Heartbeat Classification Using Feature Selection Driven by Database Generalization Criteria

Mariano Llamedo* and Juan Pablo Martínez

***Abstract*—In this paper, we studied and validated a simple heartbeat classifier based on ECG feature models selected with the focus on an improved generalization capability. We considered features from the RR series, as well as features computed from the ECG samples and different scales of the wavelet transform, at both available leads. The classification performance and generalization were studied using publicly available databases: the MIT-BIH Arrhythmia, the MIT-BIH Supraventricular Arrhythmia, and the St. Petersburg Institute of Cardiological Technics (INCART) databases. The Association for the Advancement of Medical Instrumentation recommendations for class labeling and results presentation were followed. A floating feature selection algorithm was used to obtain the best performing and generalizing models in the training and validation sets for different search configurations. The best model found comprehends eight features, was trained in a partition of the MIT-BIH Arrhythmia, and was evaluated in a completely disjoint partition of the same database. The results obtained were: global accuracy of 93%; for normal beats, sensitivity (S) 95%, positive predictive value (P^+) 98%; for supraventricular beats, S 77%, P^+ 39%; and for ventricular beats S 81%, P^+ 87%. In order to test the generalization capability, performance was also evaluated in the INCART, with results comparable to those obtained in the test set. This classifier model has fewer features and performs better than other state-of-the-art methods with results suggesting better generalization capability.**

***Index Terms*—Feature selection, heartbeat classification, linear classifier, wavelet transform (WT).**

I. Introduction

THE ANALYSIS of the ECG signal provides a noninvasive and inexpensive technique to analyze the heart function for different cardiac conditions. In the past decades, the computerized analysis of the ECG became a well-established practice, and many improvements were achieved to aid cardiologists in the task of analyzing long-term ECG recordings. One important analysis performed in the ECG is the classification of heartbeats, which is important for the study of arrhythmias. Arrhythmias are understood as any disturbance in the rate, regularity, site of origin, or conduction of the electrical impulses through the heart [1]. While some types of arrhythmias represent a life threat in the very short term (e.g., ventricular fibrillation), there are other types that appear less frequently and represent a long-term threat without proper treatment. It is in these later cases, which require carefully inspection of long-term ECG recordings, where the use of automatic algorithms represents an important help for the diagnostic.

Many algorithms for ECG classification were developed in the past decade [2]–[11], but only few of them have completely comparable methodologies, and therefore, results [4], [8], [10]. The Association for the Advancement of Medical Instrumentation (AAMI) recommendations [12] for class labeling and results presentation have eased this problem, and at this time, it is broadly accepted [4], [5], [8]–[11]. From the different classification approaches presented in these papers, some of them classify beats without any local expert (LE) assistance [2]–[4], [8], [10], but others take advantage from a LE to improve the classification performance [2], [3], [7], [8]. Regarding to the classes of interest, the AAMI recommendation suggests five classes: supraventricular (S), ventricular (V), fusion (F), beats that cannot be classified (Q), and normal (N) [12]. It is remarkable that all previous works were interested in discriminating between N and V classes, but only few of these works studied the multiclass classification problem [3], [4], [8], [10]. In terms of the data division, some works performed a beat-oriented division, no matter which subject the heartbeats belong to, so that both the training and testing datasets contain heartbeats from the same subjects [5], [9], [11]. It was shown in [4] that this approach leads to an optimistic bias of the results, being more advisable a patient-oriented division, as it will also happen in the application scenario, where the algorithm is to be used.

Concerning the features used for classification (classification model), the surrounding RR intervals were considered in almost all published works. Other choices were the decimated ECG samples (mostly from the QRS complex or T wave) [4], or transformed by Hermite polynomials [3], or wavelet decomposition (WT) [8]. Some works use features that integrate information present in both leads, like the vectocardiogram maximal vector (VCG_M) and angle (VCG_ϕ) [6]. Another multilead strategy can be seen in [4], where a final decision is taken from several posterior probabilities calculated from single-lead features. In the same work, features derived from the delineation of the ECG, like the QRS complex and T wave duration, proved to be useful for classification. In some works, where the

Manuscript received April 27, 2010; revised July 8, 2010; accepted July 26, 2010. Date of publication August 19, 2010; date of current version February 18, 2011. This work was supported by Ministerio de Ciencia e Innovación under Project TEC2010-21703-C03-02 and by Diputación General de Aragón, Spain under Project GTC T-30. The Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER) is an initiative of Instituto de Salud Carlos III. *Asterisk indicates corresponding author.*

*M. Llamedo was with the Electronic Department, National Technological University, C1179AAQ Buenos Aires, Argentina. He is now with the Communications Technology Group, Aragón Institute of Engineering Research, University of Zaragoza, 50018 Zaragoza, Spain, and also with the CIBER, Zaragoza 50018, Spain (e-mail: llamedom@electron.frba.utn.edu.ar).

