monitor self-washing/rubbing by means, for instance, of sanitizing gel).

An orthogonal approach focuses on recognizing the hand-washing activity by means of wearable inertial sensors. Here the relevant scientific contributions are reduced to a few units and most of them are based on multiple sensors with very high sensitivity and accuracy typical of scientific instrumentation [17]–[19]. In this case, even if no video data are acquired, some privacy issues should be taken into account when designing wearable applications. In fact, several authors have shown that accelerometer data can reveal the personal identity by the way each individual moves (i.e., reveal their gender, age, and other identity markers) [20], [21]. From the recognition capability point of view, these preliminary works show that the automatic classification of handwashing activity, through the use of inertial sensors (accelerometers and gyroscopes), is a feasible task but, on the other hand, they do not study the potential of commercial smartwatches in common use, nor the application of modern deep-learning techniques.

Only a few relevant works which make use of commercial smartwatches have been published from 2015 to today. The first one, presented by Moldol *et al.* in 2015, describes a handwash monitoring and reminder system which interacts with a Bluetooth-enabled soap dispenser to recognize the start of the washing procedure [22]. Thanks to these broadcast advertisements, the smartwatch can easily start processing accelerometer and gyroscope data in order to recognize each procedure step as described by WHO.

In 2020, Mondol *et al.*, and Banerjee *et al.* present two robust solutions to check the WHO compliance of the handwashing activity starting from IMU sensors signals [23], [24]. In particular they can easily identify if a component of the WHO recommendations is missing or if it is not executed as expected. More recently, also Samyoun *et al.* present a handwashing quality assessment system based on smart-watch [25]. Here the authors measure the handwashing quality in terms of likeness with the standard guidelines of WHO where the start and the end of the washing events are marked with the help of the voice interaction with the user facilitating the recognition of the activity [26]. In 2022 Wahl *et al.*,

present a detection system based on commercially available wearable device which can distinguish enacted compulsive handwashing from WHO routine [27].

Also different combined approaches have been proposed in the last years, for instance, Wu *et al.* in 2021 present a prototype for autonomous hand hygiene tracking combining different IoT technologies such as IMU, video cameras, and smart dispensers which can provide prompt feedback if the handwashing is not performed as suggested by WHO. Finally, in 2022, Fagert *et al.* introduce a new sensing modality for handwashing monitoring by measuring the handwashing activity-induced vibration responses of sink structures [28].

Although in recent years several studies concerning hand-washing recognition have been proposed, none of these cover all the issues addressed in the present work. In particular, the contributions of the article can be summarized as follows:

- we evaluate the feasibility of IMU sensors embedded on common smartwatches in recognizing the handwashing activity
- we focus on unstructured handwashing rather than trying to recognize WHO suggested procedure steps
- we try to distinguish between handwashing and han-drubbing
- we compare traditional machine learning techniques with the most modern deep learning.

III. BACKGROUND

In this section, we report some background information about the machine learning tools investigated with the proposed method. In particular two *standard* machine learning tools and two *deep* learning tools have been tested. For what concerns standard learning, we evaluated Support Vector Machines (SVM) and Ensemble subspace with k-nearest neighbors (ES-KNN), while, in the deep-learning domain we have considered a Convolutional Neural Network (CNN) and a Long short-term Memory network (LSTM).

A. ENSEMBLE SUBSPACE WITH K-NEAREST NEIGHBORS (ES-KNN)

The k-nearest neighbors (KNN) is one of the most simple and easy to implement supervised machine learning algorithms that can be used in regression and classification problems. It assigns an unknown observation to the class most common among its *k* nearest neighbors observations, as measured by a distance metric, in the training data [29], [30]

Despite its simplicity, KNN gives competitive results and in some cases even outperforms other complex learning algorithms. However, one of the common problems which affect KNN is due to the possible presence of non-informative features in the data which can increase miss-classification errors. This is more likely in the case of high-dimensional data.

To improve KNN classification performances, ensemble techniques have been proposed in the literature. In general, the ensemble method entails the process of creating multiple

models and combining them (for instance by averaging) to produce the desired output, as opposed to creating just one model. Several studies show that, frequently, an ensemble of models performs better than any individual model, because the various errors of the models average out [31].

One way to generate an ensemble in machine learning is to train the classifiers on different sets of data, obtained by creating several subsets from the original training set. This technique, which is commonly called *Ensemble subspace* has been widely explored by several authors among which the contributions of [32] and [33], which are known respectively as *bagging* and *boosting* subspace ensemble, certainly stand out.

In this work, we focus on a particular class of ensemble subspace tools applied to KNN algorithms which are called Ensemble Random Subspace KNN (ERS-KNN). According to this technique, the features are randomly sampled, with replacement, for each learner forcing it to not over-focus on features that appear highly predictive/descriptive in the training set, but which can fail in unknown data [34], [35].

B. SUPPORT VECTOR MACHINES (SVM)
SVM is another class of supervised learning models traditionally used for regression and classification problems with a reduced number of samples.

An SVM model represents the input data as points in space, in such a way that the data belonging to the different classes are separated by a margin as large as possible. The new data are then mapped in the same space and the prediction of the category to which they belong is made on the basis of the side on which it falls. From the practical point of view, an SVM defines a hyperplane that best divides the dataset into the desired classes.

Moreover, in addition to a simple linear classification, it is possible to make use of the SVM to effectively carry out nonlinear classifications using nonlinear kernel methods which implicitly map input data in a multi-dimensional feature space [36].

C. CONVOLUTIONAL NEURAL NETWORK (CNN)
A CNN is actually a kind of multi-layer neural network following a computer vision approach to make use of any spatial or temporal information in the data. The CNN, in fact, was inspired by the biological process that occurs in the animal visual cortex, where neurons handle responses only from separate regions of the visual field. In the same way, CNN makes use of convolving filters to handle local regions within the data. A CNN is mainly composed of an input layer, several convolutional layers, pooling layers, and fully connected layers. The input layer has the task of collecting data and of forwarding it to the subsequent layer. The convolutional layer represents the main core of a CNN as it contains several convolution filters, called kernels, which convolve with the input data. The operation of convolution automatically extracts useful features from the input data and reduces its dimension. Moreover, the pooling layer, also

called subsampling layer, is also inserted to further reduce the number of parameters and the resulting computational cost. It includes max-pooling and/or average-pooling operations which sample, respectively, the max and the average value from the input. Finally, one or more fully connected layers act as a traditional Perceptron network which takes as input the features that originated from the previous layer.

A CNN is traditionally built using several layer batteries and it is used in the deep-learning approach also thanks to its characteristic of eliminating the requirement of feature extraction and feature selection often at the cost of an increase in computational complexity and memory usage [37].

D. LONG SHORT-TERM MEMORY (LSTM)
Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. LSTM networks are mostly designed to recognize patterns inside sequences of data such as numerical time series. RNN and LSTM differ from classical artificial neural networks due to the fact that they have a temporal dimension and they can not only process single data points, such as images, but also entire sequences of data such as speech or video.

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. It also makes decisions about what to store, and when to allow reads, writes, and erasures, via gates that open and close [38].

LSTM networks are well-suited to classifying, processing, and making predictions based on time series data, and they have been used in many complex problems such as handwriting recognition, language modeling and translation, speech synthesis, audio analysis, and protein structure prediction, and many others [39].

IV. THE PROPOSED METHOD
In this work, we evaluate the suitability of four different supervised classification methods, namely SVM, ERS-KNN, CNN, and LSTM, for classifying handwashing and handrubbing activities starting from gyroscopic and accelerometer data sampled in real-life conditions by means of common smartwatches.

A. EXPERIMENTAL PROTOCOL AND DATA GATHERING
The classification accuracy of the proposed machine learning models has been evaluated on top of two different datasets which are an ad-hoc collected dataset and the Daily Living Activities (DLA) dataset [40], [41].

