This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



# Symmetrical Compression Distance for Arrhythmia Discrimination in Cloud-based Big-Data Services

J.M. Lillo-Castellano, I. Mora-Jiménez, R. Santiago-Mozos, F. Chavarría-Asso, A. Cano-González, A. García-Alberola, and J.L. Rojo-Álvarez, *Senior Member, IEEE*

***Abstract*—The current development of cloud computing is completely changing the paradigm of data knowledge extraction in huge databases. An example of this technology in the cardiac arrhythmia field is the SCOOP platform, a national-level scientific cloud-based Big Data service for Implantable Cardioverter Defibrillators (ICD). In this scenario, we here propose a new methodology for automatic classification of intracardiac electrograms (EGMs) in a cloud computing system, designed for minimal signal preprocessing. A new Compression-based Similarity Measure (CSM) is created for low computational burden, so-called weighted fast compression distance, which provides better performance when compared with other CSMs in the literature. Using simple machine learning techniques, a set of 6848 EGMs extracted from SCOOP platform were classified into seven cardiac arrhythmia classes and one noise class, reaching near to 90% accuracy when previous patient arrhythmia information was available and 63% otherwise, hence overcoming in all cases the classification provided by the majority class. Results show that this methodology can be used as a high-quality service of cloud computing, providing support to physicians for improving the knowledge on patient diagnosis.**

***Index Terms*—Cardiac Arrhythmia Classification, Implantable Defibrillator, Intracardiac Electrogram, Weighted Fast Compression Distance, Big Data Analytics.**

## I. Introduction

Cardiovascular diseases are the primary cause of death worldwide, and they remain directly or indirectly responsible for more than 30% of reported deaths [1]. Arrhythmias are common cardiac conditions in which the contraction rate of the heart is abnormally fast, slow or irregular, usually resulting in an abnormal mechanical function of the heart. Typically, physicians classify arrhythmias into two classes, namely, supraventricular (with atrial origin and usually nonfatal), and ventricular (severe arrhythmias originated at ventricules). Based on their mechanisms and origin, fast arrhythmias are further classified into monomorphic tachycardias (ventricular or supraventricular), flutter (either atrial or ventricular), and fibrillatory rhythms (atrial and ventricular fibrilllation) [2].

JMLC, IMJ, RSM, and JLRA are with the Department of Signal Theory and Communications, Telematics and Computing, Rey Juan Carlos University. Dep. III, Camino del Molino s/n, 28943, Fuenlabrada, Spain. eMail: {josemaria.lillo, inmaculada.mora, ricardo.santiago.mozos, joseluis.rojo}@urjc.es. JLRA is Prometeo Researcher with Electric and Electronic Department, Universidad de las Fuerzas Armadas ESPE, Ecuador.

FCA and ACG are with Hospital Solutions at Medtronic Ibérica® S.A., C/ María de Portugal 9, 28050, Madrid, Spain. eMail: {fernando.chavarria, alicia.cano}@medtronic.com.

AGA is with Arrhythmia Unit, Hospital Universitario Virgen de la Arrixaca. Ctra. Madrid-Cartagena s/n, 20120, El Palmar, Spain. eMail: arcadi@secardiologia.es.

The usual treatment for patients with high risk of suffering a severe arrhythmia is the Implanted Cardioverter Defibrillator (ICD) [3], which is a battery-powered device placed under the skin that tracks the heart electrical signals by using thin wires (leads) directly lodged in the apex (sometimes in the septum) of the right ventricle. Though the ICD has limited memory and computational resources, it is able to: (1) record the intracardiac electrical signals, known as electrograms (EGM); (2) automatically detect arrhythmic episodes; and (3) provide with a therapy depending on the severity of the detected arrhythmic episode, such as pacing, cardioversion, or defibrillation [3]. Whenever the ICD detects an arrhythmia, it records the EGM for subsequent clinical inspection and future research, which allows to determine, during the follow-up, whether the treatment was adequate. Thus, when the patient visits the physician, all these arrhythmic episodes are downloaded into a database to be inspected and analyzed.

To date, despite the huge amount of research, Cardiac Arrhythmia Classification (CAC) is still an open and active research field [4]. The design of CAC algorithms have to deal with: (1) EGMs of different duration and absence of recorded signal in temporal segments; (2) databases with a small number of records, which hinders good generalization for CAC; (3) strong dependence of the algorithm with the usual preprocessing and feature extraction stages, determined sometimes by a specific ICD; and (4) lack of an appropriate gold standard in the arrhythmia episode labelling, which requires expert consensus and involves high economic effort.

Recently, the Spanish company Medtronic Ibérica® S.A. has developed the SCOOP platform, a pioneer scientific repository system in the cloud involving 48 national hospitals and conveying around 12,000 EGM records of cardiac arrhythmia episodes stored by ICDs and subsequently labelled by a scientific committee of expert physicians. The aim of this platform is to provide researchers with a high number of intracardiac episodes, thus planting the seed to generate new scientific knowledge in cardiac electrophysiology. In the context of the SCOOP cloud platform, we here propose a new Big Data based methodology for CAC, characterized by: (1) minimal EGM preprocessing; (2) independence of the EGM duration; (3) integrable into a high-quality cardiology cloud computing system. Specifically, the proposed methodology classifies a new unlabelled EGM by using the similarities to the labelled EGMs available in the Big Data repository.

In order to provide with a system addressing the previous requirements, we focus here on similarity measures based on Information Theory. This kind of measures, such as Normal-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



ized [5] and Fast [6] Compression Distance (NCD and FCD), provide the similarity between two signals based on their common patterns by using concepts related to compression length and dictionaries. In addition, they allow to design parameter-free methodologies and to compare signals with different lengths, even when temporal segments have no data or when signals are sampled at different rates. However, neither NCD nor FCD are symmetrical measures, and NCD requires a high computational burden [6]. To overcome both drawbacks, we define a new similarity measure, so-called Weighted Fast Compression Distance (WFCD), which is compared with NCD and FCD in terms of classification performance and computational burden. This new measure enables a fast arrhythmia episode classification in SCOOP, opening the road towards new advanced support to diagnosis and clinical knowledge discovery.

The remaining of the paper is organized as follows. Section II presents the basis on data compression that is necessary to describe the similarity measures based on the dictionary matching concept, later presented in Section III. Section IV describes the SCOOP platform and the cardiac signals dataset. The proposed arrhythmia classification methodology and experiment description are described in Section V. Results and discussion are presented in Sections VI and VII, respectively.

## II. Background on Compression-based Similarity

Lossless data compression in Information Theory refers to the reduction in the number of bits required to represent some specific information without loss [7]. A Compression-based Similarity Measure (CSM) is characterized by its ability to exploit the amount of information shared by two elements [8]. To date, many clustering and classification problems have been tackled by different systems whose core was a CSM, hence allowing to design parameter-free methodologies and to deal with data of diverse nature, such as genomics and phylogeny, images, biomedical signals, or text documents [6], [9], [10], [11].

Given a bit string $\boldsymbol{x}$, the Kolmogorov complexity (denoted by $K_{\boldsymbol{x}}$) is defined as the number of bits of the shortest computer program of the fixed reference computing system capable of producing $\boldsymbol{x}$ [12]. Hence, $K_{\boldsymbol{x}}$ can be seen as the ultimate compressed version of $\boldsymbol{x}$ from which $\boldsymbol{x}$ can be exactly recovered by a decompression program. Intuitively, $K_{\boldsymbol{x}}$ corresponds to the minimum amount of information required to generate $\boldsymbol{x}$.

The *Information Distance* between two strings $\boldsymbol{x}$ and $\boldsymbol{y}$, denoted by $ID_{\boldsymbol{xy}}$, is defined as the length of the shortest program computing $\boldsymbol{x}$ *from* $\boldsymbol{y}$ [13]. It can be expressed by using the Kolmogorov Complexity as follows,

$$ID_{\boldsymbol{xy}} = K_{\boldsymbol{xy}} - min\{K_{\boldsymbol{x}}, K_{\boldsymbol{y}}\}, \quad (1)$$

where $K_{\boldsymbol{xy}}$ is the Kolmogorov complexity of string $\boldsymbol{xy}$ (concatenated pair $\boldsymbol{x}$ and $\boldsymbol{y}$), and $min\{\cdot,\cdot\}$ denotes the minimum operator. It can be proven that $ID_{\boldsymbol{xy}}$ is actually a metric which depends on the strings length [13]. As a consequence, if the value of $ID_{\boldsymbol{xy}}$ between two short strings is large with respect to their lengths, then both strings are very different; however, if

TABLE I
Dictionary obtained with LZW algorithm on $\boldsymbol{x} = 0110011001$.

| **Word** | 0 | 1 | 01 | 11 | 10 | 00 | 011 | 100 |
|---|---|---|---|---|---|---|---|---|
| **Code** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

the same value of $ID_{\boldsymbol{xy}}$ is obtained for two long strings, then those strings are very similar. Thus, a relative measure was defined [5] to express similarity regardless of the string length, called normalized information distance (NID) and given by

$$NID_{\boldsymbol{xy}} = \frac{ID_{\boldsymbol{xy}}}{max\{K_{\boldsymbol{x}}, K_{\boldsymbol{y}}\}}, \quad (2)$$

where $max\{\cdot,\cdot\}$ denotes the maximum operator. Note that NID is in the range from zero to one [5].

