Applied Soft Computing 68 (2018) 784–794

Contents lists available at ScienceDirect

Applied Soft Computing

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

Real-time electrocardiogram streams for continuous authentication☆

Carmen Camara[a], Pedro Peris-Lopez[a,b,*], Lorena Gonzalez-Manzano[a], Juan Tapiador[a]

[a] Carlos III University of Madrid, Avda. de la Universidad 30, 28911, Leganes, Spain
[b] Aalto University, Konemiehentie 2, 02150 Espoo, Finland

ARTICLE INFO

Article history:
Received 31 January 2017
Received in revised form 7 July 2017
Accepted 13 July 2017
Available online 21 July 2017

Keywords:
Datastreams
Healthcare
Identification
Electrocardiogram

ABSTRACT

Security issues are becoming critical in modern smart systems. Particularly, ensuring that only legitimate users get access to them is essential. New access control systems must rely on continuous authentication (CA) to provide higher security level. To achieve this, recent research has shown how biological signals, such as electroencephalograms (EEGs) or electrocardiograms (ECGs), can be useful for this purpose. In this paper, we introduce a new CA scheme that, contrarily to previous works in this area, considers ECG signals as continuous data streams. The data stream paradigm is suitable for this scenario since algorithms tailored for data streams can cope with continuous data of a theoretical infinite length and with a certain variability. The proposed ECG-based CA system is intended for real-time applications and is able to offer an accuracy up to 96%, with an almost perfect system performance (kappa statistic >80%).

1. Introduction

Security applications are gaining momentum in modern societies. With the advent of information technologies, data and resources are available almost anytime, anywhere. One key aspect is to ensure that the access to these elements is provided for authorized users only. This need is usually referred to as access control [1].

As a prerequisite of access control, the identity of the user has to be established. This process is called *authentication* and is especially critical when sensitive data is at stake. For instance, access to medical records can be forbidden until being authenticated [2].

Authentication can be carried out by means of something the user *knows*, something the user *has* and/or something the user *is* [3]. Among these three alternatives, the latter is receiving particular attention as a consequence of the evolution of *biometrical* systems, i.e., the acquisition of body-related variables called *biosignals* [4]. For example, entering a building after fingerprint recognition or accessing a smartphone application after facial scanning are two cases of these systems [5,6]. Recent developments for medical devices open up the door to the access of biosignals in almost real-time. These devices can be placed over the skin (e.g., a external heart rate monitor), semi-implanted (e.g., an insulin pump) or within the body (e.g., a pacemaker or a neurostimulator). Implantable medical devices (IMDs) is the general term used to refer to electronic devices implanted within the body. IMDs are designed to provide a medical treatment, to monitor the patient's status, or to enable a particular capability in the patient [7].

Different biosignals have already been considered for authentication purposes, including the electroencephalogram (EEG) [8], the photoplethysmograph (PPG) [9] and the Electrocardiogram (ECG) [10]. Likewise, the continuous availability of biosignals enables performing an advanced form of authentication, called continuous authentication (CA). This variant is different from non-continuous authentication (NCA). In NCA, the user is authenticated once at time T, for example when s/he is logged in a system with authentication checking. On the contrary, in a CA setting the user is authenticated every period of time T_i, thus ensuring the continued presence of the user.

Biosignal-based CA approaches have a direct benefit: users cannot transfer their privileges to other parties, since it must be the very same user who is periodically authenticated. Despite this benefit, one drawback is that biosignals evolve over time and may be slightly different from time to time. As a consequence, the authentication mechanism should be continuously enhanced and not static as time goes by [11].

☆ This work was supported by the MINECO grant TIN2013-46469-R (SPINY: Security and Privacy in the Internet of You); by the CAM grant S2013/ICE-3095 (CIBERDINE: Cybersecurity, Data, and Risks), and by the MINECO grant TIN2016-79095-C2-2-R (SMOG-DEV - Security mechanisms for fog computing: advanced security for devices).

* Corresponding author at: Carlos III University of Madrid, Avda. de la Universidad 30, 28911, Leganes, Madrid, Spain.
E-mail addresses: macamara@pa.uc3m.es (C. Camara), pperis@inf.uc3m.es, pedro.peris-lopez@aalto.fi (P. Peris-Lopez), lgmanzan@inf.uc3m.es (L. Gonzalez-Manzano), jestevez@inf.uc3m.es (J. Tapiador).

https://doi.org/10.1016/j.asoc.2017.07.032
1568-4946/

Such a continuous enhancement and the adaptation to changes makes artificial intelligence (AI) techniques particularly suitable. In particular, as the process requires telling apart the legitimate user from other subjects, it can be considered a classification problem. This problem has been frequently solved through AI and, particularly, data mining and machine learning techniques. Machine learning focuses on the design of algorithms to make predictions after the identification of structural patterns in data. In general a model is created, trained with part of the existing dataset and evaluated with the remaining part of the dataset [12].

Since biosignals can be retrieved in real-time, this can be taken as an advantage to permanently refine the authentication mechanism. To this end, beyond machine learning, *data stream mining* can be applied. data stream mining (DSM) is a recent IA technique that leverages data streams to adapt the classification model when a change is detected [13]. This is especially beneficial for the case of biosignals, due to their aforementioned evolution over time. Interested readers are urged to consult [14–16] for a detailed introduction to DSM and related concepts.

Contribution & organization: Despite the potential of DSM for biosignal-based CA, this approach has not been previously explored. To this end, this paper presents the use of DSM for a particular type of cardiac signal, namely ECG data. The proposed solution allows the use of ECG streams in real-time applications in which the credentials of the users are validated in a continuous-fashion. For the generation of the ECG streams in the CA setting different approaches have been assessed. For completeness, the NCA scenario has been also evaluated.

The rest of this paper is structured as follows. Section 2 introduces related work in this area. The main differences between data mining and data stream mining are explained in Section 3. Our system is first introduced in Section 4 and the details are provided in Section 5. In Section 6, we present the results of our experimentation for NCA and CA approaches and a discussion is provided in Section 7. Finally, in Section 7 we draw some conclusions and outline some future work.

2. Related work

The use of biometrics is widespread nowadays, from the use of the touchscreen in smart devices [17] to a more common approach like fingerprint-based identification [18]. Biological signals are currently taking an important role in the authentication field [19] and they are considered useful biometric traits. Multiple physiological signals are used in this context, such as the EEG signal [8], the PPG signal [9], or the ECG signal [10].

Though many existing works deal with classical authentication using assorted biometric traits [20], continuous authentication systems and applications have been also extensively developed. For instance, Niinuma et al. [21] use the facial skin and color clothes to authenticate users. In the context of mobile devices, facial [22] and touch screen recognition [23] have been applied. Signal processing has also been used in this field, particularly PPG and ECG signals. Although some proposals work with PPG [24,25], here we focus on those related to ECG signals since electrocardiograms are a richer signal from the information point of views—PPG signals only provide beats and average heart rate.

Focusing on ECG signal authentication, several works are devised. In [26] the QRS complex, the most stable component of ECG signal, is applied in the continuous authentication process. After preprocessing the QRS complex and extracting the cross-correlation of QRS signals between a pair of templates, the matching score is computed through different strategies, including using the mean, median, percentile and maximum values. Experiments attempt to analyze the permanence and stability of the biometric features extracted from the QRS complex in ECG signals on a time period of a day. Guennoun et al. [27] use several ECG features to perform continuous authentication. The Mahalanobis distance is then calculated between a heartbeat and a previously stored one such that results depend on a threshold when the process has been repeated for 35 heartbeats. The main limitation of this interesting proposal is that each ECG record only lasts 15 min and an experiment consumes around 30 s. In [28], the autocorrelation/linear discriminant analysis (AC/LDA) algorithm is applied for the design of the biometric features extracted from the ECG signal. Each time an authentication is performed the signal is preprocessed and the result is matched with the stored one. The proposal was tested with a population of 10 individuals and the length of each ECG record is only 5 min. A different approach is proposed in [29], the ECG signal is converted into strings to be later classified. This proposal shows promoting results but, as in previous works, the used ECG signals were recorded during a short time interval (<15 min).