J. P. Martínez is with the Communications Technology Group, Aragón Institute of Engineering Research, University of Zaragoza, 50018 Zaragoza, Spain, and also with the CIBER, Zaragoza 50018, Spain (e-mail: jpmart@unizar.es).

Digital Object Identifier 10.1109/TBME.2010.2068048

TABLE I
CLASS DISTRIBUTION OF THE DATABASES USED AND DIVISION OF THE MIT-BIH-AR DATABASE INTO TRAINING (DS1) AND TESTING (DS2) SETS

MIT-BIH-AR							
Dataset	Purpose	N	S	V	F	Q	#Rec
$DS1$	train	45784	940	3783	413	8	22
$DS2$	test	44188	1835	3218	388	7	22
Totals		89972	2775	7001	801	15	44

Other Databases							
Database	Purpose	N	S	V	F	Q	#Rec
MIT-BIH-SUP	validation	161902	12083	9897	193	78	78
INCART	test	153517	1958	19991	219	5	75

Dataset	MIT-BIH-AR recordings
$DS1$	101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230
$DS2$	100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234

Heartbeat classes are N: normal, S: supraventricular, V: ventricular and F: fusion.
Recordings with paced beats were excluded.

dimensionality of the problem was an issue, feature transformations like principal components analysis (PCA) were used to keep the dimension of the model as low as possible [11]. However, none of the reviewed papers considered the use of a feature selection algorithm to retain the most relevant features.

Several classifiers were adopted in the reviewed papers, from simple linear discriminant functions based on the Gaussian assumption of the data [4], [8] to more elaborated ones, as artificial neural networks (ANN's), self-organizing maps (SOM), and learning vector quantization (LVQ) among others [2], [3], [5], [6], [9]–[11].

The database used without exception by all groups was the MIT-BIH Arrhythmia database (MIT-BIH-AR) [13] for training and testing purposes. None of the reviewed papers reported the generalization properties of the proposed algorithms outside the MIT-BIH-AR database.

The objective of this paper is to develop and evaluate a heartbeat classification algorithm according to the following conditions.

1) Perform fully automatic ECG classification (without LE intervention).
2) Follow AAMI recommendations for class labeling and results presentation.
3) Use a simple classifier (as linear or quadratic discriminant functions) to ensure that the classification performance is due to the features selected.
4) Features should have a physiological meaning, being simple to compute and robust to the typical kind of noise present in the ECG.
5) Use a multidatabase validation approach for feature selection to ensure better generalization properties of the selected feature set.

II. METHODOLOGY

A. ECG Databases

In this paper, we used the well-known MIT-BIH-AR [13] for training and testing purposes. Additionally, the MIT-BIH Supraventricular Arrhythmia database (MIT-BIH-SUP) [14] and the St. Petersburg Institute of Cardiological Technics (INCART) database were used for evaluation and testing purposes, in order to assess the generalization achieved by the classification models developed in the MIT-BIH-AR. All databases are freely available on Physionet [15] and their details are summarized as following.

1) MIT-BIH Arrhythmia Database: The database consists of 48 two-lead recordings of approximately 30 min. and sampled at 360 Hz. The first 23 recordings were extracted from routine ambulatory practice, while the remaining 25 were selected because of the presence of less common complex ventricular, junctional, and supraventricular arrhythmias. The two recorded leads are not the same in all recordings, depending on the arrhythmia and physical limitation of the subject's body. The annotations provided with the database were used for training and testing purposes, following the recommendations and class labeling of AAMI ([12, Sec. IV-B] and [4, Tab. I]). We adopted the training (DS1) and test (DS2) set division scheme used in [4] for comparative purposes of the results. The four recordings with paced beats were discarded in this paper in accordance with AAMI [12]. The AAMI Q class (unclassified and paced heartbeats) was discarded, since it is marginally represented in the database. This limitation occurs to a lesser extent with the fusion (F) AAMI class, but instead of discarding the heartbeats of this class, a class-labeling modification to the AAMI recommendation is proposed here and was adopted. It consists in merging fusion (of normal and ventricular beats) and ventricular classes, as the same ventricular class (V'). We will refer to this modification as AAMI2 labeling. The division scheme is summarized in Table I.