The need for ad-hoc collecting data is due to the fact that to date there is no publicly available dataset for handwashing and handrubbing quality assessment through accelerometers and gyroscopic signals. For this reason, we collected sensors data from a wearable Inertial Measurement Unit (IMU) positioned on the wrist of the dominant hand of four participants during real-life activities. In particular, each subject was

asked to annotate the start and the end of each handwashing or handrubbing activity performed during the day. Each subject was wearing the IMU sensor for several hours during different days leading to a total of about 40 hours of recording containing about 1 hour and 40 minutes of total time spent washing hands and about 2 hours and 10 minutes of time spent in rubbing. The wearable device was programmed to sample its triaxial accelerometer and gyroscope at a frequency of 100 Hz and to store the collected data on the internal SD card. In order to remove sensors bias, the device was calibrated once at the start of the study by placing it on a stable surface and the accelerometers and gyroscopic measurements were recorded for 30 seconds.

Notice that the subjects were not instructed on how to wash or rub their hands leaving them completely free to use their usual way so to collect data about the unstructured way people normally use to wash their hands. Table 1 shows the average duration, together with the standard deviation, of each activity performed by the four subjects.

TABLE 1. Recorded activities duration in seconds.

subject	handwashing	handrubbing
0	$66.68s \pm 18.69s$	$23.66s \pm 6.26s$
1	$31.92s \pm 8.97s$	$26.09s \pm 3.67s$
2	$39.47s \pm 8.52s$	$19.18s \pm 4.29s$
3	$30.54s \pm 6.17s$	$25.44s \pm 8.59s$
avg	**50.92 ± 22.29**	**23.59 ± 7.33**

As we collect data in an unstructured way, the average duration and repeatability of each activity significantly depend on the subject. If this can represent an advantage in recognizing a particular subject, since her/his way of washing the hands could represent a kind of fingerprint, it could also represent a problem by reducing the ability to generalize the true activity recognition.

To collect data about daily activities we use a Shimmer3 IMU unit equipped with triaxial accelerometers and gyroscope [42]. This unit is a reference prototype designed for wearable applications frequently used in activity monitoring and sports science and it is representative of the IMU family that currently equips commercial smartwatches.

The internal accelerometer is a wide range sensor sampled at 14 bits which can be configured to operate in a full-scale range from ± 2.0 g up to ± 16.0 g with a resulting sensitivity from 1 mg/LSB in the ± 2.0 g range up to 12 mg/LSB in the ± 16.0 g range typical of a sensor which most smartwatches are equipped with.

Finally, in order to further evaluate the ability to generalize of the proposed classification models we conduct a set of experiments starting from the DLA dataset which is one of the few public datasets available that contains handwashing data sampled through inertial sensors. This dataset is based on samples recorded from different parts of the body and using different wearable sensors. In particular, data are recorded on the wrist, hip, and ankle while 8 healthy volunteers, aged

between 23-37, perform 17 different daily-life activities but, since our purpose is to evaluate the ability to classify handwashing through a smartwatch, in our experiments, we have used only the data collected on the person's wrist. For the same reason, the activities that involve the whole body, such as walking, upstairs, downstairs, etc, have been removed so that only the "hands activities" such as handwriting, handwashing, sweeping, etc., have been used for our experiments.

Notice that, this dataset does not contain data sampled during handrubbing and, moreover, the sensor placed on the person's wrist was not equipped with a gyroscope. Thanks also to these differences, this experiment can be very useful in demonstrating the generalization capabilities of the proposed models and in evaluating the usefulness of the signals coming from the gyroscope in handwashing recognition.

B. SIGNAL WINDOWING

The recorded tracks, composed of six distinct signals (i.e. 3 accelerometer and 3 gyroscope waveforms) for the ad-hoc dataset, and of only three accelerometer signals in the case of DLA dataset, have been divided into time windows and each of these has been considered as a sample to be used to train and test the classifiers. Furthermore, each sample has been labeled using the annotations provided by each subject in accordance with the following three categories: *washing*, *rubbing*, and *other*, for the first dataset, and with the following ten *handwriting, handwashing, facewashing, teethbrush, sweeping, vacuuming, eating, dusting, rubbing,* and *other* in the case of DLA dataset.

Obviously, deciding the size of the time window is a non-trivial task because it can influence the performance of classification models in different ways. In fact, it must be large enough to capture the "fingerprint" of the particular activity that we want to recognize, but it must not be too large to include consecutive activities. For what concerns human activity recognition (HAR), different window lengths have been used in the literature: starting from 1s up to 30s [9], [11], [12]. In particular, for what concerns the handwashing recognition both [19] and [25] use a very tiny time window (only 0.06 seconds), with 70% overlap between subsequent windows, due to the fact that they aim at recognizing each step of the structured handwashing procedure. In 2015, both [17] and [22] use a larger window respectively of 0.5 and 1 second while [2] found that a window with 0.2 seconds of amplitude and 75% overlap gave the best classification accuracy. To better highlight the impact of the window length on the overall classification performance, we present in this work an extensive sensitivity analysis of the classifiers with respect to this parameter.

Notice that, due to the proposed gathering protocol, which plans to continuously record sensors data during real-life activity, the number of samples containing non washing/rubbing events is much greater than that which contains them. For this reason, the samples labeled as *other* have been randomly undersampled in order to rebalance the occurrence of each class.

C. THE CLASSIFIERS

As machine learning classifier models we used multi-class SVM and ERS-KNN, CNN and LSTM (for neural networks). For what concerns the SVM tool, a cubic polynomial kernel has been choosen for performance reasons [36]. We also considered other kernels (i.e. linear, quadratic, or Gaussian functions), however, these did not reach the performances of the cubic kernel.

In order to train and test the classification accuracy of standard (i.e. not based on neural network approaches) machine learning tools, the input signal needs to be processed to extract synthetic features. In particular, in this work, for each window three sets of descriptors have been computed. The first set, hereafter referred to as *Base*, contains basic statistical descriptors aimed at capturing data tendency and variability. These are the following classical descriptors: *i)* average; *ii)* maximum value; *iii)* standard deviation; *iv)* median value. The second set contains the so called *Hjorth* parameters which are: i) *activity*; ii) *mobility*; iii) *complexity*. Finally, the last set is built with *Kurtosis* and *Skewness* parameters aimed at capturing the *Shape* of the data.

While the *Base* set easily describes the sample tendency, *Hjorth* parameters can capture the main characteristics of the signal in the frequency domain. In fact, *Hjorth* activity represents the power of the signal, the mobility its mean frequency, and the complexity measures its change in frequency [43].

Kurtosis and Skewness are used to describe, respectively, the degree of dispersion and symmetry of the data. In particular, Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution while Skewness measures how much data differ from a completely symmetrical distribution [44].

The architecture of the CNN and LSTM networks are presented, respectively in Figure 2.(a) and Figure 2.(b). In the case of deep learning approaches no feature extraction is needed and the samples of the signals, that make up the time window, can be directly used as input for the classification tool. In the case of CNN, which is commonly applied to analyzing visual imagery and which has been designed to work properly with bidimensional data, a preprocessing step has been added to represent time series data by means of visual cues. This possibility has recently attracted widespread attention so that in literature we can count several strategies aimed at re-coding time series into images to enable computer vision techniques and to perform classification [45]–[47]. In this paper, in particular, we investigate the method proposed by [45] which encodes time series as images called Gramian Angular Summation/Difference Field (GASF/GADF). This method represents time series in a polar coordinate system instead of the typical Cartesian coordinates with the advantage of preserving spatial and temporal relations. Because this method leads to the production of two distinct images: one for the Gramian Angular Summation (GASF) and one for the Gramian Angular Difference (GADF), in the case of the

ad-hoc dataset, we obtain 12 images (six from the accelerometer and six from the gyroscopic data), which reduce to only six for DLA dataset.