Nonetheless, despite the use of ECG signals for continuous authentication, existing works are evaluated over cardiac signals of a few minutes length at maximum. The variability of the signals and, hence, that of the model, is not considered. An authentication model, created from an observed set of samples does not have to be always the same and it may evolve. Data streams are a useful way to manage this issue. In fact, they are already used in the context of data outsourcing [11] but, to the best of our knowledge, this contribution is the first time continuous authentication with ECG streams is applied.

3. Data analysis: data mining vs. data stream mining

Data mining refers to the set of technologies to handle larger datasets to find patterns, trends or rules and explain data behavior [12]. These technologies have consolidated due to the huge amount of data which is everyday collected and handled. Indeed, this is a trend which continues growing at a fast pace in different areas, for instance in the healthcare context [30].

However, given the amount of data often available the question is: What if data cannot be fitted in memory? In this case smaller training sets are demanding such that algorithms process subsets of data at a time. Then, the goal is the development of a learning process linear in the number of samples. In other words, the problem is that while data mining handles too much data, it does not consider the continuous supply of data. Models cannot be updated when new information arrives and the complete training process has to be repeated. Furthermore the length of the data feed is hugely larger—for instance, imagine a cardiac signal monitored during the entire life of an individual.

Opposed to traditional data mining, data stream mining (DSM) has emerged as a paradigm to address the continuous data problem. The core assumption is that training samples arrive very fast and should be processed and discarded to make room for new samples, thus being processed one time only. More specifically, DSM presents a set of different requirements [31]:

- **Uniqueness:** Each sample must be processed before a new one arrives and it has to be done only once, without being possible the retrieval of any previous samples.
- **Limitation of resources:** Use a limited amount of memory and work in a limited amount of time. Concerning the limitation of memory, this is one of the main motivations of using data streams because memory can be overloaded when too much data is stored in it. This restriction is physical and though it can be addressed using external storage, algorithms should scale linearly in the number of samples to work in a limited amount of time.

- **Immediacy:** An algorithm should be ready to produce the best model at any time regardless of the number of processed samples.

Concerning data mining algorithms, lazy and eager algorithms are key types to be distinguished [12]. In the former type no action is performed during the training phase, such that training data is stored and it waits until testing starts. By contrast, in eager algorithms a model is constructed from training data to apply testing on its regard. In the context of DSM, lazy and eager approaches are available but existing algorithms must be adapted to the data stream setting [13].

Other noteworthy types are parametric and non-parametric algorithms [32]. Parametric algorithms are those in which parameters are of fixed size and the model does not change regardless of the amount of data. Despite their simplicity, speed and less data required in the training phase, they are appropriate for simple problems and they are not well-fitted. Some examples are logistic regression [33] or naive Bayes [34]. On the contrary, non-parametric algorithms are useful for learning when there is too much data and no prior knowledge. Flexibility to fit to a number of functional forms and high performance are some of their main advantages, while they require a substantial amount of training data, they are slower in the training phase and the training data could be overfitted if not carefully performed. Some common algorithms in this class are support vector machines [35], decision trees [36] and the K-nearest neighbour (K-NN) [12]. Among all these algorithms, K-NN is often used for its simplicity and efficiency [37]. Basically, the problem K-NN solves is to identify the point in a dataset closer to a set of given ones. In the training phase, any assumptions about the classification of samples is performed. In the classification, K samples belonging to the training set that are closest to the sample are used as a good indicator to determine an unknown class, generally using a majority voting.

4. System overview

An application of our system is depicted in Fig. 1. Imagine an air traffic control tower where there are controllers who should be permanently monitoring planes and, thus, verifying that everything runs smoothly. In this situation we have to consider that: (1) an intruder may enter into the tower trying to cause some damage; (2) a controller may try to do the work of another; and (3) physiological indices such as the heart rate is not constant and may vary according to each situation (e.g., too many plains about to take off or landing). In this regard, an authentication system requires the continuous authentication of each controller verifying that no impersonation attacks are performed and that each controller is in the work place no matter ECG fluctuations.

According to the example in Fig. 1, the system works in the following way assuming that captured ECG signals are sent to a central unit (e.g., a smartphone or a nearby computer). Firstly, in the set-up phase, the ECG signal of each controller is observed (collected) for some time (i.e., 30 min) and, once cleaned and pre-processed, a reference model is constructed (similarity module). Secondly, in the operating phase, the system is prepared to start the authentication process, in this case verifying that each controller is in the tower throughout the office hours and feels well. In a first step, ECG records are cleaned, features are extracted and each ECG sample passes (or is discarded) by the similarity module. After that, the observed ECG signal of each controller is compared against the reference model (learner), also using part of the signal for learning and adjusting the model accordingly. Note that in case a change is produced, e.g., due to stress caused by 10 planes landing in a sort period of time, the ECG signal may change. However, as the model adapts dynamically to the situation, no alarm will be activated. In contrast, in case a big change in the ECG signal is detected, due to someone impersonating a controller or s/he feeling suddenly very dizzy, the authentication fails and an alarm is activated. The steps followed during the set-up and operation phases are summarized in Algorithm 1.

Algorithm 1. ECG-based authentication

procedure SET-UP PHASE
1. Capture ECG records
2. Pre-process &Extract features
3. Build reference model for each user

end procedure
procedure OPERATION PHASE
1. Capture ECG records
2. Pre-process &Extract features
3. Pass/discard ECG samples (similarity module)
4. Authenticate samples (learner module)
5. Update the learner (if necessary).

end procedure

5. System description

The general architecture of an ECG-based authentication system is displayed in Fig. 2. Firstly, in this paper we assume that the cardiac signal is acquired by an IMD (e.g., a pacemaker or an implantable cardioverter defibrillator), or perhaps by external sensors attached to the body of the individual. Once ECG signals are recorded, they need to be preprocessed before feature extraction. To do this the ECG signal is split into time windows and, for each window, a set of numerical features are extracted. Then, the similarity module filters samples discarding those that do not seem to come from the user. Finally, the samples are classified using a classifier such as a decision tree, a support vector machine, a nearest neighbor algorithm, and so on. In fact, nearly all data mining algorithms can be tailored for coping with the data stream problem.

More details about each component of the ECG-based authentication system (see Fig. 2) are explained in the following sections.

5.1. Dataset and pre-processing

Pacemakers and implantable cardioverter defibrillators are the most extended class of implantable devices (the first pacemakers date from the early 1950s [38]). An electrocardiogram (ECG) represents the electrical activity of the heart during a period of time. In particular, an ECG chart is composed of a set of waves: P, Q, R, S and T [26].

The ECG records are cleaned as the first step. For that, the zero-frequency component (DC bias) is eliminated and then the records are passed through a pass-band filter. An entire-raw ECG signal from a user U_j is divided into windows of L seconds:

$$ECG^{\mathcal{U}_j} = \left\{ ECG^{\mathcal{U}_j}_{w(1)}, ECG^{\mathcal{U}_j}_{w(2)}, \ldots, ECG^{\mathcal{U}_j}_{w(N)} \right\} \quad (1)$$

where $N \gg 1$. Subsequently each $ECG^{\mathcal{U}_j}_{w(i)}$ is the input of the feature extraction module.