2) MIT-BIH Supraventricular Arrhythmia Database: The database consists of 78 two-lead recordings of approximately 30 min. and sampled at 128 Hz. The recordings were chosen to supplement the examples of supraventricular arrhythmias in the MIT-BIH-AR. The annotations of the recordings were first automatically performed, by the Marquette Electronics 8000 Holter scanner, and later, reviewed and corrected by a medical student [16]. The original labeling was also adapted to the AAMI recommendations and to the AAMI2 modification. This database will be considered for validation and model selection purposes. The class distribution is shown in Table I.

3) INCART 12-Lead Arrhythmia Database: This database consists of 75 annotated recordings extracted from 32 Holter records. Each record is 30 min. long and contains 12 standard leads, each sampled at 257 Hz. The annotations were produced by an automatic algorithm, and then, corrected manually, containing over 175 000 beat annotations in all. The original records were collected from patients undergoing tests for coronary artery disease (17 men and 15 women, aged 18–80, and mean age 58). None of the patients had pacemakers; most had ventricular ectopic beats. In selecting records to be included in the database,

preference was given to subjects with ECG's consistent with ischemia, coronary artery disease, conduction abnormalities, and arrhythmias. From the 12 standard leads, the two more frequent leads in the MIT-BIH-AR database (lead II and V1) were selected to perform the experiments presented in this paper. This database will be considered only for testing purposes. More details about the database are shown in Table I.

B. Signal Processing

The ECG recordings of the MIT-BIH-SUP and INCART databases were first resampled to 360 Hz, which is the sampling frequency of the MIT-BIH-AR. This was performed with a tenth-order low-pass finite-impulse response (FIR) filter without observing any notorious distortion (resample function, Signal Processing Toolbox of MATLAB, The Mathworks Inc., Natick, MA). All recordings in all databases were first preprocessed to remove artifacts as described in [4]. No energy or amplitude normalization was done, as we were interested in some amplitude-related features.

1) Wavelet Transform: Many of the considered features (explained in following sections) were based on the wavelet transform (WT) of the ECG signal. The WT is defined for a continuous signal $s(t)$ as follows:

$$W_s s(b) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} s(t)\psi\left(\frac{t-b}{s}\right) dt,\ s > 0. \qquad (1)$$

This transformation maps the ECG signal into a time-scale plane (understanding scale as a surrogate of frequency). The responsible of the mapping is the prototype-wavelet function $\psi(t)$, affected by both scaling and translation parameters s and b, respectively. The WT allows to locate details or fast transitions when scale parameter s is small, and coarser aspects or trends for higher values. The translation parameter b indicates the location of these finer or coarser details. As this continuous representation is computationally unfeasible, a typical choice is to discretize the time-scale map using a dyadic sampling, where $s = 2^k$ and $b = 2^k l$ for $k,\ l \in Z$, resulting the discrete WT (DWT). By using this restriction, lower scales have greater sampling frequency than higher scales. But as in our application, we are interested in keeping the time accuracy as high as possible (at the expense of redundancy), we relax the restriction to $b = l$ for $l \in Z$, resulting in a time-scale plane with the same sampling rate at each scale (algorithme á trous). It is worth to mention that the DWT can be efficiently implemented as a filter bank. We used the derivative of a smoothing function (quadratic spline) as the prototype wavelet $\psi(t)$, resulting the different scales of the DWT as a smoothed derivative of the ECG. As a result, the DWT retains at certain scales, the useful information present in the ECG in form of absolute maxima and zero crossings (as we will see later in Fig. 2). For background and implementation details, the interested reader is referred to [17] for a more detailed description of the WT and its implementation for ECG delineation. Following the conclusions of [17], the resulting DWT framework allows an analysis robust to the typical interferences present in routine ECG recordings, so the features derived from the DWT are expected to inherit this desirable property.