As a consequence, the CNN model takes in input a 12-channel or a 6-channel square image, depending on the dataset used, reconstructed starting from these data, whose height and width depend on the chosen window processing size (*WS*). The image is then convolved by four subsequent convolutional layers with decreasing size and numbers of filters. Furthermore, each convolutional layer is followed by a batch normalization layer and by a rectified linear activation function (*ReLu*). Batch normalization is used to standardize the input before forwarding it to the following layer and it has the effect of stabilizing the learning process and reducing the number of training epochs required to train the network [48]. The results of the ReLu layer are then processed by a subsequent pooling layer which selects the most activated features (*max pooling*). At the end of the convolutional structures, three fully connected layers, with decreasing number of neurons, have been added. The output of the last layer is then processed by a softmax function which assigns to each class a probability proportional to the output signal.

The LSTM network, on the other hand, receives in input six (three for DLA) sequences extracted from the original time series the length of which is the size of the window processing (*WS*). The input is then processed by three subsequent LSTM layers with decreasing number of hidden units. The output of the last LSTM layer is then forwarded to three fully connected layers as in the case of CNN.

D. CLASSIFICATION PERFORMANCE METRICS

For the proposed classifiers we calculate several performance metrics, together with the standard deviations, during a k-fold cross-validation test with k = 5. Dealing with multi-class classifiers, entails the evaluation of the following quantities for each of the N classes ($i \in [1 \cdots N]$ is an index that identifies a specific class): TP_i, the number of true positives predicted for class i; TN_i, the number of true negatives predicted for class i; FP_i, the number of false positives predicted for class i; FN_i, the number of false negatives predicted for class i.

Subsequently, these indicators have been used to compute the following metrics (corresponding to the so called *macro-averaging* measures) [49]:

$$Precision = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$Recall = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{2}$$

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{4}$$

(a) CNN architecture
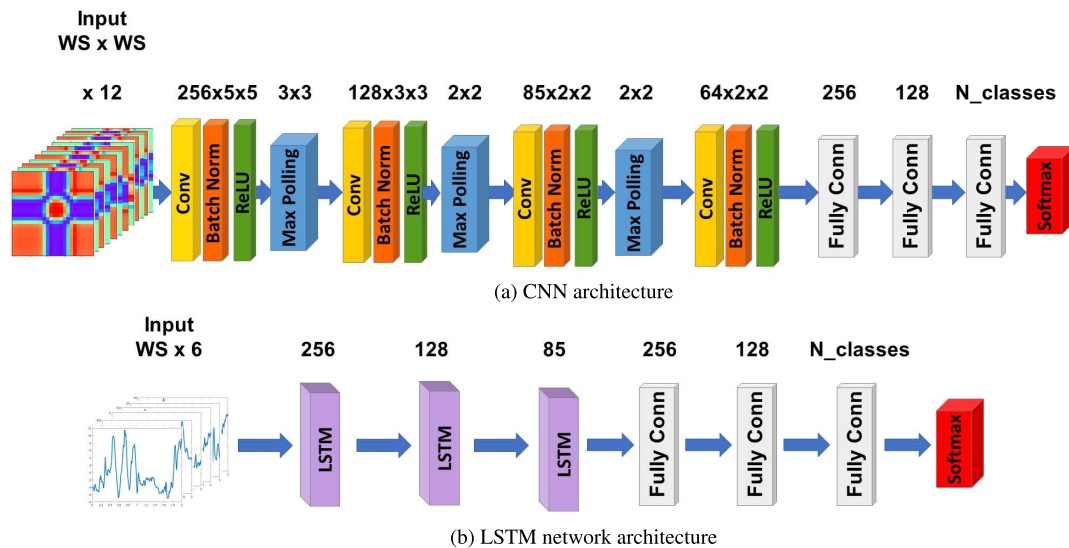

(b) LSTM network architecture

FIGURE 2. The architecture of the proposed CNN (a) and LSTM (b) networks.

V. RESULTS AND DISCUSSION

In this section, we report and discuss the obtained results. First of all, we show the best classification metrics calculated with the four machine learning tools described in section IV-C on the ad-hoc built dataset. Then, for each classifier, we report the sensitivity analysis with respect to the processing window length and the feature selection results for SVM and ERS-KNN.

Finally, in the last part of the section, we show the results obtained while running two sets of experiments starting from the publicly available DLA dataset.

A. CLASSIFICATION RESULTS

In this section, the results related to the classification of the human activity and of the subject identity obtained on the ad-hoc built dataset are reported.

1) HUMAN ACTIVITY CLASSIFICATION

Table 2 reports the best value of the classification metrics obtained when using the four proposed models. These values refer to the higher results obtained for each model when varying the window processing size and, for the standard learning tools, also the number of selected features. Each value is reported as the average value calculated during the 5-fold cross-validation test. For each metric, the highest value

TABLE 2. Best activity classification results obtained with the proposed models.

	SVM	ERS-KNN	LSTM	CNN
Accuracy	0.942	0.946	**0.947**	0.909
Precision	0.936	**0.941**	0.911	0.898
Recall	**0.934**	0.932	0.908	0.917
F1-score	0.935	**0.936**	0.910	0.908

obtained ever is highlighted in bold. For instance, the SVM classifier obtains the highest Recall value (about 0.934) while the ERS-KNN shows the highest Precision and F1-score (respectively 0.941 and 0.936). LSTM, on the other hand, reaches the best accuracy value of about 0.947. This suggests that the classification of the handwashing/handrubbing activities using signals gathered from a common smartwatch is a well feasible task which can be accomplished both with standard or deep learning techniques. Notice that, the best results reported here have been obtained with the following size of the processing window: SVM = 12s; ERS-KNN = 8s; LSTM = 2s; CNN = 6s. Moreover, in the case of SVM and ERS-KNN tools, all the proposed features have been used.

Figure 3 reports the average confusion matrices calculated on top of the results obtained during the 5-fold cross-validation tests. All four models used show a great ability to correctly classify the *other* activity. For instance, the LSTM network reaches the higher value of about 97.2%. Furthermore, also the *washing* activity has been correctly classified reaching the higher values of about 95% using standard tools while deep learning models do not exceed 91%. The *rubbing* activity, on the other hand, appears to be the most difficult to classify with a lower value of about 82% obtained with the CNN model. Here, in fact, the *rubbing* class has been misclassified as *other* about 16% of the time. This is probably due to the fact that handrubbing is a less dynamic activity with respect to hand washing which therefore produces fewer accelerations and rotations of the wrist. Furthermore, the use of running water for washing could introduce vibrations that are more easily identifiable by the classifiers.