In practice, the Kolmogorov complexity cannot be computed [12], hence data compressors are used to approximate the Kolmogorov complexities involved in Eqs. (1) and (2). For a given compressor $C$, let us denote by $C_{\boldsymbol{x}}$ and $C_{\boldsymbol{xy}}$ the length, in bits, of the compressed versions for strings $\boldsymbol{x}$ and $\boldsymbol{xy}$, respectively. Using this approximation in Eq. (2), the NCD can be defined [5] as

$$NCD_{\boldsymbol{xy}} = \frac{C_{\boldsymbol{xy}} - min\{C_{\boldsymbol{x}}, C_{\boldsymbol{y}}\}}{max\{C_{\boldsymbol{x}}, C_{\boldsymbol{y}}\}}. \quad (3)$$

Ideally, NCD values close to zero (one) correspond to similar (non-similar) strings. In practice and when real-world compressors are used, $C_{\boldsymbol{xy}} \neq C_{\boldsymbol{yx}}$ and the symmetry property is not fulfilled by the NCD [5]. In addition, NCD values can be occasionally slightly greater than one.

## III. Proposed Similarity Measurement

The working principle of a compressor consists of finding repeated patterns in the data (known as *words*), assigning each pattern to a description (known as *code*) of a size inversely proportional to its occurrence frequency [14], and creating a table (known as *dictionary*) with the word-code assignation. The *cardinality* (denoted by $|\cdot|$ ) of a dictionary is the number of words in it, and it depends on three factors: first, the string lengths; second, the maximum word-size allowed; and third, the compression algorithm used to find words in the string. Some of the mostly used compressors are LZW, zip, Gzip, LZMA, and PPMZ [7].

For a simple compression example using the LZW algorithm, let us consider the bits string $\boldsymbol{x} = 0110011001$. The result is a string of codes $\boldsymbol{x}_c = 0110242$, whose dictionary is shown in Table I (see [15]). The set of words $W_{\boldsymbol{x}}$ of this dictionary, with $|W_{\boldsymbol{x}}| = 7$, corresponds to the first row of Table I, and it gives us an idea of the information in $\boldsymbol{x}$.

The relationship between two strings is determined by their common patterns. Let us consider strings $\boldsymbol{x}$ and $\boldsymbol{y}$, and their associated sets of words $W_{\boldsymbol{x}}$ and $W_{\boldsymbol{y}}$, extracted from their dictionaries. Topological relations between $W_{\boldsymbol{x}}$ and $W_{\boldsymbol{y}}$ can be represented by a Venn diagram, see Fig. 1, which allows to easily visualize: (1) the total amount of information, i.e., $W_{\boldsymbol{x}} \cup W_{\boldsymbol{y}}$, referring to the total number of dictionary words (area inside the thick line); (2) the shared information, i.e., $W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}$, referring to the shared dictionary words (gray zone);

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



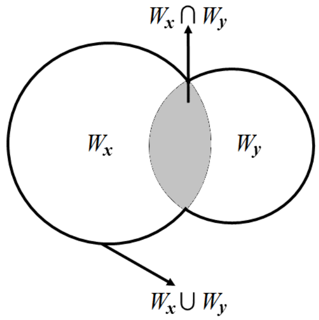



Fig. 1. Venn diagram illustrating the topological relationship among the words of two dictionaries ($W_{\boldsymbol{x}}$ and $W_{\boldsymbol{y}}$). The intersection set is shown as a gray zone, and black thicker line delimits the union set.

and (3) the properties always satisfied by their cardinalities, given by

$$min\{|W_{\boldsymbol{x}}|, |W_{\boldsymbol{y}}|\} = p_s \cdot max\{|W_{\boldsymbol{x}}|, |W_{\boldsymbol{y}}|\} \ , \tag{4}$$

$$\begin{aligned} |W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}| &= p_i \cdot min\{|W_{\boldsymbol{x}}|, |W_{\boldsymbol{y}}|\} \\ &= p_s \cdot p_i \cdot max\{|W_{\boldsymbol{x}}|, |W_{\boldsymbol{y}}|\} \ , \end{aligned} \tag{5}$$

$$\begin{aligned} |W_{\boldsymbol{x}} \cup W_{\boldsymbol{y}}| &= |W_{\boldsymbol{x}}| + |W_{\boldsymbol{y}}| - |W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}| \\ &= max\{|W_{\boldsymbol{x}}|, |W_{\boldsymbol{y}}|\} \cdot (1 + p_s - p_s \cdot p_i) \ , \end{aligned} \tag{6}$$

where parameters $p_s, p_i \in [0, 1]$ depend on the compression scenario. Last equations are the basics to understand the idea of similarity based on dictionary matching.

Two well-known CSMs based on dictionary matching are the Normalized Dictionary Distance (NDD) [16] and the FCD [6]. The NDD is a similarity metric in the range $[0, 1]$, defined as

$$NDD_{\boldsymbol{xy}} = \frac{|W_{\boldsymbol{x}} \cup W_{\boldsymbol{y}}| - min\{|W_{\boldsymbol{x}}|, |W_{\boldsymbol{y}}|\}}{max\{|W_{\boldsymbol{x}}|, |W_{\boldsymbol{y}}|\}} \ . \tag{7}$$

Note that NDD and NCD have a formally similar equation, and by comparing Eqs. (3) and (7), it comes that $C_{\boldsymbol{x}}$ ($C_{\boldsymbol{y}}$) is approximated by $|W_{\boldsymbol{x}}|$ (by $|W_{\boldsymbol{y}}|$), whereas $C_{\boldsymbol{xy}}$ is approximated by $|W_{\boldsymbol{x}} \cup W_{\boldsymbol{y}}|$. This means that $NDD_{\boldsymbol{xy}} \approx NCD_{\boldsymbol{xy}}$. Regarding FCD, it is defined as

$$FCD_{\boldsymbol{xy}} = \frac{|W_{\boldsymbol{x}}| - |W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}|}{|W_{\boldsymbol{x}}|} \ , \tag{8}$$

where the cardinality of the intersection set $|W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}|$ is used. Given that FCD does not fulfill the symmetry property, we propose a simple way to convert it into a metric by using the maximum and minimum operators as follows,

$$\begin{aligned} mFCD_{\boldsymbol{xy}} &= min\{FCD_{\boldsymbol{xy}}, FCD_{\boldsymbol{yx}}\} \\ &= \frac{min\{|W_{\boldsymbol{x}}|,|W_{\boldsymbol{y}}|\} - |W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}|}{min\{|W_{\boldsymbol{x}}|,|W_{\boldsymbol{y}}|\}} \ , \end{aligned} \tag{9}$$

$$\begin{aligned} MFCD_{\boldsymbol{xy}} &= max\{FCD_{\boldsymbol{xy}}, FCD_{\boldsymbol{yx}}\} \\ &= \frac{max\{|W_{\boldsymbol{x}}|,|W_{\boldsymbol{y}}|\} - |W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}|}{max\{|W_{\boldsymbol{x}}|,|W_{\boldsymbol{y}}|\}} \ . \end{aligned} \tag{10}$$

Using Eqs (6) and (7), we get $NDD_{\boldsymbol{xy}} = MFCD_{\boldsymbol{xy}}$. Since NDD just accounts for one FCD measure between $\boldsymbol{x}$ and $\boldsymbol{y}$, it is reasonable to assume that NDD is disregarding some information conveyed by $mFCD_{\boldsymbol{xy}}$ that can be relevant for estimating the similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$. We define then a new CSM, so-called *Weighted Fast Compression Distance* (WFCD), taking into account both $FCD_{\boldsymbol{xy}}$ and $FCD_{\boldsymbol{yx}}$ and weighting them according to the relative cardinality of each dictionary as follows,

$$WFCD_{\boldsymbol{xy}} = \frac{|W_{\boldsymbol{x}}| \cdot FCD_{\boldsymbol{xy}} + |W_{\boldsymbol{y}}| \cdot FCD_{\boldsymbol{yx}}}{|W_{\boldsymbol{x}}| + |W_{\boldsymbol{y}}|} \ . \tag{11}$$

For a better interpretation, and according to dictionary matching, Eq. (11) can be rewritten into:

$$WFCD_{\boldsymbol{xy}} = 1 - 2 \cdot \frac{|W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}|}{|W_{\boldsymbol{x}}| + |W_{\boldsymbol{y}}|} \ . \tag{12}$$

Note that WFCD is also in the range [0,1].