5.2. Feature extraction and similarity module

Fiducial and non-fiducial approaches can be followed for extracting features of a physiological signal. Fiducial-based approaches are those in which characteristic points (e.g., Q, R and S peaks in ECG signals [39,26]) in the time-domain can be used for the feature extraction. On the other hand, non-fiducial approaches obtain features from a transformed domain (e.g., Fourier or wavelet [40,41]). In terms of performance, fiducial-based and transform-based approaches achieve similar results as reported in the comparative analysis by Odinaka et al. [42]. In our particular case, we opt for avoiding any sort of manipulation of

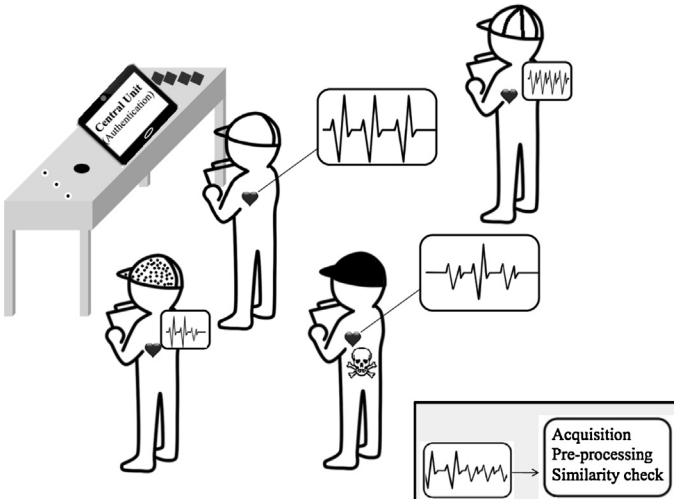

Fig. 1. Example of an scenario for continuous ECG-based stream authentication.

the ECG signal in the time domain. The efficiency and simplicity of the system are the main goals that justify to work in a transform domain.

In particular, in the time-domain the ECG signal is only segmented into windows—this is the minimal possible manipulation of the signal. After that, a transform (TF) is applied over each ECG window and a set of M coefficients is obtained:

$$F^{\mathcal{U}_j}_{w(i)} = TF\left(ECG^{\mathcal{U}_j}_{w(i)}\right) = \left\{f^{\mathcal{U}_j}_{w(i)}(0), f^{\mathcal{U}_j}_{w(i)}(1), \ldots f^{\mathcal{U}_j}_{w(i)}(M)\right\} \quad (2)$$

Each of these feature vectors $F^{\mathcal{U}_j}_{w(i)}$ is passed through the similarity module which discards bad ECG samples, that is, samples which are considered to be too far from the reference model (outliers). The reasoning behind this is that the learner only analyzes "good" feature vectors and there is a benefit in the performance of the system in comparison with the obtained without this filtering.

Each user is responsible for the similarity checking module. For that, the user computes a reference matrix, which is called the reference module. To do so, in the set-up phase, the ECG signal is observed during a T_R time interval (e.g., half an hour of continuous cardiac signal motorization). A matrix of N average vectors is computed. Each of these vectors (Y_i where $i = \{1, \ldots, N\}$) represents an average value of ECG windows (L seconds) in the Hadamard domain:

$$\bar{Y}_i = \frac{1}{T_R/(L \times N)} \cdot \sum_{q=1}^{\frac{T_R}{L \times N} \times i} F^{\mathcal{U}_j}_{w\left(q + \frac{T_R}{L \times N} \times (i-1)\right)} \quad (3)$$

Then, the similarity module works in the following way. A set of N new observed ECG windows ($F^{\mathcal{U}_j}_{w(i^*)}$, where $i^* = \{1, \ldots, N\}$) are discarded or not depending on their similarity to the user's reference model. We use the Pearson's linear correlation coefficient (*corr*) to measure similarity between two matrices. Other similarity metrics could be used but we chose this due to its invariant behaviour to scale and shift changes. Mathematically, the module is described by the following equation:

$$\begin{cases} \text{Discard ECG samples} & \text{If} \quad |\, corr\left(\begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \cdots \\ \bar{Y}_N \end{bmatrix}, \begin{bmatrix} F^{\mathcal{U}_j}_{w(1^*)} \\ F^{\mathcal{U}_j}_{w(2^*)} \\ \cdots \\ F^{\mathcal{U}_j}_{w(N^*)} \end{bmatrix}\right) |\, < \delta \\ \text{Transmit ECG samples to the learner} \left(\begin{bmatrix} F^{\mathcal{U}_j}_{w(1^*)} & F^{\mathcal{U}_j}_{w(2^*)} & \cdots & F^{\mathcal{U}_j}_{w(N^*)} \end{bmatrix}\right) & \text{Otherwise} \end{cases} \quad (4)$$

where the parameter δ is tuned through experimentation.

Finally, the ECG streams are sent to the learner. We have evaluated two approaches: (1) buffered solution; (2) unbuffered solution. In the former, each ECG stream represents an average value of feature vectors (Hadamard domain) during an observation period T_O:

$$\bar{F}^{\mathcal{U}_j}_{w(i)} = \frac{1}{T_O/L} \cdot \sum_{q=1}^{\frac{T_O}{L} \times i} F^{\mathcal{U}_j}_{w\left(q + \frac{T_O}{L} \times (i-1)\right)} \quad (5)$$

The unbuffered approach is the most demanding scenario as each ECG stream represents the feature vector of an ECG window, i.e., $F^{\mathcal{U}_j}_{w(i)}$.

Fig. 3 depicts the creation of ECG streams for both proposed approaches. Also, considering that samples of different users can be received at different time and thus no order is expected, an illustrative example of several samples of a data stream of two users ($\{j, j^*\}$) is shown below:

$$\begin{cases} \left[\bar{F}^{\mathcal{U}_j}_{w(i)} \bar{F}^{\mathcal{U}^*_j}_{w(i)} \bar{F}^{\mathcal{U}^*_j}_{w(i+1)} \cdots \bar{F}^{\mathcal{U}_j}_{w(N)} \bar{F}^{\mathcal{U}_j}_{w(N)}\right] & \text{Buffered approach} \\ \left[F^{\mathcal{U}_j}_{w(i)}, F^{\mathcal{U}_j}_{w(i+1)}, F^{\mathcal{U}^*_j}_{w(i)} \cdots F^{\mathcal{U}_j}_{w(N)}, F^{\mathcal{U}_j}_{w(N)}\right] & \text{Unbuffered approach} \end{cases} \quad (6)$$

5.3. Learner

As introduced in Section 3, a wide set of methods (e.g., decision trees, Bayesian methods, lazy, ensemble, etc.) can be used for the classification problem. Regardless of the used algorithm, a relevant aspect is how data is treated. Two approaches have been considered depending on whether the data is acquired in a continuous way or not and thus used in real-time or no real-time applications. In a real-time application in which cardiac records arrive continuously in a non-predefined order, an on-line analysis is used and

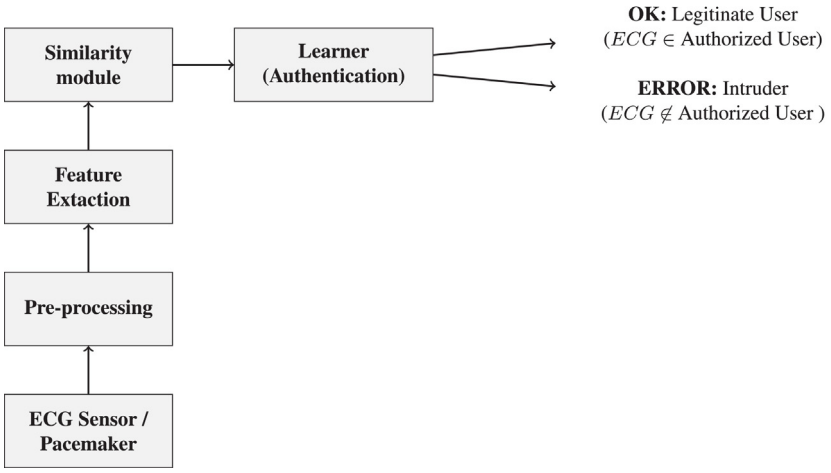

Fig. 2. General structure of an ECG-based authentication system.