2) SUBJECT CLASSIFICATION

The second set of classification experiments has been carried out aimed at identifying the person washing or rubbing their
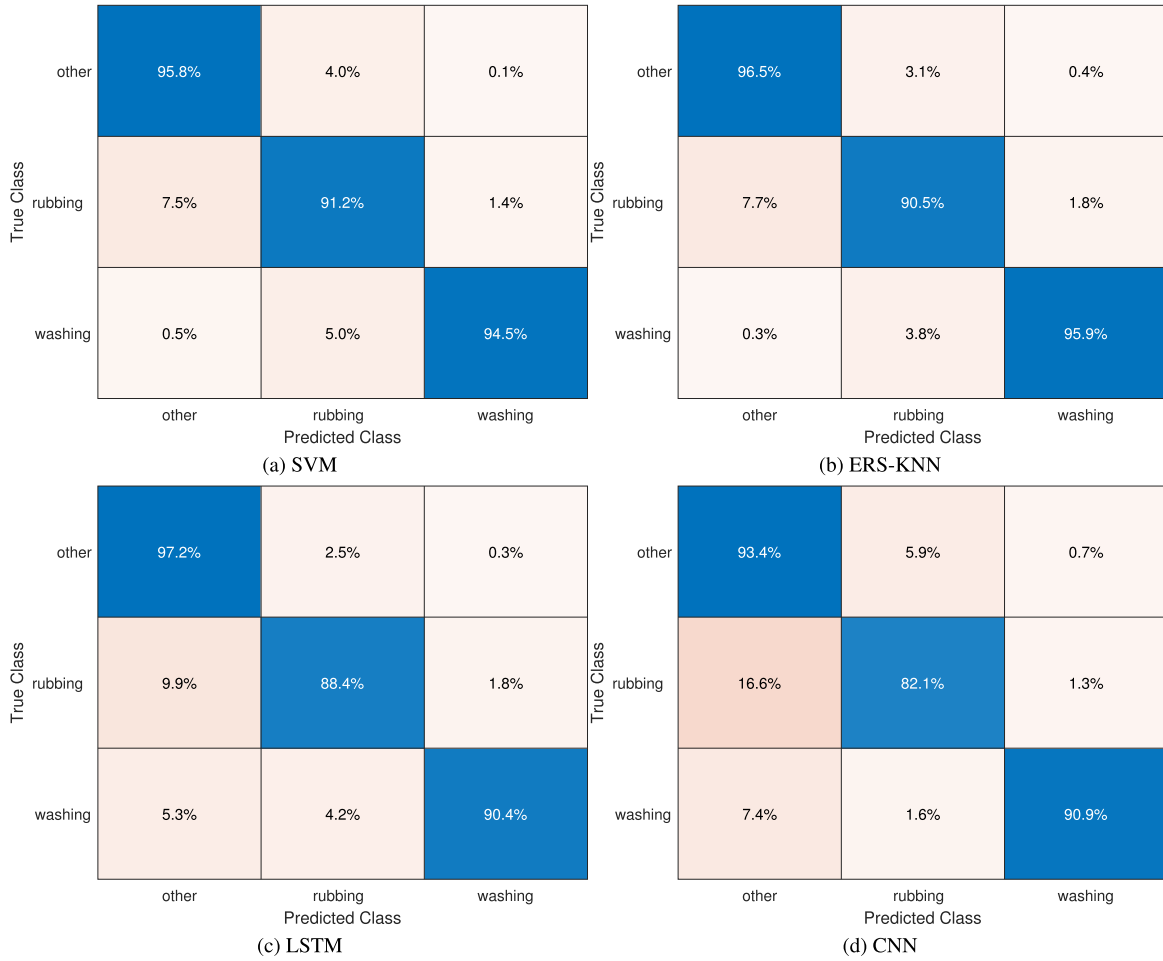
FIGURE 3. Average confusion matrices calculated on the 5-fold cross-validation test.

hands instead of the performed activity. For this purpose, each sample related to the *other* activity has been removed from the database while *washing* and *rubbing* samples have been merged into a single class to which a label containing a unique person identifier has been added.

TABLE 3. Best subject classification results obtained with the proposed models.

	SVM	ERS-KNN	LSTM	CNN
Accuracy	**0.991**	0.988	0.966	0.958
Precision	**0.990**	0.985	0.959	0.945
Recall	**0.989**	0.986	0.956	0.952
F1-score	**0.990**	0.985	0.957	0.948

Table 3 shows the best value of the classification metrics obtained with the four models. Also in this case, the best results have been obtained with the following size of the processing window: SVM = 12s; ERS-KNN = 8s; LSTM = 2s; CNN = 6s. In the case of standard tools, all the proposed features have been used. As expected, recognizing the person who is washing/rubbing the hands is a much easier task thanks to the fact that the ad-hoc dataset contains data collected in an

unstructured way where each subject is free to wash its hands as he/she is used to. Our results, with the higher accuracy of about 0.99 obtained with the SVM classifier, in fact, suggest that the hand washing/rubbing activity can represent a kind of subject fingerprint. Another interesting result from this experiment concerns the fact that SVM and ERS-KNN seem to exceed the deep learning methods by almost 5 percentage points.

B. MEMORY FOOTPRINT AND TIMING

In order to indirectly evaluate the complexity of the proposed machine learning models, during each experiment, we measured the timing performances and the memory footprint of each classifier. In particular, each classifier has been implemented on Matlab2021a® platform by means of Machine Learning tools and get executed on Intel® Core i9 desktop PC equipped with 32GB of RAM and with an NVIDIA® Quadro® RTX™ 4000 GPU. Both training and inference phases have been conducted by setting GPU as the execution environment in order to fully exploit the CUDA® parallelism of the video adapter.

TABLE 4. Timing performance and memory footprint.

	SVM	ERS-KNN	LSTM	CNN
Training time (s)	7.191	**5.803**	472.069	299.551
Inference time (ms)	**0.016**	0.331	1.749	0.505
Memory footprint (MB)	**4.168**	41.996	5.569	6.389

Table 4 shows the timing performance and the memory footprint of each model with highlighted the best values. As expected the training phase of deep learning models is the most time expensive phase with a maximum value of about 470 seconds measured in the LSTM network. On the other hand, the training phase of the ERS-KNN results to be the fastest with only about 5.8 seconds. From the inference time point of view, the SVM significantly exceeds other models performances as it needs only about 160 μ seconds to infer a label. In particular, it results in an improvement in performance in the order of about 20x, 100x, and 40x with respect to, respectively, ERS-KNN, LSTM, and CNN. Notice that, the reported inference time for SVM and ERS-KNN also includes the feature extraction time. If in the cloud-based approaches the computational cost of the training phase can be easily overcome by resorting to powerful GPU/TPU, the inference phase, on the other hand, should be done directly on the wearable device to explore several appealing benefits. For instance, executing the inference phase on the edge node avoids latency issues due to communications to and from the cloud, it enables higher levels of privacy and security by keeping most of the data on remote devices, and, finally, it can improve energy efficiency by trading off computation and communication energy requirements. From this point of view, Table 4 shows that SVM, thanks to its low memory footprint and low inference time, can be the most suitable model for a real-time wearable application.

C. SENSITIVITY TO THE WINDOW LENGTH

The size of the processing window influences the performance of the classification models in several ways. In this section, the results of the in-depth analysis of this dependence are reported. In particular, Figure 4 plots the classification metrics obtained by the four classifiers when varying the size of the processing window. Notice that, in each experiment, we use a processing window with 75% of overlap which leads to a total number of samples ranging from more than 200,000 to about 17,000 when the window increases from 2 to 20 seconds. Each point represents the average value together with the standard deviation calculated over a 5-fold cross-validation test. Both SVM and ERS-KNN (Figure 4.(a) and Figure 4.(b)) show an almost flat trend of the measured accuracy even if at some point the other metrics (precision, recall, and f1-score) begin to deteriorate as the window size increases. In particular, the SVM classifier increases its performance until when using a window of about 12 seconds. Further increasing the size of the window leads to an average

decrease in Precision, Recall, and F1-score and to more unstable results (higher standard deviations).

Similarly, the performances of ERS-KNN increase until a window size of about 8 seconds beyond which they markedly decrease together with results stability.

An opposite trend is found, however, regarding the results obtained by the two deep learning classifiers (Figure 4.(c) and Figure 4.(d)). In this case, in fact, the four performance metrics show an almost monotonous decreasing trend for increasing values of the window size. Moreover, for the CNN classifier, we found a local maximum at a window size of about 6 seconds.

D. NON-PARAMETRIC SIGNIFICANCE TESTS

In order to statistically assess whether the accuracies of the four classification models are different, three variations of the McNemar test have been performed: i) asymptotic test; ii) exact-conditional test; iii) mid-p-value test [50]. These tests compare two classifiers by analyzing their predicted labels against the true labels and then detect whether the difference between the misclassification rates is statistically significant.

Table 5 reports the results of the three tests conducted for each couple of classifiers. The possible combinations of the four classifiers produce a symmetric matrix which, for readability, is reported only in its upper part. In the same way, the comparison between the classifiers and itself has not been carried out (the matrix diagonal) for the obviousness of the results.

For each test, the logical value h is reported which represents the test decision when testing the null hypothesis that the two classifiers have equal accuracy for predicting the true class. So, a false value indicates that the null hypothesis is not rejected with a confidence level of 95% ($p < 0.05$). Moreover, also the p value is reported which represents how strong is the evidence to reject or not the null hypothesis.