As it will be shown in Section VI, WFCD is the CSM providing the best CAC performance in our EGM database. In our implementation, each EGM signal is coded as a string of bits where each voltage value is associated to 8 bits (1 byte). The length of each string depends on the number of signal samples. The LZW algorithm is then applied to each string to get a set of dictionaries used to compute the WFCDs among EGMs.

## IV. Cardiac Arrhythmia Database

SCOOP is a Spanish platform developed by Medtronic Ibérica® S.A. for supporting the cooperative knowledge generation in the ICD field. It is technically based in the remote monitoring capabilities of current devices, which makes possible the automatic transmission of their information and structured storage in a remote server. This way, each remote follow-up for each ICD patient is incorporated into a growing database, within an observational framework research study so-called UMBRELLA, ensuring the legal, normative, and scientific data exploitation, as well as privacy requirements [17]. UMBRELLA also complements the digitally stored information from the device with clinical patient data and with their prognosis.

SCOOP started in 2011 and it has stored since then more than 20,000 two channels-EGMs of cardiac arrhythmia episodes from 50 hospitals, with an average follow-up around 2.5 years. Any Spanish hospital can currently join at any time and propose scientific studies with the database, and it is likely that SCOOP will expand to an international environment in the near future. Nowadays, SCOOP platform supports more than 40 scientific running proposals, showing the viability and relevance of a complementary model for knowledge generation. Besides the unprecedented database size in ICD recordings, SCOOP has a quality ensured by a systematic clinical evaluation process of each episode by a scientific committee consisting of 6 expert cardiologist and arrhythmologist (i.e., reviewers) with large knowledge and medical background on reviewing EGMs of ICD stored arrhythmia episodes. The scientific committee defined a set of labels (8-class) with the following criteria.

First, atrial origin arrhythmias consisted of: *Sinus Tachycardia* (ST), with regular P-R intervals, normal P and R-wave, and rate between 100 and 180 bpm; *Atrial Fibrillation* (AF), with varying P-R intervals, irregular P-P intervals, changing P-wave morphology, and atrial rate of 400 bpm or more;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



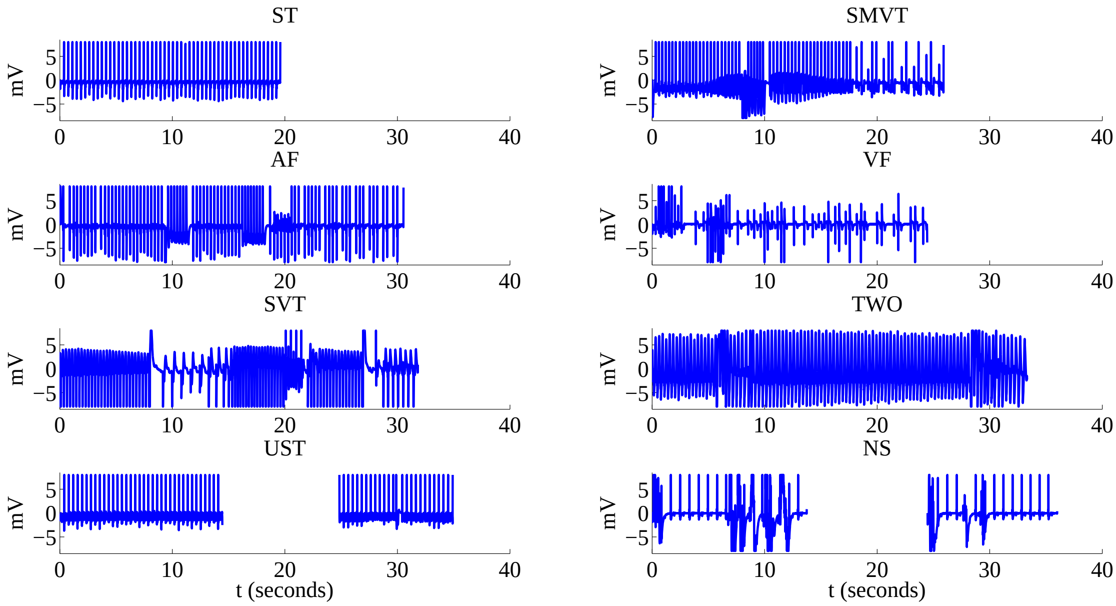



Fig. 2. Overview of the EGM waveforms recorded by the *Vtip to Vring* ICD lead configuration. Each panel shows a random episode among those of the same class.

*Supraventricular Tachycardia or Flutter* (SVT), with regular P-P intervals, stable P-wave morphology, ventricular rate may change and atrial rate $>$ 100 bpm (in atrial flutter, atrial impulses are conducted to the ventricles in various ratios as 2:1 or 4:1, and atrial rate is usually between 250 and 400 bpm); and *Uncertain Supraventricular Tachycardia* (UST), given by fast arrhythmia whose origin can not be determined with the information recorded in the EGM.

Second, ventricular origin arrhythmias consisted of: *Sustained Monomorphic Ventricular Tachycardia* (SMVT), given by regular and sustained rhythms that meet ICD programmed criteria and are finished by one therapy or they last more than 30 s; and *Sustained Polymorphic or Ventricular Fibrillation* (VF), given by episodes with variable morphology and chaotic rhythm, which meet ICD programmed criteria and are finished by one therapy or they last more than 30 seconds.

And third, other consisted of: *T-wave Oversensing* (TWO), given by an inappropriate detection of the T-wave by the device, interpreting this T-wave as R-wave, and resulting in a double counting during a normal sinus rhythm; and *Noise* (NS), given by an irregular and chaotic rhythm, with short duration and rate $>$ 500 bpm, which starts and finishes spontaneously.

If any episode is labeled as *uncertain* by any reviewer during the classification process, it is again labelled by all 6 reviewers in a final review.

For this work, EGMs recorded from January 2012 to December 2013 were selected and labeled, yielding 6848 EGMs from 629 patients. Table II shows the relative occurrence for each arrhythmia in this database, as well as the grouping of these labels into three major sets (3-class), namely, atrial arrhythmias, ventricular arrhythmias, and other. The episodes had lengths of 24.42$\pm$ 16.64 s, with median 19.82 s and interquartile range 13.23 s. The number of episodes per patient was 10.8$\pm$ 22.9, median 4 and interquartile range 10. Both episode length and the number of episodes per patient followed an asymmetric and non-Gaussian distribution (confirmed by standard Kolmogorov-Smirnov normality test [18]).

TABLE II
RHYTHM TYPES DETERMINED BY THE COMMITTEE AND THEIR RELATIVE OCCURRENCES.

| Ocurrence | 3-class | 8-class | Ocurrence |
|---|---|---|---|
| 34.19% | Atrial | ST | 13.96% |
| | | AF | 11.73% |
| | | SVT | 4.48% |
| | | UST | 4.02% |
| 64.06% | Ventricular | SMVT | 60.06% |
| | | VF | 4.00% |
| 1.75% | Other | TWO | 1.20% |
| | | NS | 0.55% |

Due to the diversity of ICD models in the national scenario, several lead configurations were present both for far-field and near-field EGMs. The most usual configurations were pairs [19] *Can to HVB - Vtip to Vring* (3334 episodes), *Atip To Aring - Vtip to Vring* (2899), and *Aring to HVB - Can to Vring* (200). Note that *HVB* denotes the can (coil) electrode, *A(V)tip* denotes atrial (ventricular) tip of the sensing electrode, and similar notation for the ring sensing electrode. Two simultaneously recorded channels were available for each episode. All episodes were sampled at 128 samples per second in the $\pm$8 mV range with an amplitude resolution of 0.063 mV (1 byte of dynamic range). For an overview of the signals recorded by the ICD, we present in Figure 2 eight EGM waveforms (one per class), randomly chosen among those with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



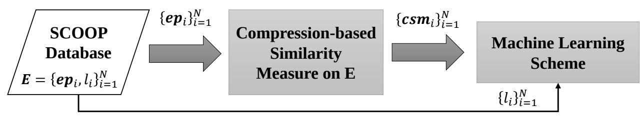



Fig. 3. Diagram of the proposed methodology to classify cardiac arrhythmia episodes in the SCOOP platform. This work focuses on the CSM, working with a subsequent simple machine learning classifier.

the *Vtip to Vring* lead configuration. It is interesting to remark the different episode durations and the absence of recorded signal in several temporal intervals. Both characteristics were dependent on the ICD programming criteria by the clinician.