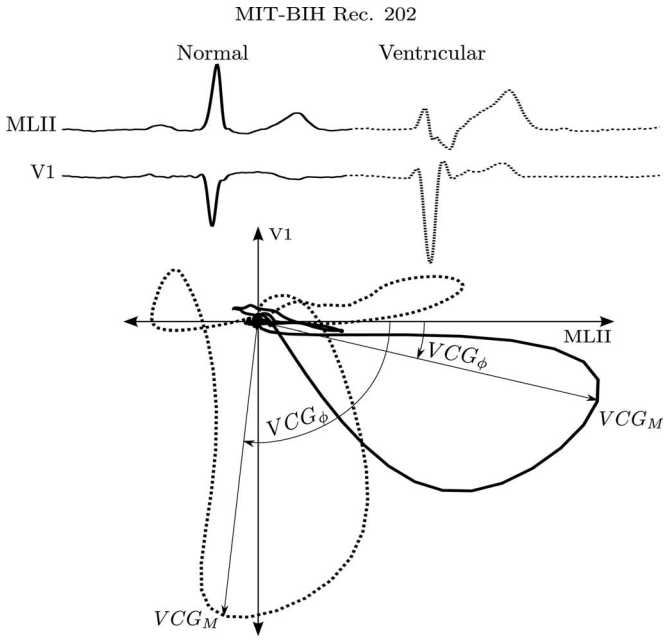

Fig. 1. Illustration of the features calculated from the VCG loop computed with the two available leads, for a normal (continuous line) and ventricular (dotted line) beats. The maximum value of the loop and the angle at this point are shown.

2) ECG Delineation: As our objective is the evaluation of a heartbeat classifier, the QRS location is assumed to be known and we use the annotations included in the databases. Following the QRS complex detection positions, the delineation of each heartbeat was performed with the delineator described in [17]. Both the delineation result and the DWT of the ECG signals (which are intermediate signals for the delineator) were used to calculate some features described in the following sections.

C. Classification Features

Following the conclusions of previous works [2], [4], we included in our model both interval and morphological features. As interval features, we used features from the RR sequence $\mathrm{RR}[i-1]$, $\mathrm{RR}[i]$, and $\mathrm{RR}[i+1]$ to describe the local-time evolution of the heart rhythm. In order to assess the local variation of the heart rhythm, the feature $\mathrm{RR}_V[i] = \sum_{j=-1}^{1} |d\mathrm{RR}[i-j]|$ (being $d\mathrm{RR}[i] = \mathrm{RR}[i] - \mathrm{RR}[i-1]$) characterizes the variation in the surrounding heartbeats. We also included estimates of the local and global rhythm by the mean RR interval in the last 1, 5, 10, and 20 min. (RR_P being $P \in \{1, 5, 10, 20\}$, the interval in minutes of aggregation).

The morphological features used can be grouped in three categories depending on whether they were calculated in the ECG signal, the 2-D VCG loop formed by both available leads or in the DWT of the ECG signal.

1) The QRS width ($\mathrm{QRS}_W = \mathrm{QRS}_{\mathrm{OFF}} - \mathrm{QRS}_{\mathrm{ON}}$) is obtained from the delineation of the ECG.
2) From the 2-D VCG loop constructed with the two available leads, we calculated two features: the maximal vector of the QRS loop (VCG_M) and the angle of this vector (VCG_ϕ), as shown in Fig. 1.

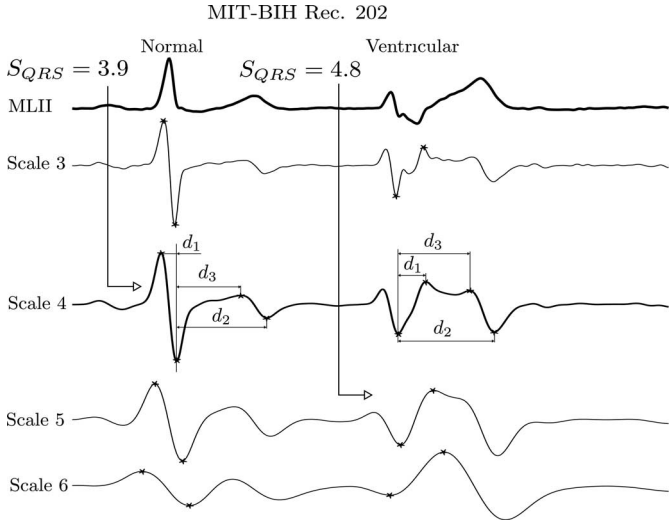

Fig. 2. Illustration of the features calculated from the wavelet transform for the same normal and ventricular beat in Fig. 1. The two most important peaks from the QRS complex and T wave are indicated with an asterisk, and the relative distances (d_i) to the most important peak in the fourth scale. Also, the scale, where the QRS complex is centered (S^L_{QRS}) is shown for both types of heartbeats used for its calculation (only for one lead).