For instance, when comparing ERS-KNN with SVM, the three variants of the McNemar test agree not to reject the null hypothesis while, for both ERS-KNN Vs LSTM and ERS-KNN Vs CNN, the null hypothesis needs to be rejected so that the accuracies of the two classification models can not be considered equivalent. Furthermore, in these cases, the p-value for each test is close to zero, which indicates strong evidence to reject the null hypothesis that the two classifiers have equal predictive accuracies.

Also the comparison between SVM and LSTM, SVM and CNN, and LSTM and CNN, lead to strong rejections of the null hypothesis with p-values very close to zero so that, in this scenario, only standard machine learning tools, namely SVM and ERS-KNN, can be considered statistically equivalent from the classification performance point of view.

E. FEATURE SELECTION RESULTS

In order to evaluate the relative influence of the proposed features on the classification performances, we use the forward feature selection method [51]. Forward feature selection is
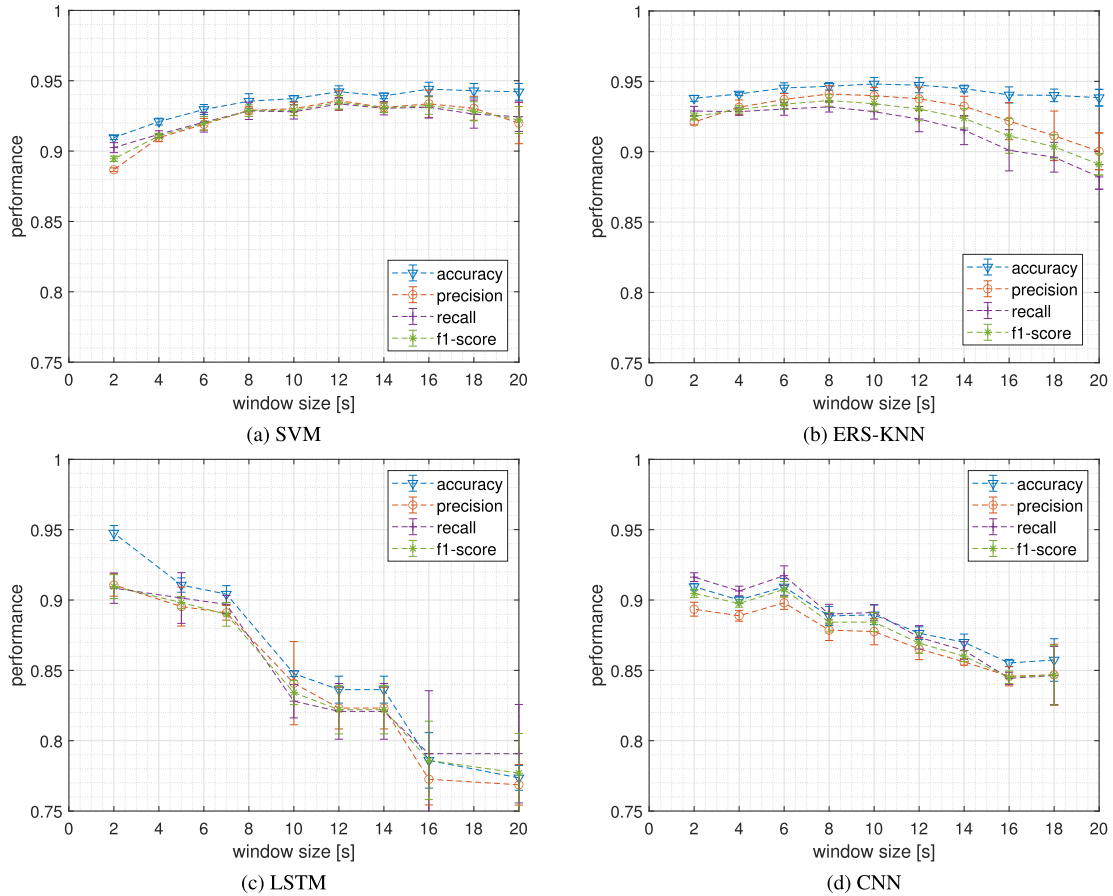
(a) SVM

(b) ERS-KNN

(c) LSTM

(d) CNN

FIGURE 4. Performances of the proposed classifiers when varying the size of the processing window.

based on an objective function (e.g. the accuracy) which is used as a criterion to evaluate the impact of adding a feature from a candidate subset, starting from an empty set until adding other features doesn't induce any improvement in the objective function. We applied this strategy to highlight how the proposed features contribute to the overall performance of the two standard classifiers. In particular, each group of features, namely *Base (B)*, *Hjorth (H)*, and *Shape (S)* has been treated as an atomic unit that can be added or removed as a whole. First of all, we tested each classifier using only one of the three groups, and then we added the other groups to explore all possible combinations.

Table 6 shows the activities classification performances, together with its standard deviations, of the SVM when varying the adopted features. For each performance metric, the maximum value achieved has been highlighted in bold. All metrics showed a monotone increasing trend when consecutively adding one of three groups of features reaching the higher performances when all the proposed features are used together (*Base+Hjorth+Shape*). This suggests that all features provide original information content useful for the classification process. Furthermore, the *Hjorth* group seems to contain the most informative group of features producing the highest classification

performance with respect to the other groups when tested alone.

The same experiment conducted with the ESR-KNN classifier produces comparable results, reported in Table 7, with the only difference that, in this case, the measured performances are slightly higher.

Notice that, for both classifiers, these results suggest that a good trade-off between classification performances and real-time computation complexity can be represented by the design choice of calculating only *Base+Hjorth* giving up only about 0.2% of classification performance decrease. Moreover, if for the purpose of a particular real-time application, an accuracy of about 92% could be considered acceptable, it even would be sufficient to calculate the Hjorth features, saving many computational resources and power.

F. CLASSIFICATION RESULTS ON DLA DATASET
The models described in the previous sections have been trained and tested on the publicly available DLA dataset. A first set of experiments has been made to classify the following ten activities: *handwriting, handwashing, facewashing, teethbrush, sweeping, vacuuming, eating, dusting, rubbing, other*, and, a second set has been made in order to recognize the 8 subjects of the dataset.

TABLE 5. Results of three variations of the McNemar test.

		SVM		LSTM		CNN	
		h	p	h	p	h	p
ERS-KNN	asymptotic	**false**	0.65	true	4.14×10^{-25}	true	2.04×10^{-21}
	mid-p	**false**	0.65	true	1.33×10^{-25}	true	3.12×10^{-27}
	exact	**false**	0.69	true	1.82×10^{-32}	true	1.12×10^{-22}
SVM	asymptotic			true	3.94×10^{-05}	true	4.53×10^{-22}
	mid-p			true	3.87×10^{-05}	true	1.27×10^{-18}
	exact			true	4.46×10^{-05}	true	3.43×10^{-22}
LSTM	asymptotic					true	1.23×10^{-25}
	mid-p					true	5.13×10^{-25}
	exact					true	2.26×10^{-24}

TABLE 6. Results of the forward feature selection method applied to SVM classifier.

	Accuracy	Precision	Recall	F1-s
B	0.901 ± 0.007	0.892 ± 0.008	0.891 ± 0.014	0.891 ± 0.006
H	0.923 ± 0.003	0.912 ± 0.005	0.913 ± 0.006	0.891 ± 0.005
S	0.339 ± 0.098	0.500 ± 0.069	0.480 ± 0.067	0.490 ± 0.067
B+H	0.941 ± 0.006	0.932 ± 0.008	0.932 ± 0.006	0.932 ± 0.006
B+S	0.916 ± 0.004	0.910 ± 0.006	0.906 ± 0.010	0.908 ± 0.006
H+S	0.926 ± 0.011	0.917 ± 0.009	0.921 ± 0.067	0.919 ± 0.010
B+H+S	$\mathbf{0.942 \pm 0.008}$	$\mathbf{0.936 \pm 0.008}$	$\mathbf{0.934 \pm 0.006}$	$\mathbf{0.935 \pm 0.007}$

TABLE 7. Results of the forward feature selection method applied to ERS-KNN classifier.