## V. Experimental Setup

Let us consider a set $E$ of $N$ labeled EGM episodes or instances $E = \{(\boldsymbol{ep}_i, l_i)\}_{i=1}^N$ from $P$ different patients. Each episode, represented by $\boldsymbol{ep}$, is composed of the two channels of the cardiac arrhythmia EGM signal. A label $l$ (ST, AF, SVT, UST, SMVT, VF, TWO, or NS) is associated to each episode according to the experts committee. In order to design a cardiac arrhythmia discriminator from dataset $E$, we propose a two stages methodology (see Fig. 3). First, the CSM is computed for each pair of episodes in $E$, obtaining a vector $\boldsymbol{csm}$ of $N$ components per episode. Second, similarity vectors and labels are used to design the CAC system by following a machine learning approach.

### A. Performance Evaluation

In this work, two different techniques based on resampling are used for estimating the generalization performance. Both techniques are based on the *Leave One Out - Cross Validation* (LOOCV) [20] strategy and estimate performance by considering those episodes that are not used for designing the arrhythmia discriminator.

In LOOCV, the original dataset is divided into $N$ subsets (as many subsets as available instances) and $N$ statistical models (here, machine learning classifiers) are constructed. Each model is designed using a different subset of $N-1$ instances, evaluating the performance on the remaining instance (validation subset). The model performance is estimated as the average performance on the $N$ validation subsets. LOOCV has been shown to give an almost unbiased estimator of the performance in machine learning schemes, hence it provides with a reasonable criterion for model selection and comparison.

In our context, the LOOCV strategy can be applied in two different ways. The first one, by setting the equivalence between instance and episode, what leads to the *Leave One Episode Out - Cross Validation* (LOEOCV) technique. The second one considers the equivalence instance-patient and separates the original dataset $E$ into $P$ different subsets, where each subset contains the episodes of just one patient. We have called this technique *Leave One Patient Out - Cross Validation* (LOPOCV), and it estimates the classification performance on episodes of those patients not considered during the classifier design. Note the difference between LOPOCV and LOEOCV techniques. In LOEOCV, the model can be designed with episodes from the same patient considered in the validation subset (same patient, but not same episode). In LOPOCV, the model is designed with episodes from patients who were not considered in the validation subset (so, every validation subset will have a different number of episodes).

In order to evaluate the ability of CSMs presented in preceding sections for cardiac arrhythmias discrimination, three kinds of experiments were performed with the same machine classification scheme. Each experiment was differenced by the CSM input space (see below) and the number of classes to discriminate (8-class and 3-class cases). Performance was evaluated using both LOEOCV and LOPOCV techniques.

### B. Input Space for CSM Featuring

Each arrhythmic episode consisted of two registers, i.e. two EGM channels denoted as $\boldsymbol{ch}_1$ and $\boldsymbol{ch}_2$. Both channels were simultaneously recorded from different lead configurations and have the same number of samples per episode, whereas different episodes usually have different number of samples. There are several possibilities for processing data from both channels. Though the most direct way is to concatenate both channels to have a signal of twice samples, another option is to convert each episode $\boldsymbol{ep}_i$ into a complex signal $\boldsymbol{c}_i = \boldsymbol{ch}_{1,i} + j \cdot \boldsymbol{ch}_{2,i}$, where $j$ is the imaginary unit. The second possibility, inspired on phase portraits [21], allows to work with the magnitude and phase signals. Previous studies have shown the usefulness of phase portraits for discriminating SVT versus VT [22].

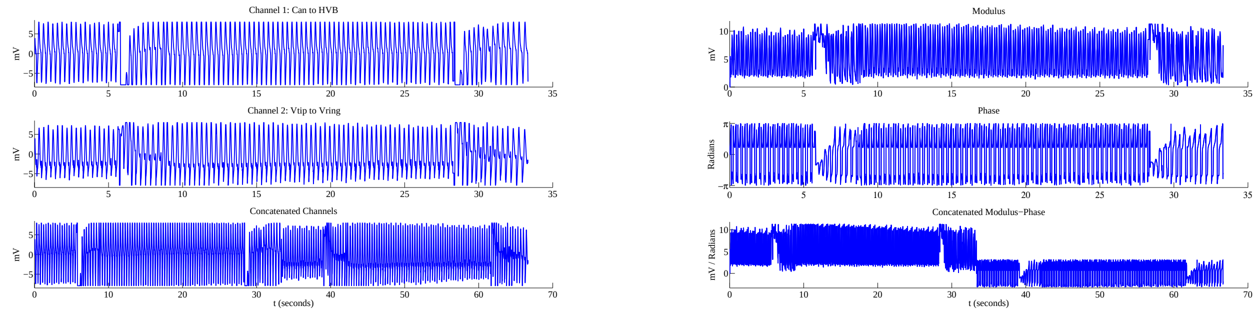
In our methodology, the use of the phase signal can provide a reasonable performance while reducing the computational burden of the whole classification process. Thus, we used four different input spaces for CSM featuring, namely, concatenated channels (Ccat), just modulus (Mod), just phase (Pha), and concatenated modulus-phase (Xcat). Note that the number of samples corresponding to Ccat and Xcat is twice that of Mod and Pha spaces. Figure 4 shows the original waveforms (channels 1 and 2) for the TWO episode (located at 3rd row, 2nd column in Figure 2), as well as the subsequent signals for CSM featuring.

### C. k-Nearest Neighbors Classification Algorithm

In this paper we focus on a methodology using different CSMs for classifying an heterogeneous Big-Data set of signals, rather than benchmarking a wide variety of classification machines. This is the reason to use a classification algorithm as simple and effective as possible, the voting $k$-Nearest Neighbors ($k$-NN) classifier. It is a nonparametric algorithm classifying an instance according to the most frequent label among the labels of its $k$ nearest instances, by using a suitable distance measure [20]. It can be shown that when the number of instances tends to infinity, the voting $k$-NN classifier tends to the optimal one (Bayesian classifier), what is a quite desirable consistency property even though $k$-NN does not assume any statistical model for the data. Furthermore, $k$-NN has been shown to provide a quite reasonable performance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics






(a) (b)

Fig. 4. CSM featuring signals for the TWO episode in Fig. 2: (a) Individual EGM channels (two upper panels) and Ccat (lower panel); (b) Mod, Pha, and Xcat signals.

in comparison to more complex machine classifiers, with the additional advantage of its interpretability.

An essential point in the $k$-NN classifier design is the selection of parameter $k$ (best number of neighbors for voting). To avoid poor performance, the $k$ value must be chosen so that the classifier is not sensitive to atypical instances while avoiding data over-fitting. For this reason, the best $k$ value not only depends on the data distribution, but also on the size and dimensionality of the dataset [20].

## VI. Results

### A. Classification Performance

Real world applications usually have imbalanced datasets, this is, scenarios where instances of different classes are not equally represented in the dataset used for designing the classifier. This fact is specially relevant for evaluating performance in multi-class imbalanced problems. Several merit figures are used to quantify the classifier performance [23], among them, accuracy rate and Cohen's Kappa coefficient have been considered in this paper. The term *baseline accuracy* is used here to denote the accuracy obtained when classifying all instances as belonging to the class with highest a priori value (i.e., the majority class).