3) Regarding the features calculated from the DWT of the ECG, three types can be defined.
   a) The first type includes seven features (per lead) that were calculated from peak amplitudes and positions from the fourth scale of the DWT ($W_4 s(l)$), since this scale (between 12.25–22.5 Hz) has good projection of the ECG information. These seven features are the two greatest absolute values of the QRS complex, the two greatest absolute values of the T wave, and their three relative positions (to the position of the greatest peak in the heartbeat, see Fig. 2).
   b) The second type is also calculated from the fourth scale of the DWT. The autocorrelation signal for both leads ($r_x(k)$ and $r_y(k)$) and the interlead cross-correlation signal ($r_{xy}(k)$) were calculated within a time window, which starts 130 ms before the fiducial point and ends 200 ms after. One remarkable aspect is that features calculated from the correlation signals will essentially be synchronized in time, even if the fiducial point is not accurately determined. We calculated for the three signals, the location and value of the absolute maximum, and for r_x and r_y, the location of the first zero crossing, as shown in Fig. 3.
   c) The feature is the wavelet scale, where the QRS complex is centered for each lead, since fast evolving signals (like a normal beat) tend to be centered in lower wavelet scales (higher frequency content). The QRS center scale for each lead ($S^{\mathrm{Lead}}_{\mathrm{QRS}}$) is calculated as the weighted sum

$$S^L_{\mathrm{QRS}} = \frac{\sum_{s=1}^{6} A^L_s s}{\sum_{s=1}^{6} A^L_s} \tag{2}$$

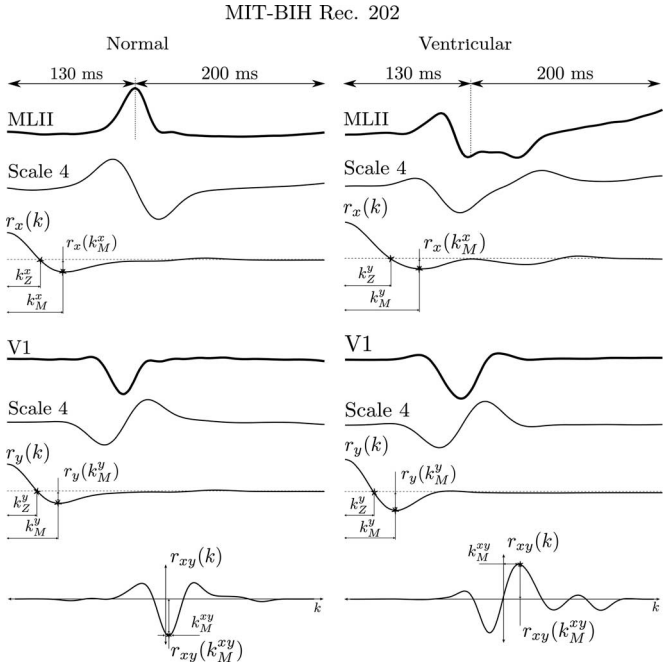

Fig. 3. Illustration of the features calculated from the wavelet correlation signals for the same normal and ventricular beats. The autocorrelation signal of the QRS complex at scale four is shown for both leads (r_x and r_y) as well as the cross-correlation signal (r_{xy}) at the bottom. The zero crossings and peaks of interest are indicated with an asterisk.

where A^L_s is the mean absolute amplitude of the QRS peaks at scale s of the DWT, and lead L

$$A^L_s = \frac{1}{D}\sum_{d=1}^{D} \left| W^L_s s(l_d) \right|, \quad s = 1, 2, \ldots 6 \tag{3}$$

being D the number of detected peaks (1 or 2) and l_d is the positions of the peaks.