	Accuracy	Precision	Recall	F1-score
B	0.904 ± 0.007	0.899 ± 0.003	0.873 ± 0.009	0.886 ± 0.002
H	0.908 ± 0.004	0.908 ± 0.005	0.889 ± 0.006	0.886 ± 0.005
S	0.719 ± 0.007	0.683 ± 0.010	0.674 ± 0.009	0.678 ± 0.010
B+H	0.944 ± 0.002	0.937 ± 0.003	0.929 ± 0.002	0.933 ± 0.002
B+S	0.908 ± 0.006	0.909 ± 0.006	0.886 ± 0.009	0.897 ± 0.007
H+S	0.922 ± 0.004	0.920 ± 0.006	0.902 ± 0.009	0.911 ± 0.006
B+H+S	$\mathbf{0.946 \pm 0.003}$	$\mathbf{0.941 \pm 0.003}$	$\mathbf{0.932 \pm 0.003}$	$\mathbf{0.936 \pm 0.003}$

1) HUMAN ACTIVITY CLASSIFICATION

Table 8 reports the classification results obtained with the models configuration and parameters used in section V-A1. Each value is reported as the average value calculated during the 5-fold cross-validation test. For each metric, the highest value obtained ever is highlighted in bold.

TABLE 8. Best activity classification results obtained with the proposed models on DLA dataset.

	SVM	ERS-KNN	LSTM	CNN
Accuracy	0.891	0.914	0.866	**0.923**
Precision	0.907	0.931	0.882	**0.939**
Recall	0.867	0.914	0.874	**0.925**
F1-score	0.886	0.922	0.878	**0.932**

In this experiment, the CNN outperforms other models in each of the calculated metrics. Regarding the ERS-KNN performance, the gap is about 1% or 2% (depending on the considered metric) while, it increases up to 6% for some metrics calculated for SVM and LSTM models Despite this, the performance of ERS-KNN and CNN remains well above

91% in each computed metric demonstrating a high capability in classifying wrist human activities using signals from wearable IMU devices.

Compared to the performances of the experiment conducted on the ad-hoc dataset, the average values suffered a decrease of about 1% or 2% for ERS-KNN and CNN and of about 7% and 8% for SVM and LSTM. A generalized decrease in classification performance is certainly due to the greater number of classes used in this second dataset, which can lead to a greater dispersion of the results. Furthermore, the lack of gyroscopic signals may have had a significant impact on classification capabilities, especially for SVM and LSTM which show the lowest values.

Figure 5 reports the average confusion matrices calculated on top of the results obtained during the 5-fold cross-validation tests. ERS-KNN and CNN demonstrate a great ability to correctly classify the activity of interest (i.e. the *handwashing*). For instance, CNN correctly recognizes it over 95% of the time while the ERS-KNN about 94% of the time. On the other hand, for the SVM and LSTM models, this rate decreases up to 88% and 87% respectively. Moreover, the
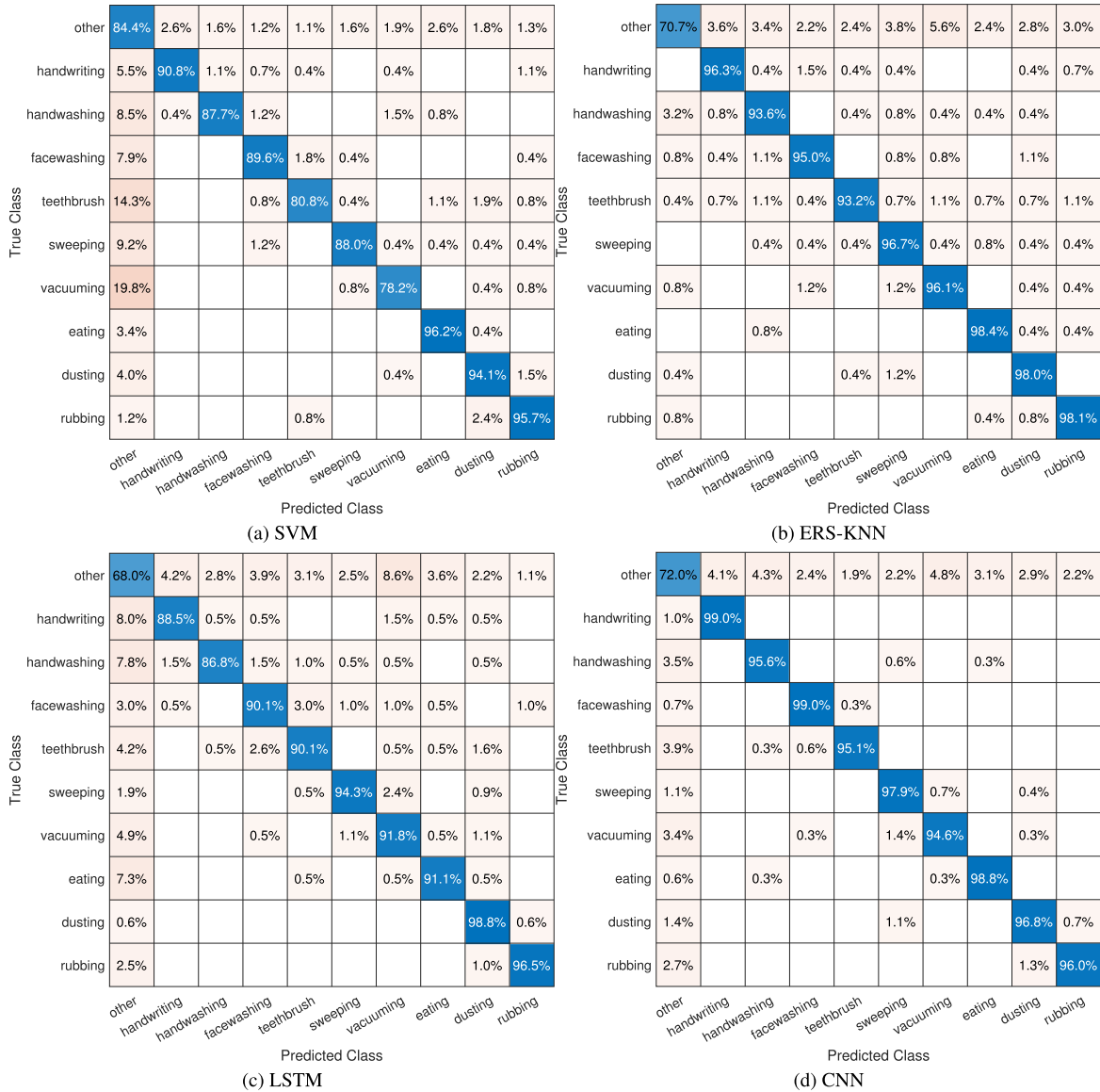
(a) SVM

True Class \ Predicted	other	handwriting	handwashing	facewashing	teethbrush	sweeping	vacuuming	eating	dusting	rubbing
other	84.4%	2.6%	1.6%	1.2%	1.1%	1.6%	1.9%	2.6%	1.8%	1.3%
handwriting	5.5%	90.8%	1.1%	0.7%	0.4%		0.4%			1.1%
handwashing	8.5%	0.4%	87.7%	1.2%			1.5%	0.8%		
facewashing	7.9%			89.6%	1.8%	0.4%				0.4%
teethbrush	14.3%			0.8%	80.8%	0.4%		1.1%	1.9%	0.8%
sweeping	9.2%			1.2%		88.0%	0.4%	0.4%	0.4%	0.4%
vacuuming	19.8%					0.8%	78.2%		0.4%	0.8%
eating	3.4%							96.2%	0.4%	
dusting	4.0%						0.4%		94.1%	1.5%
rubbing	1.2%			0.8%					2.4%	95.7%