Results provided by our CAC approach are shown in Table III, where each cell presents the accuracy rate (first value) and Kappa coefficient (second value) for different evaluation techniques (LOEOCV and LOPOCV), number of classes (8 and 3), input spaces for CSM featuring, and CSMs (NCD, NDD, FCD and WFCD). The classification scheme was $k$-NN voting, and parameter $k$ was tuned for the best accuracy according to the validation strategies in Section V-A. A wide range of $k$ values was explored for each classifier design ($k \in$[1,100]), selecting $k = 1$ for all cases in LOEOCV, and values around $k = 13$ for 8-class and $k = 5$ for 3-class in LOPOCV. Best performances are denoted in bold for every evaluation technique, number of classes, and quality indicator (accuracy rate and Kappa coefficient). The best accuracies were always for concatenated channels and WFCD, and among them, the worst accuracy was for the 8-class case with LOPOCV (62.67%), close to the baseline accuracy (60.06%, see Table II).

A key result from Table III is the dramatically different accuracy for LOEOCV and LOPOCV, the former being a much more optimistic scenario. This fact indicates that considering previous episodes from one patient when designing a classifier improves significantly the accuracy when another episode of the same patient is classified. Comparison of the accuracy rates obtained for WFCD-Ccat-8-class in LOEOCV (89.78%) and LOPOCV (62.67%) shows that LOPOCV accuracy is about 30 percentage points below that of LOEOCV. This reduction is not so dramatic for the 3-class case, which outperforms the baseline accuracy (64.06%) in 20 percentage points in spite of its reduction with respect to the LOEOCV technique. In general, results in Table III indicate that the 3-class scenario is much better than the 8-class case for all CSMs.

Regarding the Kappa coefficient, it allows us to interpret the reliability of the CAC schemes, specially for imbalanced datasets, by providing a numerical rating of the degree to which sometimes correct classification is just by chance (Kappa equal to 0). In this setting, it is possible to conclude from Table III that the accuracy provided by the LOEOCV technique is far from being by chance, because Kappa value is greater than 0.65 except when the FCD measure is used. The opposite conclusion could be inferred for the 8-class LOPOCV classification schemes, with Kappa values close to 0 and accuracy rates close to the baseline (60.05%), suggesting the strong relevance of imbalance in the CAC schemes. Note that Kappa values for 3-class LOPOCV classifiers were significantly higher than those for the 8-class schemes. The best case was 0.49 for WFCD, indicating that part of the classification structure is captured by the CAC scheme (accuracy 77.96%).

With respect to CSM featuring, the best results were obtained for channels concatenation, what may be reasonable since the use of just modulus and just phase has an implicit loss of information. Regarding the CSMs performance comparative, the proposed WFCD always provided the best results both in 8 and 3 classes scenarios, with concatenated channels. Table IV shows the confusion matrices for 8 classes (A) and 3 classes (B) when using the LOEOCV technique with WFCD. Penultimate row indicates the accuracy rate (in %) per class, while last row shows the total accuracy rate (see Table III). Since all the classifiers in LOEOCV corresponded to 1-NN,

TABLE III
ACCURACY RATE (FIRST VALUE, IN %) AND KAPPA COEFFICIENT (SECOND VALUE) IN EPISODES CLASSIFICATION (8-CLASS AND 3-CLASS CASES) WHEN USING CONCATENATED CHANNELS, MODULUS, PHASE, CONCATENATED MODULUS-PHASE WITH FOUR DIFFERENT CSMS (IN ROWS) AND TWO EVALUATION STRATEGIES (LOEOCV AND LOPOCV).

| | $k$-NN LOEOCV | | | | | | | | $k$-NN LOPOCV | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 8-class | | | | 3-class | | | | 8-class | | | | 3-class | | | |
| | Ccat | Mod | Pha | Xcat | Ccat | Mod | Pha | Xcat | Ccat | Mod | Pha | Xcat | Ccat | Mod | Pha | Xcat |
| **NCD** | 85.66<br>0.76 | 81.07<br>0.68 | 86.52<br>0.77 | 85.63<br>0.76 | 93.27<br>0.85 | 89.63<br>0.78 | 92.49<br>0.84 | 92.29<br>0.83 | 62.00<br>0.27 | 61.04<br>0.23 | 60.53<br>0.20 | 61.02<br>0.23 | 75.80<br>0.47 | 74.09<br>0.40 | 71.67<br>0.34 | 72.44<br>0.36 |
| **NDD** | 87.40<br>0.79 | 83.15<br>0.72 | 87.50<br>0.79 | 87.54<br>0.79 | 93.71<br>0.86 | 90.71<br>0.80 | 93.12<br>0.85 | 93.35<br>0.86 | 62.41<br>**0.28** | 61.04<br>0.25 | 60.79<br>0.20 | 62.22<br>0.26 | 77.12<br>0.48 | 74.93<br>0.42 | 71.27<br>0.33 | 74.06<br>0.38 |
| **FCD** | 34.32<br>0.11 | 36.74<br>0.08 | 42.36<br>0.18 | 35.41<br>0.12 | 57.35<br>0.22 | 58.31<br>0.20 | 61.78<br>0.28 | 62.59<br>0.29 | 40.73<br>0.08 | 48.51<br>0.02 | 47.56<br>0.07 | 35.71<br>0.03 | 48.93<br>0.15 | 48.29<br>0.07 | 55.25<br>0.12 | 53.46<br>0.15 |
| **WFCD** | **89.78**<br>**0.83** | 85.32<br>0.75 | 89.54<br>0.82 | 89.71<br>0.82 | **95.43**<br>**0.90** | 92.35<br>0.84 | 94.62<br>0.88 | 95.33<br>0.89 | **62.67**<br>0.27 | 61.36<br>0.22 | 60.24<br>0.19 | 61.25<br>0.23 | **77.96**<br>**0.49** | 74.05<br>0.39 | 70.14<br>0.30 | 73.47<br>0.37 |

TABLE IV
CONFUSION MATRIX USING THE LOEOCV STRATEGY, CCAT AND WFCD IN (A) 8-CLASS AND (B) 3-CLASS CLASSIFICATION SCENARIOS.

| | Actual | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **ST** | **AF** | **SVT** | **UST** | **SMVT** | **VF** | **TWO** | **NS** |
| **ST** | **855** | 6 | 12 | 39 | 16 | 1 | 1 | 0 |
| **AF** | 6 | **663** | 35 | 15 | 52 | 4 | 2 | 2 |
| **SVT** | 24 | 27 | **216** | 5 | 22 | 4 | 1 | 0 |
| **UST** | 42 | 17 | 7 | **195** | 9 | 1 | 0 | 1 |
| **SMVT** | 26 | 74 | 36 | 17 | **3925** | 69 | 9 | 3 |
| **VF** | 1 | 14 | 1 | 3 | 83 | **195** | 0 | 2 |
| **TWO** | 2 | 2 | 0 | 1 | 4 | 0 | **69** | 0 |
| **NS** | 0 | 0 | 0 | 0 | 2 | 0 | 0 | **30** |
| **Class accur.** | 89.4 | 82.6 | 70.4 | 70.9 | 95.4 | 71.2 | 84.1 | 78.9 |
| **Accur.** | 89.78 | | | | | | | |

(A)

| | Actual | | |
|---|---|---|---|
| | **Atrial** | **Ventricular** | **Other** |
| **Atrial** | **2164** | 109 | 7 |
| **Ventricular** | 172 | **4272** | 14 |
| **Other** | 5 | 6 | **99** |
| **Class accur.** | 92.44 | 97.38 | 82.50 |
| **Total accur.** | 95.43 | | |

(B)

the confusion matrix for 3 classes can be directly obtained by aggregating results of the confusion matrix for 8 classes. In these tables, the number of episodes correctly classified for each class is shown in bold (main diagonal). Confusion matrices are informative about the clinical performance and behavior of the classifier. For instance, note that the number of misclassifications among atrial rhythms in Table IV (A) is in general larger than in the others. Also, the absolute number of misclassifications between SMVT and VF is relevant, however, one has to keep in mind that there is not a widespread used cut-off criterion for discriminating both arrhythmias. The capabilities for noise identification are extremely high and reliable, with a very low number of misclassifications. With respect to the 3-class discrimination scheme, still a number of errors are present among atrial and ventricular origin arrhythmias.