ECG streams are evaluated by interleaving testing and training (i.e., prequential evaluation) and following a sliding window strategy in which the size of the window is fixed and the buffer keeps the newest instances. Similar to the first-in, first-out data structures [43], and illustrated in Fig. 4, whenever a new instance is inserted into the window, another instance $j - S$ is forgotten—S represents the window size. In particular, each new instance is used to test the model prior being used for training. Regarding the generation of data streams, buffered and unbuffered approaches are considered. Nonetheless, in a non real-time application, referred to as batch setting, the dataset can be split intro training and testing and there is not memory restrictions.

In terms of security, non real-time applications correspond to a NCA system. On the other hand, the analysis in real-time of the ECG data streams (user credentials) conforms with the requirements of a CA system.

6. Experimental validation

Established parameters and results achieved after experimentation are presented in the following sections.

6.1. Experimental settings

Table 1 provides the experimental setting used to validate our approach. The experiments were performed using the recordings of 10 individuals from the MIT-BIH Normal Sinus Rhythm Database [44]. The individuals under study do not show any relevant medical problem and were observed during a long time period. Besides, Table 1 provides a brief motivation for each choice of values.

The Walsh–Hadamard transform (HT) is the chosen transformation in our system. The HT performs a projection of a signal onto a set of square waves, called Walsh functions (WAL). Mathematically, the forward and inverse HT of a signal $x(t)$ of length W are defined as

$$y_n = \frac{1}{W}\sum_{i=0}^{M-1} x_i WAL(n,i), n = 1, 2, \ldots, W-1 \quad (7)$$

$$x_i = \frac{1}{W}\sum_{i=0}^{M-1} y_i WAL(n,i), n = 1, 2, \ldots, W-1 \quad (8)$$

Using the HT is justified by two main reasons. On the one hand, this transform is computationally efficient as it just consists of a matrix multiplication (that of the signal and the Walsh matrix). On the other hand, it has the ability of compressing the input signal, with the majority of the signal information being kept on the lower coefficients in the transformed domain. Therefore HT is efficient is terms of computation and memory requirements. Note here that the usage of other transforms (e.g., Fourier or Wavelet) were evaluated and discarded, mainly, due to their complexities in terms of computational requirements.

As for the learner, the K-nearest neighbour (K-NN) is the algorithm used [12]. In the NCA setting, the dataset is divided into training and testing—60/40 and 80/20 are the splits commonly used in this area [45]. In the CA analysis, a tailored K-NN is employed as the learner, which uses a buffer memory to keep a small portion of the instances (training ones). For updating, this buffer follows a sliding window strategy with a first-in-first-out (FIFO) rule. We refer the reader to [43] for a detailed introduction to data streams and drift concept.

The reasoning of using this learner is twofold: (1) efficiency; and (2) simplicity [37]. Regarding efficiency, a K-NN often outperforms more complex learners [12]. In relation to simplicity, a K-NN does not require complex computations. In detail, new samples are classified taking into account the class to which a set of training samples (N nearest instances) belong. Although more complex learners could have been used, we consider the K-NN the most appropriated since it offers a high performance and its simplicity facilitates the implementation of the system in portable devices with constrained resources.

6.2. Results

The main goal of the system is to achieve a high performance in the identification of the users enrolled in the system. Two approaches have been evaluated depending whether the data is sent o not in a continuous fashion to the learner (i.e., core of the CA system):

Non-continuous authentication (NCA) A data mining approach makes sense when we deal with non-real time applications and there is not severe memory restrictions. During a first phase (training), data of all the enrolled users is recorded and stored in memory. The classifier is then trained using these samples. After that, credentials (ECG streams) of the users are checked (testing phase)—for instance, user credentials are verified each time she/he attempts to unlock the touchscreen of her/his smartphone. Note that, motivated by the need to achieve high performance, the buffered approach is applied in our experimentation.

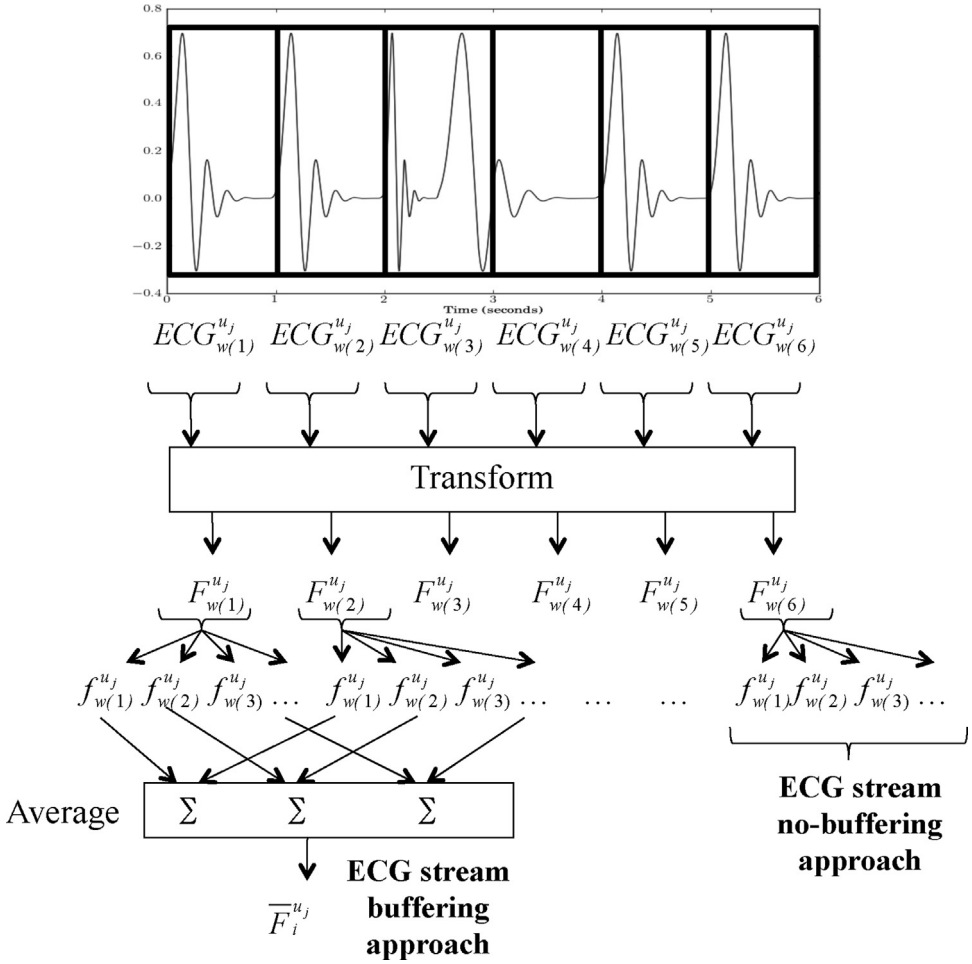

Fig. 3. ECG stream samples of buffered and unbuffered approaches.

Table 1
Established parameters.