(b) ERS-KNN

True Class \ Predicted	other	handwriting	handwashing	facewashing	teethbrush	sweeping	vacuuming	eating	dusting	rubbing
other	70.7%	3.6%	3.4%	2.2%	2.4%	3.8%	5.6%	2.4%	2.8%	3.0%
handwriting		96.3%	0.4%	1.5%	0.4%	0.4%			0.4%	0.7%
handwashing	3.2%	0.8%	93.6%		0.4%	0.8%	0.4%	0.4%	0.4%	
facewashing	0.8%	0.4%	1.1%	95.0%		0.8%	0.8%		1.1%	
teethbrush	0.4%	0.7%	1.1%	0.4%	93.2%	0.7%	1.1%	0.7%	0.7%	1.1%
sweeping			0.4%	0.4%	0.4%	96.7%	0.4%	0.8%	0.4%	0.4%
vacuuming	0.8%			1.2%		1.2%	96.1%		0.4%	0.4%
eating			0.8%					98.4%	0.4%	0.4%
dusting	0.4%				0.4%	1.2%			98.0%	
rubbing	0.8%							0.4%	0.8%	98.1%

(c) LSTM

True Class \ Predicted	other	handwriting	handwashing	facewashing	teethbrush	sweeping	vacuuming	eating	dusting	rubbing
other	68.0%	4.2%	2.8%	3.9%	3.1%	2.5%	8.6%	3.6%	2.2%	1.1%
handwriting	8.0%	88.5%	0.5%	0.5%			1.5%	0.5%	0.5%	
handwashing	7.8%	1.5%	86.8%	1.5%	1.0%	0.5%	0.5%		0.5%	
facewashing	3.0%	0.5%		90.1%	3.0%	1.0%	1.0%	0.5%		1.0%
teethbrush	4.2%		0.5%	2.6%	90.1%		0.5%	0.5%	1.6%	
sweeping	1.9%				0.5%	94.3%	2.4%		0.9%	
vacuuming	4.9%			0.5%		1.1%	91.8%	0.5%	1.1%	
eating	7.3%				0.5%		0.5%	91.1%	0.5%	
dusting	0.6%								98.8%	0.6%
rubbing	2.5%								1.0%	96.5%

(d) CNN

True Class \ Predicted	other	handwriting	handwashing	facewashing	teethbrush	sweeping	vacuuming	eating	dusting	rubbing
other	72.0%	4.1%	4.3%	2.4%	1.9%	2.2%	4.8%	3.1%	2.9%	2.2%
handwriting	1.0%	99.0%								
handwashing	3.5%		95.6%			0.6%		0.3%		
facewashing	0.7%			99.0%	0.3%					
teethbrush	3.9%		0.3%	0.6%	95.1%					
sweeping	1.1%					97.9%	0.7%		0.4%	
vacuuming	3.4%			0.3%		1.4%	94.6%		0.3%	
eating	0.6%		0.3%			0.3%		98.8%		
dusting	1.4%				1.1%				96.8%	0.7%
rubbing	2.7%								1.3%	96.0%

FIGURE 5. Average confusion matrices calculated on the 5-fold cross-validation test on DLA dataset.

remaining wrong classifications of the *handwashing* activity are due overall to its misclassification as *other* and they are not due to another wrist activity in particular, effectively representing false negatives.

2) SUBJECT CLASSIFICATION

As for the subject classification on ad-hoc dataset experiments, each sample related to the *other* activity has been removed, while the remaining classes have been merged into a single class to which a label containing a unique person identifier has been added.

Table 9 shows the value of the classification metrics obtained with the four models. Also for the DLA dataset, recognizing the person who is performing the wrist-based action is an easier task with respect to recognizing the activity.

TABLE 9. Best subject classification results obtained with the proposed models on DLA dataset.

	SVM	ERS-KNN	LSTM	CNN
Accuracy	0.946	**0.959**	0.902	0.941
Precision	0.949	**0.957**	0.902	0.940
Recall	0.942	**0.949**	0.903	0.943
F1-score	0.946	**0.959**	0.903	0.939

Moreover, also in this experiment, the standard classifiers (i.e. SVM and ERS-KNN) overcome the performance of the deep learning methods even if, with respect to CNN, the gap is appreciably reduced.

In general, despite this experiment is showing a high average classification accuracy of the proposed models (greater

than 95%), it also highlights an average reduction of performances while switching from the ad-hoc dataset to the DLA. Also in this case, in all probability, the loss of performance can be attributed to the higher number of subjects to be recognized and to the lack of the gyroscopic signals.

VI. CONCLUSION

Hands hygiene is extremely important in breaking the chain of pathogens transmission by contact. In fact, contaminated hands are a privileged way to get to the mucous membrane of the mouth, nose, or eyes. Also regarding the COVID-19, it is estimated that a non-negligible part of infections occurs due to contact, through our hands, with contaminated surfaces.

In this work, we proposed and evaluated four classification machine learning models to distinguish the unstructured handwashing/handrubbing gestures from the rest of the daily activities starting from commonly used wearable devices. The proposed models lay the foundations for the creation of a system that provides users with automatic and continuous indirect monitoring of hands hygiene in an attempt to reduce the contact transmission of pathogens including SARS-CoV-2 coronavirus.

The experimental results, obtained over two different datasets containing in total more than 50 hours of recording of daily activities registration performed by 12 different subjects, show that both standard and deep learning techniques can be considered a viable solution to the classification problem reaching, respectively, an average accuracy of about 94% and 95%. Furthermore, considering the design of a low-cost low-power wearable device as a possible target of this work, our results show that SVM, thanks to its low memory footprint and low inference time, could allow executing the inference phase on the edge without recurring to the cloud infrastructure. This could, in principle, avoid latency and energy issues due to communications to and from the cloud, leading to higher levels of privacy and security by keeping most of the data on remote devices. Finally, as a last noteworthy result, we also show that by making use of the gyroscopic signals coming from the IMU devices it is possible to increase the recognition capability of standard learning tools, such as SVM and ERS-KNN, achieving the performances of the most complex model without impairing the computational cost.

REFERENCES

[1] J. L. Santarpia, D. N. Rivera, V. L. Herrera, M. J. Morwitzer, H. M. Creager, G. W. Santarpia, K. K. Crown, D. M. Brett-Major, E. R. Schnaubelt, M. J. Broadhurst, J. V. Lawler, S. P. Reid, and J. J. Lowe, "Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care," *Sci. Rep.*, vol. 10, no. 1, pp. 1–8, Dec. 2020.

[2] C. Wang, Z. Sarsenbayeva, X. Chen, T. Dingler, J. Goncalves, and V. Kostakos, "Accurate measurement of handwash quality using sensor armbands: Instrument validation study," *JMIR mHealth uHealth*, vol. 8, no. 3, Mar. 2020, Art. no. e17001.

[3] M. Zhang and A. A. Sawchuk, "Human daily activity recognition with sparse representation using wearable sensors," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 3, pp. 553–560, May 2013.

[4] T. Sztyler and H. Stuckenschmidt, "On-body localization of wearable devices: An investigation of position-aware activity recognition," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2016, pp. 1–9.

[5] T. Sztyler, H. Stuckenschmidt, and W. Petrich, "Position-aware activity recognition with wearable devices," *Pervasive Mobile Comput.*, vol. 38, pp. 281–295, Jul. 2017.

[6] G. Bhat, R. Deb, V. V. Chaurasia, H. Shill, and U. Y. Ogras, "Online human activity recognition using low-power wearable devices," in *Proc. Int. Conf. Comput.-Aided Design*, Nov. 2018, pp. 1–8.

[7] L. Köping, K. Shirahama, and M. Grzegorzek, "A general framework for sensor-based human activity recognition," *Comput. Biol. Med.*, vol. 95, pp. 248–260, Apr. 2018.