From a clinical and technical point of view, the higher performance of the LOEOCV validation strategy is in accordance with the fact that a patient has individual physiopathological mechanisms characterizing his/her EGM episodes. Thus, when a patient suffers from a specific type of arrhythmia, it is likely that the patient will suffer again from the same type of arrhythmia. Then, 1-NN will likely be among the best $k$-NN classifier because other episodes of the same patient have been considered for the classifier design. On the other hand, the same patient may have overlapped segments between contiguous episodes, often for patients suffering from an arrhythmic storm (those segments seldom exceed 2 s). Accordingly, an additional experiment was carried out to check the accuracy bias when episodes with shared segments are considered in the classifier design. For this purpose, the LOEOCV accuracy was achieved by discarding from the classifier design (in each realization) those episodes of the same patient in a 6, 12, and 24 hour window. Ccat CSM features and WFCD were used. For the 8-class cases, accuracy(%)/kappa was 84.11/0.73, 83.85/0.73 and 83.88/0.72 for 6, 12 and 24 hour window, respectively. For the 3-class cases, results were 91.91/0.82, 91.77/0.82 and 91.59/0.82 for 6, 12 and 24 hour window, respectively, showing a moderate and acceptable bias due to this effect. These results also stress the importance of considering patient prior information in the classifier design, which significantly improve the performance.

It is interesting to compare the performance of the proposed compression-based with morphology-based discrimination algorithms as the Wavelet Algorithm (WA) [24], which is implemented in several ICD models in SCOOP database. The activation of WA depends on the clinician criteria, and different settings can be adjusted for its programming. WA only discriminates between the supraventricular or ventricular anatomical origin of the arrhythmia, and its automatic diagnosis is not always available for all episodes, given that cycle-based discrimination algorithms could inhibit it. For these reasons, the comparison between our CSM method and morphology discriminators used by ICDs is only possible when WA was activated in ICDs, in total 2837 episodes of the SCOOP database. For this subset, the EGMs in SCOOP database were retrieved from 4 different ICD models. We compared the ICD labels with those from the experts, obtaining a 76.91% of concordance with a kappa value of 0.44. The same comparison was made using the labels provided by our algorithm with the LOEOCV strategy with those of the experts, yielded a 95.28% of concordance with kappa of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



0.9. These results evidence that the discrimination capabilities increase with the proposed learning system.

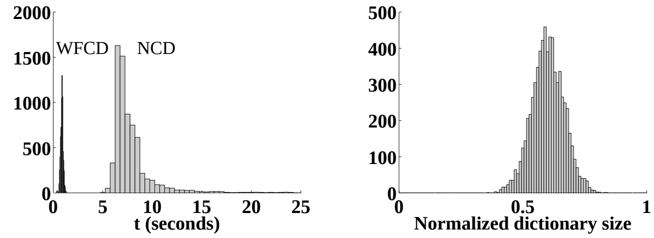
### *B. Compression Performance*

A relevant requirement in a cloud-based Big-Data service is the computational burden. A service of this nature handles a huge amount of data, hence its tools should be as efficient and simple as possible in order to avoid an increase in the response time. We benchmarked the computational burden of NCD (the most used CSM in the literature) and WFCD. When computing CSMs, the computational burden depends on the number of processors, the type of compressor, the strings length, and the number of similarities to be calculated (the number of strings in the dataset). The last one is highly relevant because the number of operations for NCD calculation is different from that for WFCD calculation, as shown next.

Let us consider the NCD computation between bits string $\boldsymbol{x}$ and a set of $N$ strings given by $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$. Let us also assume that we have previously compressed each string in $\boldsymbol{Y}$, storing the values $C_{\boldsymbol{y}_1}, C_{\boldsymbol{y}_2}, \cdots, C_{\boldsymbol{y}_N}$. As indicated in Eq. (3), apart from $C_{\boldsymbol{x}}$ it is always necessary to calculate $C_{\boldsymbol{x}\boldsymbol{y}_1}, C_{\boldsymbol{x}\boldsymbol{y}_2}, \cdots, C_{\boldsymbol{x}\boldsymbol{y}_N}$, what requires a high computational cost because the compression task is computationally expensive.

For comparison purposes, let us consider now the computational time used to get the WFCD and NCD between one string $\boldsymbol{x}$ and the set of $N$ strings $\boldsymbol{Y}$. If after compression we store the set of words $\{W_{\boldsymbol{y}_1}, W_{\boldsymbol{y}_2}, \cdots, W_{\boldsymbol{y}_N}\}$ and corresponding cardinalities, it is just necessary to perform the compression of $\boldsymbol{x}$. In this case, the highest computational cost is associated to that of obtaining the sets $W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}_1}, \cdots, W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}_N}$. Since getting the common words of two dictionaries (match operation) is faster than the compression task, it is clear that WFCD is computationally lighter than NCD. For visual illustration, Fig. 5a shows the histograms of the time spent in computing the similarity between each episode and the rest of the SCOOP dataset for WFCD and NCD using concatenated channels and the LZW compressor on an Intel Core i7 CPU Q 720 @1.60GHz. The LZW compressor was configured to find words of bytes, words with size 1 were removed (LZW algorithm initializes the dictionary with all possible words with size 1) and no maximum word-size was defined. Note that WFCD computation is about one order of magnitude quicker than that of NCD (means were 7.22 s and 0.89 s in NCD and WFCD, respectively).

Another important performance parameter to analyze is the episode compression rate, which can be represented by means of the dictionary size of each episode when normalized to the corresponding length (number of episode samples). Fig. 5b shows the distribution of these values, resulting in a Gaussian distribution with mean 0.6 and standard deviation 0.06. A compression rate of 0.6 (mean of the distribution) points out that LZW only can compress the episode information with a 40% mean rate, what indicates that episodes present high variability and that LZW does not find long patterns (the median word size in dictionaries was 3). However, and as shown by the classification performance of our CSM method, this fact is supportive for our goal, since it allows to characterize episodes according to its class and to correctly classify a high number of episodes.




(a) (b)

Fig. 5. Compression performance: (a) Histograms of the time (seconds) spent to compute NCD and WFCD measures between one episode (Ccat) and the rest of episodes in the SCOOP dataset; (b) Histogram of the WFCD dictionary size, normalized by the episode length (number of samples).

## VII. Discussion and Conclusion

CAC is still an open and active research field, with memory and computational constraints by the ICDs requiring low complexity arrhythmia discrimination algorithms. Simple criteria as cardiac cycle length or RR intervals have limitations when discriminating SVT from VT, and alternative algorithms have been proposed, as the EGM width criterion, which present limitations for wide QRS complex SVT [25], or morphological methods, such as the Correlation Waveform Analysis or the Probability Density Function, which do not improve the classification accuracy in current devices [4] and often they have not been validated in large databases. Similarity measures based on Information Theory have been successfully applied to a high number of applications, including language tree construction [5], content-based image retrieval [6], or weaning outcome prediction [10], among many others. Our CSM method aims to open an alternative approach to CAC, from an off-line scenario of Big Data Services, by pursuing the following algorithmic properties: (1) minimum signal preprocessing; (2) minimum number of free parameters; (3) possible comparison among signals with different lengths or sampling frequencies, and even when blank intervals are present; and (4) no need for feature selection or extraction stages.

When working with large databases from patients, it seems reasonable to develop automatic, fast enough and reliable procedures capable of learning the diagnosis provided by the expert. A usual classification requirement is a reasonable trade-off between reliability and efficiency. In the case of SCOOP, there is a need for continuous clinical evaluation of device-stored episodes by physicians, since the ICDs do not provide an accurate discrimination for all of them. This has special relevance in clinical studies, where automatic classification is required to have an accurate idea of the impact of incorrect episode classifications in terms of inappropriate therapies. In addition, a subsequent clinical evaluation of the episode could help in the patient follow-up for automatic discrimination and fast communication to the clinician about the actual patient's episodes and the therapies which could have been inappropriately delivered. Due to the observed high variability and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



classification difficulty, each episode in SCOOP database was analyzed by at least 2 expert reviewers and labelled only after expert consensus. However, previous research has shown that there is a wide diversity of visual evaluation methodologies in clinical studies [26], [27], and its rigor is strongly tied to the available time and resources. In addition, there exists a bias in the expert classification that can be linked to the inter- and intra-physician variability, which generates a systematic error which is also dependent on the ICD model [26], [27].

The high accuracy required for ICD episodes classification needs to be analyzed in the context of the scope of existing and ongoing algorithms for current devices. In [4], performance was benchmarked for VT vs SVT+ST discrimination in bicameral devices in four detection algorithms, yielding 98% sensitivity for VT detection with specificity from 66% to 94%, whereas previous studies [28] had benchmarked the same algorithms in a similar dataset (consisting of 71 SVT and 15 VT) and yielded extremely low specificities (11%, 12%, 20%, and 28%). Recently, the preliminary results of the PainFREE SST study [29] showed SmartShock algorithm providing with an impressively low inappropriate shock rate, namely, larger than 97% in different device models. Note that inappropriate shock reduction is associated with improved detection, but this does not mean that detection is completely accurate, as far as inappropriate shocks can be avoided by changing the device detection criteria.