Parameters	Value	Justification
Pass-band filter	0.67 Hz and 45 Hz	Using this filter, the main sources of noise such as the respiration noise or the power line noise are canceled
N	3	As a trade-off between efficiency and be able to check the variability of the ECG signal
δ	10^{-1}	A relative-low correlation threshold
L	2 s	To guarantee the observation of 2 or 3 heart beats depending on how fast each individual is beating
M	256	Number of coefficients needed to keep the ECG information at a low level (e.g., PQRST waves)
T_O	3 min	Observation period to guarantee the stability of the ECG signal
T_R	30 min	Time-interval needed to characterize the "common" samples of a user.
K-NN	$K=1$	Greater values of parameter K do not offer higher performances and increase system complexity
Max. num. instances in memory in CA approach (S)	10% of the tested dataset	Trade-off between memory efficiency and system performance

In connection to the application scenario described in Section 4, the credentials of each controller would be checked when s/he logs on the system (e.g., whenever her/his computer is turned on). **Continuous authentication (CA)** Classical approaches are not feasible when data is provided in a continuous way and memory restrictions exist. The use of an on-line analysis approach is much more suitable for processing data streams. Tools like massive online analysis (MOA) [31], VFML [46] and RapidMiner [47] are commonly used for mining data streams.

Following the scenario introduced in Section 4, the credentials of each controller would be verified at regular time intervals. In the unbuffered approach there is only a distance of few seconds between intervals, and this distance considerably increases to hundreds of seconds in the buffered approach. Accordingly, we have evaluated both approaches in Sections 6.4.1 (buffered approach) and 6.4.2 (unbuffered approach).

6.3. Non-continuous authentication (NCA)

We have a population of individuals and average feature vectors have been acquired at regular intervals. For simplicity, we use regular intervals in our experiments. This is not a limitation and

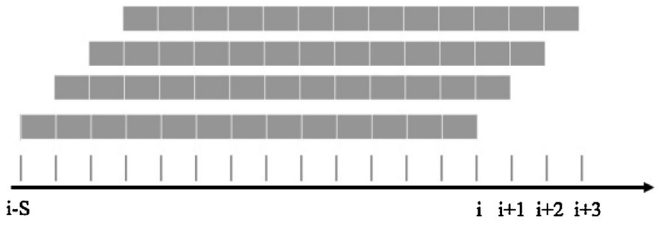

Fig. 4. Example of a sliding window strategy.

non-regular intervals might be also employed. In our experimentation, the population size is set to 10 and each individual was sensed during a period of 11 h. To assess the impact of the training dataset, the performance of the system has been evaluated for different training sizes. Fig. 5 shows the accuracy (correctly classified instances) and kappa statistic for several values of the training set. For the hard case (i.e., 10% of the whole dataset or, in other words, each individual is observed during around 1 h) the accuracy is over 93.5%. Therefore, the system performs well even when the training phase is set to minimal values. When we use common values (60% or 80% [45]) for the training set, the performance is almost perfect (97.4% and 97.9%). The value of kappa, which is greater than 0.81 for all the training sizes, shows a perfect agreement [48]—that is, the influence of "random guessing" is minimal.

To guarantee the robustness of our results, we have also tested the classifier using a 10-fold cross-validation. The accuracy and kappa statistic are 97.90% and 97.68%, respectively. Apart from showing a significantly high True Positive Rate (97.9%), as expected from a good identification system, the weighted average False Positive Rate is extremely low (0.2%). The detailed accuracy by class and the confusion matrix are summarized in Tables 2 and 3, respectively. All in all, the metrics indicate that the system is very close to an ideal identification system.

6.4. Continuous authentication (CA)

The buffered and unbuffered approaches have been tested—see Section 5 for a detailed explanation in the generation of the used ECG streams.

6.4.1. CA: buffered approach

We have tested the performance of the system for a population of individuals. According to the proposed use case (Section 4), assume that credentials (ECG streams) of controllers working in the same room of the tower are checked by a central unit in a continuous fashion at long-separated intervals. In this context, each subject generates an ECG data stream in a continuous way during a long period of time (i.e., 11 h per individual in our experiments). Each sample of the stream represents an average value in the Hadamard domain, as explained in Section 5.2 (see Eq. (5) for details). The population size has been set to 10 as in the NCA setting.

In Fig. 6 we can see the evolution of the accuracy over the time using a prequential evaluation. The learner employed is a nearest neighbor (i.e., K-NN with $K = 1$), as in the NCA setting but with a sliding window (maximum number of instances stored in memory) of reduced dimensions. In our experiments, 10% of the total instances are kept in memory. Having overcome the penalty of the first instances, the system exceeds the threshold of a 90% and swiftly stabilizes around 96% of correctly classified instances. Similarly, the kappa statistic rapidly exceeds the 80% threshold, approaching an almost perfect classifier performance.

6.4.2. CA: unbuffered approach

We have assessed the system when the ECG streams are generated in a continuous way and at a high speed rate. Again, based on the proposed use case (Section 4), imagine a controller which has a cardiac problem and the authentication should be permanently to ensure the proper functioning of the system.

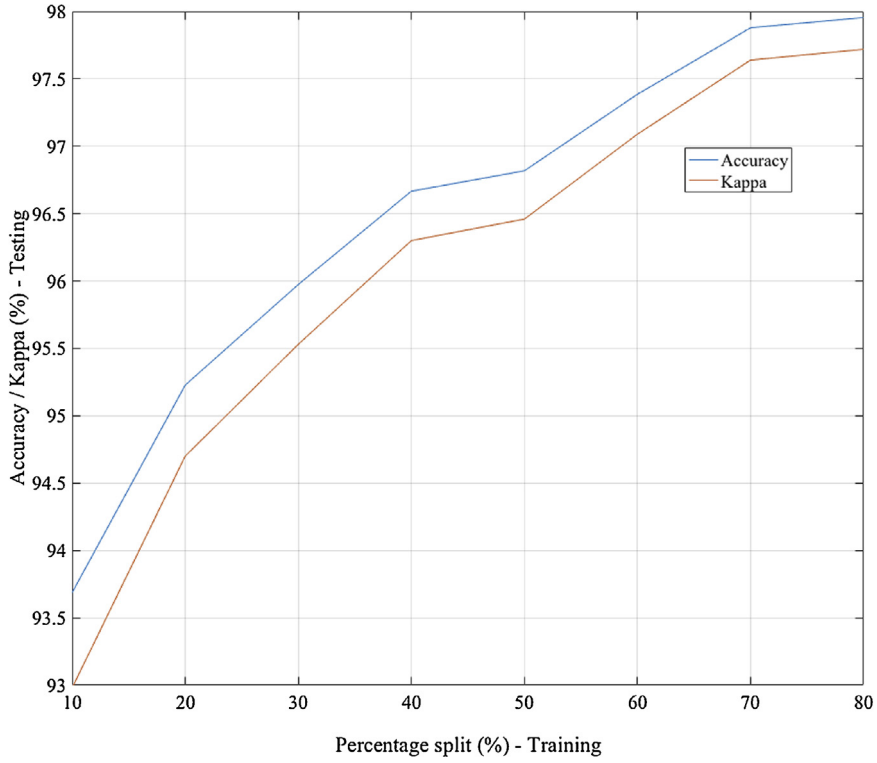

Fig. 5. Assessment of the training size for a small population (NCA).

Table 2
Accuracy by class—NCA setting.