D. Discriminant Functions

Under the assumption of normally distributed data, the MAP classification criterion leads to quadratic discriminant functions, broadly used for classification purposes [18]. In the general case, the quadratic discriminant function of the ith class and feature vector $\mathbf{x}$, can be written as follows:

$$g_i(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\log(|\boldsymbol{\Sigma}_i|) + \log(P(\omega_i)) \tag{4}$$

being $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, and $P(\omega_i)$ the mean vector, covariance matrix, and prior probability of the ith class. The classification rule assigns $\mathbf{x}$ to the class i, which results in the maximum posterior probability $g_i(\mathbf{x})$. The values of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ were computed from the training data with the sample mean and covariance matrix expressions as follows:

$$\boldsymbol{\mu}_i = \frac{1}{M_i}\sum_{m=1}^{M_i} \mathbf{x}_m \tag{5}$$

$$\boldsymbol{\Sigma}_i = \frac{1}{M_i - 1}\sum_{m=1}^{M_i} (\mathbf{x}_m - \boldsymbol{\mu}_i).(\mathbf{x}_m - \boldsymbol{\mu}_i)^{\mathrm{T}} \tag{6}$$

being M_i the number of examples ($\mathbf{x}_m$) of the *i*th class. The values for the prior probabilities $P(\omega_i)$ were considered the same for all classes. In the case that the covariance matrix $\mathbf{\Sigma}$ is considered to be the same for all classes ($\mathbf{\Sigma}_i = \mathbf{\Sigma}_j = \mathbf{\Sigma}\ \forall i \neq j$), the quadratic discriminant classifier (QDC) becomes linear in $\mathbf{x}$ leading to the linear discriminant classifier (LDC)

$$g_i(\mathbf{x}) = \boldsymbol{\mu}_i^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_i + \log(P(\omega_i)) \tag{7}$$

where $\mathbf{\Sigma}$ can be estimated as the weighted sample covariance

$$\mathbf{\Sigma} = \frac{\sum_{i=1}^{C} w_i \sum_{m=1}^{M_i} (\mathbf{x}_m - \boldsymbol{\mu}_i)(\mathbf{x}_m - \boldsymbol{\mu}_i)^T}{\sum_{i=1}^{C} w_i M_i} \tag{8}$$

being C the total amount of classes and w_i the weighting coefficients. This class-weighting possibility is of much interest due to the heavy imbalance of the class sizes inherent to this application, where the normal class is, in general, one order of magnitude (at least) more represented than other classes. We refer as LDC to the linear classifier, where $w_i = w_j\ \forall i \neq j$, any other weight scheme will be referred as compensated linear classifier (LDC-C). In this paper, all classification tasks were performed using and modifying the PRtools toolbox [19] for MATLAB (The Mathworks Inc., MA).

E. Handling of Feature Domains

As the features to be included in our model belong to diverse domains, like $\mathbb{R}$, $\mathbb{R}^+$, and S^2 (angular or directional domain), we have to transform or deal with them in order to perform classification tasks. In our case, we assume that each feature is normally distributed, and therefore, valid in the $\mathbb{R}$ domain. According to this, all interval and morphological features defined in $\mathbb{R}^+$ should first being transformed to the $\mathbb{R}$ domain by a (natural) logarithm operation. In contrast, circular (or S^2) features require a special treatment that will be briefly described. The interested reader is referred to [20] for more details. For a directional feature, ϑ is the mean direction and directional variance, counterparts of the regular mean and variance are defined as follows [20]:

$$\mu_\vartheta^c = \arg(z) \tag{9}$$

$$V_\vartheta^c = 1 - |z| \tag{10}$$

where $z = E[e^{j\vartheta}]$. Then, for a multivariate F-dimensional model, where Θ is the set of indexes of the directional features, the mean vector $\boldsymbol{\mu}_i$, and covariance matrix $\mathbf{\Sigma}_i$ are as follows:

$$\boldsymbol{\mu}_i = [\mu_i(1) \dots \mu_i(F)]^T \tag{11}$$

with

$$\mu_i(f) = \begin{cases} \dfrac{1}{M_i} \displaystyle\sum_{m=1}^{M_i} x_m(f), & \text{if } f \notin \Theta \\ \arg\left(\dfrac{1}{M_i} \displaystyle\sum_{m=1}^{M_i} e^{j x_m(f)} \right), & \text{if } f \in \Theta \end{cases} \tag{12}$$

and

$$\mathbf{\Sigma}_i = \frac{1}{M_i - 1} \sum_{m=1}^{M_i} \mathbf{x}'_m \mathbf{x}'_m{}^T \tag{13}$$

being

$$\mathbf{x}'_m = [x'_m(1) \dots x'_m(F)]^T \tag{14}$$

$$x'_f(m) = \begin{cases} x_m(f) - \mu_i(f), & \text{if } f \notin \Theta \\ (x_m(f) - \mu_i(f))_{\mathrm{mod}\ 2\pi}, & \text{if } f \in \Theta. \end{cases} \tag{15}$$

As it can be noted from (13), $\mathbf{\Sigma}_i$ can be easily calculated from the directional mean $\boldsymbol{\mu}_i$ and the raw data.