SCOOP is currently a national open project created and running in Spain, so that any Spanish hospital can join and participate on it at any time, and there is a possibility to its expansion in an international environment. Since the fast growing and development of cloud computing systems and Big Data Analytics are changing the basis of data knowledge extraction, they are increasingly present in everyday life of companies and organizations, and platforms like SCOOP can offer great possibilities to physicians for further analyzing and supporting their patient's diagnosis and treatment, as well as provide with new scenarios for developing new algorithms and tools for the daily clinical practice.

The proposed methodology would benefit from further development for its advantage use in Cloud-based Big-Data services, since it has some limitations. One of them is the presence of a bias in the predominant SVMT class detection due to unbalanced classes. Moreover, an additional bias is also present due to arrhythmic episodes of the same patient near in time, which often exhibit similar features and morphology. In our work, the use of machine learning techniques was constrained to explore the most simple and effective ones, however, other probabilistic machine learning schemes considering prior probabilities and weighted costs could probably improve the CAC performance reducing the aforementioned bias. Another limitation is the WFCD dependence on both sampling rate and amplitude resolution, which is a crucial factor for the use of the proposed approach in CAC databases with mixed sampling rates and amplitude resolutions. In this work, this dependence has not been further scrutinized because in SCOOP database EGMs were always stored by ICDs with the restriction of a fixed sampling rate and amplitude resolution (128 Hz and 0.063 mV, respectively).

The classification algorithm proposed in this work certainly should be required to provide not only with the estimated classification label, but also with the estimated confidence on this classification, hence it could be used for retrieving only those highly reliable labels. This ongoing theoretical and practical study is at this moment beyond the scope of the present work. Within the current SCOOP framework, several advantaged uses can be designed for the algorithm, which do not consist of the substitution of the scientific committee for labelling, but rather of running parallel in collaboration with some committee members.

The potential of the algorithm for large databases labelling can include other clinical studies different from SCOOP, and the algorithm could also be used in clinical studies using devices from other companies. According to current database size, and only looking at those arrhythmias with better algorithm recognition rate for the 3-class scenario (ventricular), our approach could be used for more than 50% of the episodes with a classification error of about 2.5%. This is an acceptable rate for an environment like SCOOP, where scientific clinicians are expected to review more than 60.000 episodes in the near future. Our methodology could have potential advantage as an intelligent tool for existing remote monitoring systems, by providing with priority when many episodes are pending of cardiologist evaluation, or for suggesting those episodes with discordance between device and algorithm classification for preferential analysis, or for preselecting actual ventricular episodes from inappropriate therapy episodes and sending alert messages.

## VIII. ACKNOWLEDGMENT

Authors want to thank María García de Pablo, Rafael Moreno, Germán Gutiérrez, and the Team at Medtronic Ibérica® S.A. for their kind and professional support throughout all the project.

Special thanks to the researchers and contributors in SCOOP: Miguel Ahumada (Hospital General Universitario de Elche), Miguel Álvarez (Hospital Universitario Virgen de las Nieves), Javier Alzueta (Hospital Clínico Universitario Virgen de la Victoria), María Fe Arcocha (Hospital Universitario de Basurto), Ángel Arenal (Hospital General Universitario Gregorio Marañón), José María Arizón (Hospital Universitario Reina Sofía de Córdoba), Javier Balaguer (Hospital Universitario de Guadalajara), Nuria Basterra (Complejo Hospitalario de Navarra), Juan Benezet (Hospital General Ciudad Real), Andrés Bodegas (Hospital Universitario de Cruces), Josep Brugada (Hospital Clinic i Provincial de Barcelona), David Calvo (Hospital Universitario Central de Asturias), Ernesto Díaz Infante (Hospital Universitario Virgen Macarena), Joaquín Fernández de la Concha (Hospital Universitario Infanta Cristina de Badajoz), Ignacio Fernández Lozano (Hospital Universitario Puerta de Hierro Majadahonda), María Luisa Fidalgo Andrés (Complejo asistencial Universitario de León), Adolfo Fontenla Fontenla (Hospital Universitario 12 de Octubre), Xavier Fosch (Hospital Son Llatzer), Javier Garcia (Hospital Universitario de Burgos), Enrique García Campo (Complejo Hospitalario Universitario de Vigo), Antonio Hernández Madrid (Hospital Universitario Ramón y Cajal), Javier Jiménez (Hospital de la Ribera), Juan Gabriel Martínez (Hospital General Universitario de Alicante), José Martínez Ferrer (Hospital Universitario de Áraba), Jordi Mercé (Hospital Universitario Joan XXIII de Tarragona), José Moreno Arribas (Hospital Universitario San Juan de Alicante), Roberto Muñoz Aguilera (Hospital Universitario Infanta Leonor), Juanjo Olalla (Hospital Universitario Marqués de Valdecilla), Ricardo Pavón (Hospital Universitario Nuestra Señora de Valme), Rafael Peinado (Hospital Universitario de La Paz), Luisa Pérez (Complejo Hospitalario Universitario de A Coruña), Fernando Pérez Lorente (Hospital General Univesitario Reina Sofía Murcia), Jose María Porres (Hospital Universitario Donostia), Rosa Porro (Complejo Hospitalario de Cáceres), Aurelio Quesada (Consorcio Hospital General Universitario de Valencia), Aníbal Rodríguez (Hospital Universitario de Canarias), Juan Carlos Rodríguez (Hospital Universitario Insular de Gran Canaria), Rafael Romero (Complejo Hospitalario Nuestra Señora de la Candelaria), Jerónimo Rubio (Hospital Clínico Universitario de Valladolid), Ricardo Ruiz (Hospital Clínico Universitario de Valencia), Xavier Sabaté (Hospital Universitario de Bellvitge), María José Sancho Tello (Hospital Universitario La Fe), Francisco José Tornés (Complejo Hospitalario Torrecárdenas), Julián Villacastín (Hospital Clínico Universitario San Carlos), Roger Villuendas (Hospital Universitario Germans Trias i Pujol), Xavier Viñolas (Hospital de la Santa Creu i Sant Pau) and Francisco Zumalde (Hospital de Galdakao-Usansolo).

This work has been partly supported by TEC2010-19263 and TEC2013-48439-C4-1-R projects from Spanish Government, by PRIN13_IYA12 project from Rey Juan Carlos University, and by Prometeo Project of the Secretariat for Higher Education, Science, Technology and Innovation of the Republic of Ecuador. RSM is supported by Juan de la Cierva Program of the Spanish Ministry of Science and Innovation (JCI-2011-11150). JMLC is supported by the Spanish FPU grant FPU13/03134.

## REFERENCES


[1] World Health Organization. (2011) The top 10 causes of death (Fact Sheet no. 310). [Online]. Available: http://www.who.int/mediacentre/factsheets/fs310/en/

[2] T. B. García and G. T. Miller, *Arrhythmia Recognition: The Art of Interpretation*, 1st ed. Jones & Bartlett Learning, 2004.

[3] R. X. Stroobandt, S. S. Barold, and A. F. Sinnaeve, *Implantable Cardioverter - Defibrillators Step by Step: An Illustrated Guide*, 1st ed. Wiley-Blackwell, 2009.

[4] E. Aliot, R. Nitzschéb, and A. Ripartb, "Arrhythmia detection by dual-chamber implantable cardioverter defibrillators. a review of current algorithms," *Europace, Elsevier*, vol. 6, no. 4, pp. 273–286, 2004.

[5] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi, "The similarity metric," *IEEE Trans Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[6] D. Cerra and M. Datcu, "A fast compression-based similarity measure with applications to content-based image retrieval," *J Visual Communication and Image Representation*, vol. 23, no. 2, pp. 293–302, 2012.

[7] D. Salomon, *Data Compression. The Complete Reference*, 4th ed. Springer, 2007.

[8] P. Vitanyi, "Compression-based similarity," in *Proc. of the 1st Intl Conf on Data Compression, Communications and Processing*, Palinuro, Cilento Coast, Italy, 21-24 June 2011, pp. 111–118.

[9] R. Cilibrasi and P. Vitanyi, "Clustering by compression," *IEEE Trans on Information Theory*, vol. 51, no. 4, pp. 1523–1545, 2005.

[10] J. M. Lillo-Castellano, I. Mora-Jiménez, R. Santiago-Mozos, J. L. Rojo-Álvarez, J. Ramiro-Bargueño, and A. Algora-Weber, "Weaning outcome prediction from heterogeneous time series using normalized compression distance and multidimensional scaling," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1737–1747, 2013.