Class	TP rate	FP rate	Precision	Recall	*F*-measure	ROC area
1	0.995	0.002	0.986	0.995	0.991	0.997
2	0.982	0.001	0.995	0.982	0.989	0.991
3	0.936	0.007	0.936	0.936	0.936	0.965
4	0.968	0.004	0.964	0.968	0.966	0.982
5	0.955	0.005	0.955	0.955	0.955	0.975
6	1.000	0.000	1.000	1.000	1.000	1.000
7	0.982	0.001	0.995	0.982	0.989	0.991
8	0.991	0.001	0.991	0.991	0.991	0.995
9	0.986	0.002	0.982	0.986	0.984	0.992
10	0.995	0.002	0.986	0.995	0.991	0.997
Weighted avg.	0.979	0.002	0.979	0.979	0.979	0.988

Table 3
Confusion matrix—NCA setting.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	219	0	1	0	0	0	0	0	0	0
S2	0	216	0	2	0	0	0	0	0	2
S3	2	0	206	3	6	0	0	2	1	0
S4	0	0	4	213	3	0	0	0	0	0
S5	0	1	6	3	210	0	0	0	0	0
S6	0	0	0	0	0	220	0	0	0	0
S7	1	0	0	0	1	0	216	0	2	0
S8	0	0	2	0	0	0	0	218	0	0
S9	0	0	1	0	0	0	1	0	217	1
S10	0	0	0	0	0	0	0	0	1	219

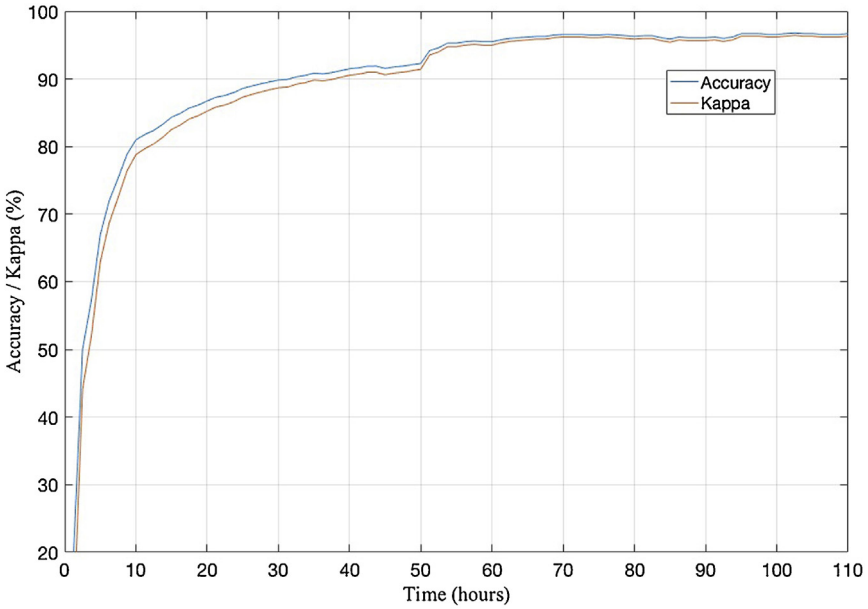

Fig. 6. System performance: CA (buffered approach).

In the previously described scenario, we have a unique legitimate individual. Therefore, the system has to distinguish between two classes. That is, the data streams belong to the legitimate user (IMD holder in our example) or to any other unauthorized/fraudulent user (the attacker, in general terms). We have tested this setting for each one of the 10 individuals of the buffered approach. Therefore, in each experiment there is a legitimate user and the other ones are categorized within a unique class (fraudulent user). The performance (accuracy) for each of these aforementioned experiments in summarized in Fig. 7. The accuracy is on average around 80%, with a standard deviation of 3.7%. Therefore, the system works well—in most cases, authorized users are correctly distinguished from intruders—and is particular well-suited to cope with the slight changes in the ECG data streams.

To assess the influence of whether the ECG streams are buffered or not, we have tested this setting using both approaches. In Tables 4 and 5 we show the obtained results. In terms of accuracy, the buffered approach offers a benefit of around 15% in comparison with the unbuffered approach. The Kappa statistic points out how the performance of the system switches from “substantial” to “almost perfect” accuracy when we move from the unbuffered to the buffered approach. Apart from performance metrics, the use of one approach or the other depends on the requirements demanded by the real-time application in question. The determining factor is

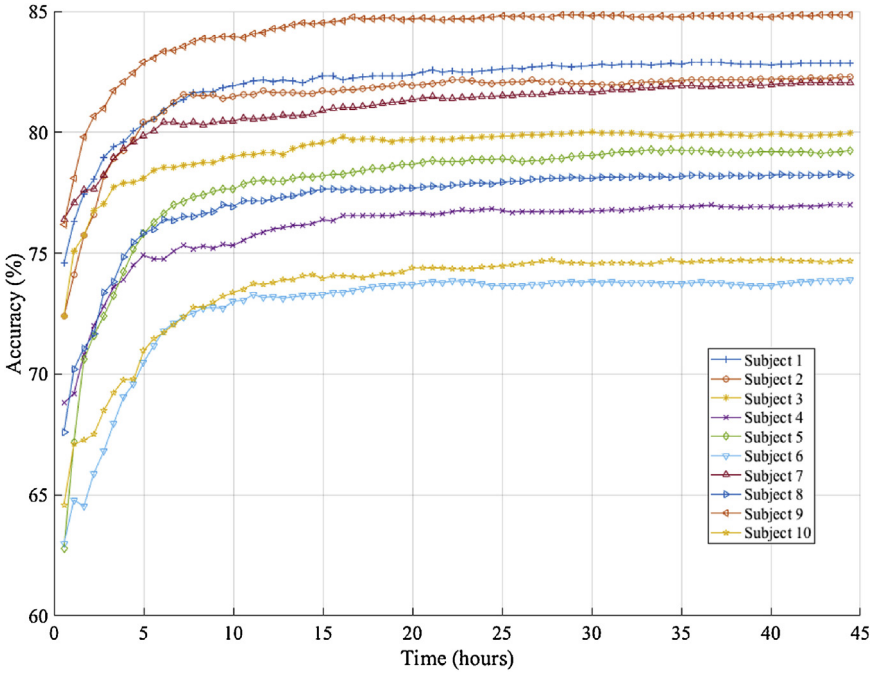

Fig. 7. System performance: CA (unbuffered approach).

Table 4
CA: unbuffered approach (two classes).

Subject	Average accuracy	Average kappa
S1	81.98	63.95
S2	81.39	62.79
S3	79.28	58.57
S4	76.00	52.05
S5	77.81	55.58
S6	72.75	45.49
S7	81.02	62.04
S8	77.17	54.35
S9	84.12	68.18
S10	73.51	47.12
Average	78.50	57.01

Table 5
CA: buffered approach (two classes).

Subject	Average accuracy	Average kappa
S1	96.80	93.59
S2	95.10	90.19
S3	91.32	82.50
S4	97.34	94.68
S5	94.06	88.13
S6	92.09	84.09
S7	94.80	89.62
S8	94.35	88.69
S9	97.95	95.90
S10	94.04	88.05
Average	94.79	89.54

the rate at which the ECG streams are examined. In the unbuffered approach, the ECG streams are provided almost instantly (i.e., intervals of 2 s). In contrast, only 20 examples/h is the sample rate used in the buffered approach (i.e., intervals of 3 min). Therefore, the particular application of the system will driven the used option.

7. Discussion

Although some authors have already explored the problem of continuous authentication with cardiac signals (e.g., ECG [27] and PPG [24], the used datasets are made up of records with length of only a few minutes), this is the first time that ECG records are interpreted and processed as data streams. In our opinion, a data stream approach fits perfectly the problem of CA, particularly in the case of ECG signals—and, more generally, physiological signals with a slight variability and a theoretical infinite length. We have considered the typical assumptions for classification in the DSM setting [49]: (1) each sample has a fixed number of attributes that are less than several hundreds; (2) the number of classes is limited and small (in our experiments, ten classes are considered at maximum); (3) we assume that the learner has a small memory; the size of the training dataset is larger than the available memory; and finally, (4) the speed rate of processing each sample is moderate high (the precise value is conditioned to the device that supports on-board the learner).