[8] E. Lattanzi, M. Donati, and V. Freschi, "Exploring artificial neural networks efficiency in tiny wearable devices for human activity recognition," *Sensors*, vol. 22, no. 7, p. 2637, Mar. 2022.

[9] J. Cheng, O. Amft, and P. Lukowicz, "Active capacitive sensing: Exploring a new wearable sensing modality for activity recognition," in *Proc. Int. Conf. Pervasive Comput.* Cham, Switzerland: Springer, 2010, pp. 319–336.

[10] D. Singh, E. Merdivan, S. Hanke, J. Kropf, M. Geist, and A. Holzinger, "Convolutional and recurrent neural networks for activity recognition in smart environment," in *Towards Integrative Machine Learning and Knowledge Extraction*. Cham, Switzerland: Springer, 2017, pp. 194–205.

[11] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Futur. Gener. Comput. Syst.*, vol. 81, pp. 307–313, Aug. 2018.

[12] C. Hou, "A study on IMU-based human activity recognition using deep learning and traditional machine learning," in *Proc. 5th Int. Conf. Comput. Commun. Syst. (ICCCS)*, May 2020, pp. 225–234.

[13] H. Müller, M. T. Mayrhofer, E.-B. Van Veen, and A. Holzinger, "The ten commandments of ethical medical AI," *Computer*, vol. 54, no. 7, pp. 119–123, Jul. 2021.

[14] *Surewash: The Science of Hand Hygiene*, Glanta, Göteborg, Sweden, 2021. Accessed: Jul. 26, 2021.

[15] C. Zhong, A. R. Reibman, H. A. Mina, and A. J. Deering, "Multi-view hand-hygiene recognition for food safety," *J. Imag.*, vol. 6, no. 11, p. 120, Nov. 2020.

[16] Z. Yue, F. Wang, and Z. Liu, "An intelligent hand-washing monitoring platform based on gesture recognition technology," in *Proc. China Autom. Congr. (CAC)*, Oct. 2021, pp. 940–944.

[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] M. Bal and R. Abrishambaf, "A system for monitoring hand hygiene compliance based-on Internet-of-Things," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2017, pp. 1348–1353.

[19] 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.

[20] A. Jain and V. Kanhangad, "Gender classification in smartphones using gait information," *Expert Syst. Appl.*, vol. 93, pp. 257–266, Mar. 2018.

[21] T. Van Hamme, G. Garofalo, E. A. Rúa, D. Preuveneers, and W. Joosen, "A systematic comparison of age and gender prediction on IMU sensor-based gait traces," *Sensors*, vol. 19, no. 13, p. 2945, Jul. 2019.

[22] M. A. S. Mondol and J. A. Stankovic, "Harmony: A hand wash monitoring and reminder system using smart watches," in *Proc. 12th EAI Int. Conf. Mobile Ubiquitous Syst., Comput., Netw. Services*, 2015, pp. 11–20.

[23] M. A. S. 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.

[24] 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.*, Jun. 2020, pp. 34–36.

[25] S. Samyoun, S. S. Shubha, M. A. S. Mondol, and J. A. Stankovic, "IWash: A smartwatch handwashing quality assessment and reminder system with real-time feedback in the context of infectious disease," *Smart Health*, vol. 19, Mar. 2021, Art. no. 100171.

[26] F. Wu, T. Wu, D. C. Zarate, R. Morfuni, B. Kerley, J. Hinds, D. Taniar, M. Armstrong, and M. R. Yuce, "An autonomous hand hygiene tracking sensor system for prevention of hospital associated infections," *IEEE Sensors J.*, vol. 21, no. 13, pp. 14308–14319, Jul. 2021.

[27] K. Wahl, P. M. Scholl, S. Wirth, M. Miché, J. Häni, P. Schülin, and R. Lieb, "On the automatic detection of enacted compulsive hand washing using commercially available wearable devices," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105280.

[28] J. Fagert, A. Bonde, S. Srinidhi, S. Hamilton, P. Zhang, and H. Y. Noh, "Clean vibes: Hand washing monitoring using structural vibration sensing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 3, pp. 1–25, Jul. 2022.

[29] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis. Nonparametric discrimination: Consistency properties," *Int. Stat. Rev.*, vol. 57, no. 3, pp. 238–247, 1989.

[30] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statistician*, vol. 46, pp. 175–185, Aug. 1992.

[31] U. Naftaly, N. Intrator, and D. Horn, "Optimal ensemble averaging of neural networks," *Netw., Comput. Neural Syst.*, vol. 8, no. 3, p. 283, 1997.

[32] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[33] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Int. Conf. Mach. Learn.*, vol. 96, Jul. 1996, pp. 148–156.

[34] T. K. Ho, "Nearest neighbors in random subspaces," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. (SPR) Struct. Syntactic Pattern Recognit. (SSPR)*. Cham, Switzerland: Springer, 1998, pp. 640–648.

[35] S. Li, E. J. Harner, and D. A. Adjeroh, "Random KNN feature selection—A fast and stable alternative to random forests," *BMC Bioinf.*, vol. 12, no. 1, pp. 1–11, Dec. 2011.

[36] I. Steinwart and A. Christmann, *Support Vector Machines*. Cham, Switzerland: Springer, 2008.

[37] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] Y. Yu, X. Si, C. Hu, and Z. Jianxun, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.

[40] M. Leotta, A. Fasciglione, and A. Verri, "Daily living activity recognition using wearable devices: A features-rich dataset and a novel approach," in *Proc. Int. Conf. Pattern Recognit.*, A. D. Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham, Switzerland: Springer, 2021, pp. 171–187.

[41] M. Leotta, A. Fasciglione, and A. Verri, "Daily living activity recognition using wearable devices: A features-rich dataset and a novel approach," in *Pattern Recognition. ICPR International Workshops and Challenges* (Lecture Notes in Computer Science), vol. 12662. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-68790-8_15.

[42] *Shimmer3 IMU Unit*, Shimmer, San Francisco, CA, USA 2019. Accessed: Jul. 26, 2021.

[43] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, Sep. 1970.

[44] T.-H. Kim and H. White, "On more robust estimation of skewness and kurtosis," *Finance Res. Lett.*, vol. 1, no. 1, pp. 56–73, Mar. 2004.

[45] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, Jan. 2015, pp. 3939–3945.

[46] G. Baldini, G. Steri, R. Giuliani, and C. Gentile, "Imaging time series for Internet of Things radio frequency fingerprinting," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2017, pp. 1–6.

[47] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K.-R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Inf. Fusion*, vol. 53, pp. 80–87, Jan. 2020.

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[49] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[50] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional," *BMC Med. Res. Methodol.*, vol. 13, no. 1, p. 91, Dec. 2013.

[51] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

EMANUELE LATTANZI received the Laurea degree *(summa cum laude)*, in 2001, and the Ph.D. degree from the University of Urbino, Italy, in 2005. Since 2001, he has been with the Information Science and Technology Institute, University of Urbino. In 2003, he was with the Department of Computer Science and Engineering, Pennsylvania State University, as a Visiting Scholar with Prof. V. Narayanan. From 2008 to 2020, he was an Assistant Professor in computer engineering with the Department of Pure and Applied Sciences (DiSPeA), University of Urbino, where he is currently an Associate Professor in computer engineering. His research interests include wireless sensor networks, energy-aware routing algorithms, artificial intelligence, and the Internet of Things.

LORENZO CALISTI received the Laurea degree in applied computer science from the University of Urbino, Italy, in 2020, where he is currently pursuing the Graduate degree. His research interests include machine learning and embedded systems.

VALERIO FRESCHI received the Graduate degree in electronic engineering from the University of Ancona, Italy, in 1999, and the Ph.D. degree in computer science engineering from the University of Ferrara, Italy, in 2006. He is currently an Assistant Professor in computer engineering with the Department of Pure and Applied Sciences (DiSPeA), University of Urbino, Italy. His research interests include wireless sensor networks, energy-efficient algorithms, optimization, and machine learning.

• • •