[11] D. Cerra, M. Datcu, and P. Reinartz, "Authorship analysis based on data compression," *Pattern Recognition Letters*, vol. 42, pp. 79–84, 2014.

[12] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, 3rd ed. Springer, 2008.

[13] C. H. Bennett, P. Gács, M. Li, P. Vitányi, and W. Zurek, "Information distance," *IEEE Trans on Information Theory*, vol. 40, no. 4, pp. 1407–1423, 1998.

[14] D. Salomon, *Variable-length Codes for Data Compression*, 1st ed. Springer Science & Business Media, 2007.

[15] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, 1st ed. Springer, 1992.

[16] A. Macedonas, D. Besiris, G. Economou, and S. Fotopoulos, "Dictionary based color image retrieval," *Journal of Visual Communication and Image Representation*, vol. 19, no. 7, pp. 464–470, 2008.

[17] A. Fontenla, M. López-Gil, J. Martínez-Ferrer, J. Alzueta, I. Fernández-Lozano, X. Viñolas, A. Rodríguez, J. Fernández-de-la Concha, I. Anguera, and F. Arribas, "Clinical profile and incidence of ventricular arrhythmia in patients undergoing defibrillator generator replacement in Spain," *Revista Española de Cardiología (English Edition)*, 2014, in press.

[18] G. Marsaglia, W. Tsang, and J. Wang, "Evaluating kolmogorov's distribution," *Journal of Statistical Software*, vol. 8, no. 18, 2003.

[19] J. L. Rojo-Álvarez, A. Arenal-Maíz, and A. Artés-Rodríguez, "Discriminating between supraventricular and ventricular tachycardias from egm onset analysis," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 1, pp. 16–26, 2002.

[20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley–Interscience, 2001.

[21] E. Wegert, *Visual Complex Functions: An Introduction with Phase Portraits*, 1st ed. Springer Science & Business Media, 2012.

[22] A. Casaleggio, P. Rossi, A. Dall'Acqua, A. Faini, G. Sartori, G. Musso, R. Mureddu, E. Casali, V. Malavasi, and S. Chierchia, "On cardiac activity characterization from implantable cardioverter defibrillator electrogram analysis: is far-field better?" in *Computers in Cardiology*, Memphis, USA, 22-25 Sept. 2002, pp. 161–164.

[23] C. Aggarwal, *Data Classification: Algorithms and Applications*, 1st ed. Chapman & Hall, 2014.

[24] C. Swerdlow, M. Brown, K. Lurie, J. Zhang, N. Wood, W. Olson, and J. Gillberg, "Discrimination of ventricular tachycardia from supraventricular tachycardia by a downloaded wavelet-transform morphology algorithm: A paradigm for development of implantable cardioverter defibrillator detection algorithms," *Journal of Cardiovascular Electrophysiology*, vol. 13, no. 5, pp. 432–441, 2002.

[25] F. Duru, M. Schönbeck, T. F. Lüscher, and R. Candinas, "The potential for inappropriate ventricular tachycardia confirmation using the intracardiac electrogram (egm) width criterion," *Pacing and Clinical Electrophysiology*, vol. 22, no. 7, pp. 1039–1046, 1999.

[26] P. Friedman, R. McClelland, W. Bamlet, H. Acosta, D. Kessler, T. Munger, N. Kavesh, M. Wood, E. Daoud, A. Massumi, C. Schuger, S. Shorofsky, B. Wilkoff, and M. Glikson, "Dual-chamber versus single-chamber detection enhancements for implantable defibrillator rhythm diagnosis: The detect supraventricular tachycardia study," *Circulation*, vol. 113, no. 25, pp. 2871–2879, 2006.

[27] B. Powell, Y. Cha, S. Asirvatham, D. Cesario, M. Cao, P. Jones, M. Seth, L. Saxon, and F. Roosevelt, "Implantable cardioverter defibrillator electrogram adjudication for device registries: Methodology and observations from altitude," *Pacing and Clinical Electrophysiology*, vol. 34, no. 8, pp. 1003–1012, 2011.

[28] F. Hintringer, S. Schwarzacher, G. Eibl, and O. Pachinger, "Inappropriate detection of supraventricular arrhythmias by implantable dual chamber defibrillators: A comparison of four different algorithms," *Pacing and Clinical Electrophysiology*, vol. 24, no. 5, pp. 835–841, 2001.

[29] C. Wollmann, T. Lawo, V. Kühlkamp, R. Becker, C. Garutti, T. Jackson, M. Brown, and H. Mayr, "Implantable defibrillators with enhanced detection algorithms: Detection performance and safety results from the painfree sst study," *Pacing and Clinical Electrophysiology*, vol. 37, no. 9, pp. 1198–1209, 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2412175, IEEE Journal of Biomedical and Health Informatics



**J.M. Lillo-Castellano** received the Telecommunication Engineering degree and the MSc in Information and Communication Technology in Biomedical Engineering from the Rey Juan Carlos University (Madrid, Spain) in 2010 and 2012, respectively. Currently, he is a PhD Candidate and working as hired researcher at the Department of Signal Theory and Communications, Telematics and Computing, Rey Juan Carlos University. His research interests include Multivariate Data Analysis, Machine Learning, Digital Image Processing, and their applications to Bioengineering.

**I. Mora-Jiménez** received the Telecommunication Engineering degree in 1998 from the Polytechnic University of Valencia (Spain), and the PhD degree in Telecommunication in 2004 from Carlos III University of Madrid (Spain). Currently, she is an associate professor in the Department of Signal Theory and Communications, Telematics and Computing at Rey Juan Carlos University (Madrid, Spain). Her main research interests include Statistical Learning Theory, Neural Networks, and their applications to Image Processing, Bioengineering, and Communications.

**R. Santiago-Mozos** received the Ph.D. degree in signal processing and communications from the Carlos III University of Madrid, Madrid, Spain, in 2009. He has just finished a Postdoctoral Fellow of the Spanish Program for Recruitment and Incorporation of Human Resources at the University Rey Juan Carlos of Madrid, Madrid, Spain. He has been a Spanish Foundation for Science and Technology Postdoctoral Fellow at the Machine Learning and Data Mining Group in National University of Ireland, Galway, Ireland, where he had also completed a Marie Curie Fellowship. His research interests include machine learning and medical applications and he has coauthored more than 20 papers appearing in refereed journals and conference proceedings.

**F. Chavarría-Asso** received the Computer Science Engineering degree in 2009 from the High Polytechnical Center in the University of Zaragoza (Spain). Currently, he is working in the area of implantable cardiac devices in the company Medtronic Ibérica® (Madrid, Spain). He has worked in the concept and development of several projects and services involving device data mining and management of patient information gathered through the Medtronic CareLink® remote monitoring network.

**A. Cano-González** received the Telecommunication Engineering degree and MsC in Telemedicine and Biomedical Engineering in 2006 and 2008 from the Polytechnic University of Madrid (Spain). Currently, she is working in the area of implantable cardiac devices in the company Medtronic Ibérica (Spain). She works as a project manager of an innovative project (SCOOP Project) focused on encouraging the design and development clinical/scientific research to enhance clinical evidence generation of implantable cardiac devices.

**A. García-Alberola** received the MD (1982) and the PhD (1991) from Universitat de Valencia, Spain. Since 1993 he has been a Cardiologist and an Professor of Medicine at Hospital Universitario Virgen de la Arrixaca and Universidad de Murcia, where he is the Chief of the Laboratory of Cardiac Electrophysiology. He has co-authored more than 120 scientific papers and more than 50 communications in cardiac electrophysiology, and his main research areas are repolarization analysis, arrhythmia mechanisms, and cardiac signal processing

**J.L. Rojo-Álvarez** received the Telecommunication Engineering degree in 1996 from the University of Vigo (Spain), and the PhD degree in Telecommunication in 2000 from the Polytechnic University of Madrid (Spain). Since 2006, he has been an Associate Professor at the Department of Signal Theory and Communications, Telematics and Computing, Rey Juan Carlos University (Madrid, Spain). He has published more than 70 papers in JCR journals and more than 100 international conference communications. He has participated in more than 50 projects (with public and private funding), and directed more than 10 of them, including several actions in the National Plan for Research and Fundamental Science. He was awarded in 2009 with the I3 Prize of Spanish Science and Innovation Ministry to the research path.