Data stream algorithms have the potential to deal with potential infinite amount of data. Regarding physiological signals, as far as we know, recordings are taken during a maximum period of 24 h in the best case [50]. The execution time of the algorithm used scales linearly with the number of examples. In our experimental setting, the learner consumes several tens of milliseconds per sample using a Quad Core 2.7 GHz Intel Core i5 with 16GB of RAM. Using this value (or the equivalent if different equipment is used), an upper bound of the time necessary for processing an arbitrary number of examples may be computed.

Although important variations on ECG streams only occur after 5 years observation period [51], we can find slight variations from time to time—that is, data is not stationary. This is often referred as concept drift. To dealt with this, old instances should become irrelevant to characterize the current state of the system and this information would have to be forgotten by the learner. The interested reader can consult [43] for a detailed explanation of the main existing approaches in the literature. In our particular case,

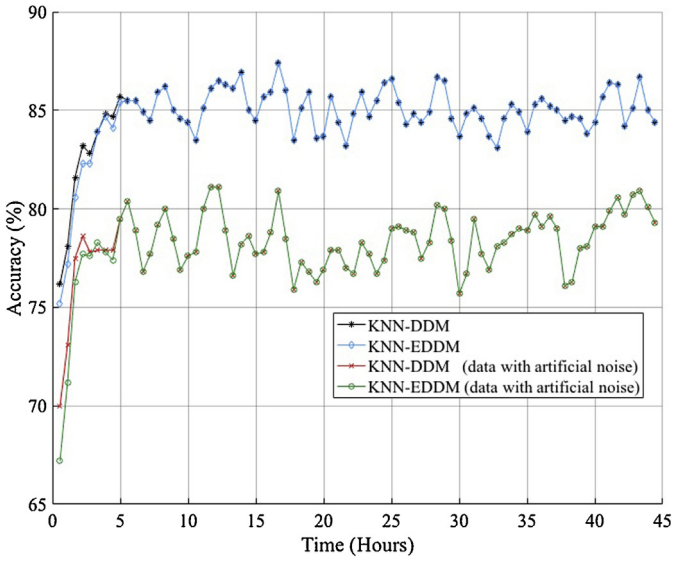

Fig. 8. System performance: CA (unbuffered approach) with drift detection.

Table 6
Average performance: CA (unbuffered approach) with drift detection.

Approach	Accuracy (average value)
KNN	84.1200 ± 1.5095
KNN-DDM	84.8000 ± 1.6290
KNN-EDDM	84.7300 ± 1.7921
KNN-DDM (with artificial noise)	78.2500 ± 1.7060
KNN-EDDM (with artificial noise)	78.1600 ± 1.9831

as explained in Section 5.3, we keep only the most recent samples in memory and the memory size is fixed—sliding window strategy.

Aside from using a limited memory, we can benefit from drift detection mechanisms that reset the learner model and trigger the learning of a new one when a significant change is detected. We have tested two well-known methods: drift detection method (DDM) and early drift detection method (EDDM) [52]. In a nutshell, DDM is based on monitoring the number of errors produced by the learner during prediction—errors are modelled by a binomial distribution. DDM performs well to detect abrupt changes and not very slow gradual changes. EEDM was proposed with the aim of improving the detection of gradual changes and keeping a good performance with abrupt changes. Instead of considering only the number of errors as in DDM, it also takes into account the distance (number of examples) between two classification errors.

The performance of the two aforesaid methods has been evaluated with one of the subjects of the CA (unbuffered approach) setting which is our more demanding scenario. The subject 9 has been selected for this experimentation without prejudice to the generality in the results. More precisely, DDM and EDDM algorithms are used as a wrapper on the KNN learner. We have tested two scenarios: (1) the original data stream; (2) artificial noise has been added to the original data—10% and 5% are the fractions of attributes values and class labels that have been disturbed, respectively. Fig. 8 displays the obtained results and Table 6 summarizes the average values. In both cases, DDM and EDDM converge to the same accuracy values which points out that the gradual changes in the ECG records are not very slow. In terms of performance, the KNN with drift detection marginally improves our previous results of only using a KNN with sliding window. In addition, drift detection methods work well even when the data streams are quite noisy—the performance only suffers a brief dip. Note that we have overstated the used example since the noise remains during the whole data stream and often it is intermittent.

Finally, a key-aspect in the processing of cardiac signals is the time period during which the ECG is observed. This aspect is examined at the end of Section 6.4.2—see Tables 4 and 5 for details. In the buffered approach, each stream is linked with the observation of the ECG during a moderate long time period with the extra benefit of achieving a very high performance. In the unbuffered approach, the sending of the examples to the learner is almost instantaneous with the penalty of a slightly degradation of the performance in comparison with the buffered approach. The choice of one approach or another would be conditioned by the processing speed rate demanded by the learner. In our particular case (a CA system), we have the possibility to check the credentials of an individual almost instantaneously (each 2 s) or just remain patient and proceed with the verification once every 3 min.

8. Conclusions

We are currently in an era in which our surrounding devices generate and transmit data in a continuous way. An example of these devices are those belonging to the Internet-of-Things (IoT) or the new generations of implantable medical devices (IMDs) with wireless connectivity. These devices receive data continuously and very frequently in a non-orderly fashion. One use of such data is user authentication. In particular, the use of biological signals has been previously studied for authentication purposes. Cardiac signals (PPG or ECG) and brain signals (EEG) collected from IMDs or body sensors, are widely used for authentication and some authors have applied them to the CA scenario [24,29]. However, given the continuous nature of the authentication process, the system has to be adapted to changes; for example, ECG signal may slightly change over time. Thus, DSM emerges as a promising technique to face this sort of problems. To the best of our knowledge, none of the existing solutions use ECG signals as data streams.

We exploit the full potential of DSM for designing a CA system using ECG streams. The proposed real-time system has been evaluated using records of 10 individuals monitored during approximately half a day. Our results show the potential of ECG streams for security purposes. In fact, the behaviour of the classifier, which is the core of the CA system, is almost perfect. The CA approach achieves an accuracy as high as the NCA approach but with the benefit of using a limited memory and being able to process data streams. Moreover, we have tested the buffered and unbuffered approaches in the CA setting to show how the use of one or another is driven by the requirements of the real time application (e.g., credentials/second that must be checked by the CA system). Finally, we have studied how drift detection techniques (e.g., DDM or EDDM) may help to deal with the existing changes in the ECG data streams—a wrapper approach has been tested. The results clearly indicate that drift detection techniques are effective to build robust CA schemes even under very noisy conditions.

As a future work, we plan to check whether the concept of ECG streams can be extended to other physiological signals. We hope this contribution can serve as seed to many other works that explore the use of biological signals for continuous authentication.

Conflict of interest

Authors declare that they have no conflict of interest.

References

[1] R.S. Sandhu, P. Samarati, Access control: principle and practice, IEEE Commun. Mag. 32 (9) (1994) 40–48.

[2] S. Krawczyk, A.K. Jain, Securing electronic medical records using biometric authentication, in: International Conference on Audio- and Video-Based Biometric Person Authentication, Springer, 2005, pp. 1110–1119.
[3] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Trans. Circ. Syst Video Technol. 14 (1) (2004) 4–20.
[4] J.L. Semmlow, B. Griffel, Biosignal and Medical Image Processing, CRC press, 2014.
[5] Q. Li, C. Jin, W. Kim, J. Kim, S. Li, H. Kim, Multi-feature based score fusion method for fingerprint recognition accuracy boosting, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) (2016) 1–4.
[6] A.A. Mandavkar, R.V. Agawane, Mobile based facial recognition using OTP verification for voting system, IEEE International Advance Computing Conference (IACC) (2015) 644–649.
[7] G. Haubrich, L. Twetan, G. Rosar. Multiple band communications for an implantable medical device, January 15 2005. US Patent App. 11/035,518.
[8] J. Hu, Z. Mu, EEG authentication system based on auto-regression coefficients, 10th International Conference on Intelligent Systems and Control (ISCO) (2016) 1–5.
[9] A. Lee, Y. Kim, Photoplethysmography as a form of biometric authentication, in: IEEE Sensors, IEEE, 2015, pp. 1–2.
[10] S.J. Kang, S.Y. Lee, H.I. Cho, H. Park, ECG authentication system design based on signal analysis in mobile and wearable devices, IEEE Signal Process. Lett. 23 (6) (2016).
[11] S. Papadopoulos, Y. Yang, D. Papadias, CADS: continuous authentication on data streams, in: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB Endowment, 2007, pp. 135–146.
[12] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.
[13] A. Bifet, R. Kirkby, Data Stream Mining a Practical Approach, 2009.
[14] M.M. Gaber, J. Gama, S. Krishnaswamy, J.B. Gomes, F. Stahl, Data stream mining in ubiquitous environments: state-of-the-art and current directions, Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 4 (2) (2014) 116–138.
[15] I. Khamassi, M. Sayed-Mouchaweh, M. Hammami, K. Ghédira, Discussion and review on evolving data streams and concept drift adapting, Evol. Syst. (2016) 1–23.
[16] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Wozniak, F. Herrera, A survey on data preprocessing for data stream mining: current status and future directions, Neurocomputing 239 (2017) 39–57.
[17] D.-H. Shih, C.-M. Lu, M.-H. Shih, A flick biometric authentication mechanism on mobile devices, in: International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS), IEEE, 2015, pp. 31–33.
[18] D. Peralta, S. García, J.M. Benitez, F. Herrera, Minutiae-based fingerprint matching decomposition: methodology for big data frameworks, Inf. Sci. 408 (2017) 198–212.
[19] K. Revett, F. Deravi, K. Sirlantzis, Biosignals for user authentication-towards cognitive biometrics? in: International Conference on Emerging Security Technologies (EST), IEEE, 2010, pp. 71–76.
[20] J. Wayman, A. Jain, D. Maltoni, D. Maio, An Introduction to Biometric Authentication Systems, Springer, 2005.
[21] K. Niinuma, U. Park, A.K. Jain, Soft biometric traits for continuous user authentication IEEE Trans. Inf. Forensics Secur. 5 (4) (2010) 771–780.
[22] U. Mahbub, V.M. Patel, D. Chandra, B. Barbello, R. Chellappa, Partial Face Detection for Continuous Authentication, 2016 arXiv:1603.09364.
[23] M. Frank, R. Biedert, E. Ma, I. Martinovic, D. Song, Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication, IEEE Trans. Inf. Forensics Secur. 8 (1) (2013) 136–148.
[24] A. Bonissi, R.D. Labati, L. Perico, R. Sassi, F. Scotti, L. Sparagino, A preliminary study on continuous authentication methods for photoplethysmographic biometrics, IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BioMS) (2013) 28–33.
[25] H. Gascon, S. Uellenbeck, C. Wolf, K. Rieck, Continuous authentication on mobile devices by analysis of typing motion behavior, in: Sicherheit, Citeseer, 2014, pp. 1–12.
[26] R.D. Labati, R. Sassi, F. Scotti, ECG biometric recognition: permanence analysis of QRS signals for 24 hours continuous authentications, in: IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2013, pp. 31–36.
[27] M. Guennoun, N. Abbad, J. Talom, S.M.M. Rahman, K. El-Khatib, Continuous authentication by electrocardiogram data, in: 2009 IEEE Toronto international conference on Science and Technology for Humanity (TIC-STH), IEEE, 2009, pp. 40–42.
[28] R. Matta, J.K.H. Lau, F. Agrafioti, D. Hatzinakos, Real-time continuous identification system using ECG signals, in: 24th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2011, pages 001313-001316.
[29] D.P. Coutinho, A.L.N. Fred, M.A.T. Figueiredo, ECG-based continuous authentication system using adaptive string matching, Biosignals (2011) 354–359.
[30] D. Miljkovic, D. Aleksovski, V. Podpečan, N. Lavrač, B. Malle, A. Holzinger, Machine learning and data mining methods for managing Parkinson's disease, in: Machine Learning for Health Informatics, Springer, 2016, pp. 209–220.
[31] A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer, MOA: massive online analysis, J. Mach. Learn. Res. 11 (August) (2010) 1601–1604.
[32] S. Russell, P. Norving, Artificial Intelligence: A Modern Approach, 3rd ed., Pearson Education Limited, 2014.
[33] R. Greiner, X. Su, B. Shen, W. Zhou, Structural extension to logistic regression: discriminative parameter learning of belief net classifiers, Mach. Learn. 59 (3) (2005) 297–322.
[34] A. McCallum, K. Nigam, et al., A comparison of event models for naive Bayes text classification, in: AAAI-98 Workshop on Learning for Text categorization, vol. 752, Citeseer, 1998, pp. 41–48.
[35] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, Support vector clustering, J. Mach. Learn. Res. 2 (Dec) (2001) 125–137.
[36] D.K. McIver, M.A. Friedl, Estimating pixel-scale land cover classification confidence using nonparametric machine learning methods, IEEE Trans. Geosci. Remote Sens. 39 (9) (2001) 1959–1968.
[37] M.S. Aldayel, K-nearest neighbor classification for glass identification problem, International Conference on Computer Systems and Industrial Informatics (2012) 1–5.
[38] O. Aquilina, A brief history of cardiac pacing, Paediatric Cardiol. 8 (2) (2008) 17–81.
[39] Y. Gahi, M. Lamrani, A. Zoglat, M. Guennoun, B. Kapralos, K. El-Khatib, Biometric identification system based on electrocardiogram data, Int. Conference on New Technologies, Mobility and Security (NTMS) (2008) 1–5.
[40] F. Agrafioti, D. Hatzinakos, ECG based recognition using second order statistics, 6th Annual Conference on Communication Networks and Services Research (CNSR) (2008) 82–87.
[41] M. Hejazi, S.A.R. Al-Haddad, Y.P. Singh, S.J.I. Hashim, A.F.A. Aziz, ECG biometric authentication based on non-fiducial approach using kernel methods, Digit. Signal Process. 52 (2016) 72–86.
[42] I. Odinaka, L. Po-Hsiang, A.D. Kaplan, J.A. O'Sullivan, E.J. Sirevaag, J.W. Rohrbaugh, ECG biometric recognition: a comparative analysis, IEEE Trans. Inf. Forensics Secur. 7 (6) (2012) 1812–1824.
[43] J. Gama, Knowledge Discovery from Data Streams, Chapman and Hall/CRC, 2010.
[44] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, Ch.P. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (June (23)) (2000) e215–e220, http://dx.doi.org/10.1161/01.CIR.101.23.e215, Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215PMID:1085218.
[45] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.
[46] G. Hulten, P. Domingos, VFML – A Toolkit for Mining High-Speed Time-Changing Data Streams, 2003.
[47] M. Hofmann, R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Chapman & Hall/CRC, 2013.
[48] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[49] G. Bifet, A. Holmes, R. Kirkby, B. Pfahringer. Data Stream Mining: A Practical Approach. howpublished = Technical Report, year=2012, institution=University of Waikato, http://www.cs.waikato.ac.nz/abifet/MOA/StreamMining.pdf.
[50] PhysioNet. PhysioBank, 2017.
[51] C. Camara, P. Peris-Lopez, J.E. Tapiador, Human identification using compressed ECG signals, J. Med. Syst. 39 (11) (2015) 1–10.
[52] P.M. Gonçalves, S.G.T. de Carvalho Santos, R.S.M. Barros, D.C.L. Vieira, A comparative study on concept drift detectors, Expert Syst. Appl. 41 (18) (2014) 8144–8156.