F. Outlier Removal for Model Parameter Estimation

The classification performance proposed strongly depends on the parameter estimation of the multidimensional Gaussians in the training datasets. The parameter estimation (or training) process can be severely disrupted by the presence of outliers. This problem can be addressed by the removal of these atypical observations in the training data prior to the parameter estimation process. In this paper, the outliers removal is performed by the algorithm described in [21], which is a projection pursuit method based on the robust estimation of the translation, scale, and kurtosis of the distribution. For the *i*th class, the centroid is estimated as the median, defined for *F*-dimensional data as follows:

$$\begin{aligned} \underset{m}{\mathrm{med}}\, \mathbf{x}_m &= \mathrm{med}(\mathbf{x}_1, \dots, \mathbf{x}_{M_i}) \\ &= \arg\min_{\boldsymbol{\mu} \in \mathbb{R}^F} \sum_{m=1}^{M_i} \|\mathbf{x}_m - \boldsymbol{\mu}\| \end{aligned} \tag{16}$$

while the dispersion is estimated as the median absolute deviation (MAD) calculated as follows:

$$D_{\mathrm{MAD}}(\mathbf{x}_1, \dots, \mathbf{x}_{M_i}) = 1.4826 \cdot \underset{m}{\mathrm{med}} \left| \mathbf{x}_m - \underset{j}{\mathrm{med}}\, \mathbf{x}_j \right| \tag{17}$$

and finally, the kurtosis is estimated as follows:

$$\kappa(\mathbf{x}_1, \dots, \mathbf{x}_{M_i}) = \left| \frac{1}{M_i} \sum_{m=1}^{M_i} \frac{(\mathbf{x}_m - \mathrm{med}(\mathbf{x}_1, \dots, \mathbf{x}_{M_i}))^4}{(D_{\mathrm{MAD}} \mathbf{x}_1, \dots, \mathbf{x}_{M_i})^4} - 3 \right|. \tag{18}$$

The presence of outliers will make the tails of a distribution heavier, increasing the kurtosis coefficient; while a large number of outliers give raise to other modes in the distribution, decreasing the kurtosis coefficient. In a first phase, the algorithm search for outliers in the directions, where the kurtosis of the data is large or small to find location outliers. Then, in a second phase, the directions of large variance are explored to address scatter outliers [21]. For both phases, each example in the distribution gets one weight, which are finally combined in a final decision weight. Based on the final weight, the data is sorted and the 5% of the most outlying examples are discarded as outliers. With this assumption of slightly contaminated data, we set an operating point for the tradeoff between discarding useful data and allowing the presence of outliers in the parameter estimation process.

G. Performance Evaluation

The performance was measured in terms of the class sensitivity (S_i) and class positive predictive value (P_i^+); and the

global accuracy (A), global sensitivity (S), and global positive predictive value (P^+) as suggested in [12] for both, training and testing datasets. In a multiclass classification problem, the confusion matrix shows the outcome achieved by a classifier and a detailed distribution of the misclassified events. For a C class problem, the confusion matrix is a square matrix of dimension C

$$\begin{array}{c} \\ \text{True classes} \end{array} \begin{array}{c} \text{Estimated classes} \\ \begin{array}{ccccccc} 1 & \dots & i & \dots & C & \end{array} \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ C \end{array} \begin{pmatrix} n_{11}^T & \dots & n_{1i}^F & \dots & n_{1C}^F \\ \vdots & \ddots & \vdots & & \\ n_{i1}^F & \dots & n_{ii}^T & \dots & n_{iC}^F \\ \vdots & & \vdots & \ddots & \\ n_{C1}^F & \dots & n_{Ci}^F & \dots & n_{CC}^T \end{pmatrix} \begin{array}{c} N_1 \\ \vdots \\ N_i \\ \vdots \\ N_C \end{array} \\ \begin{array}{ccccc} P_1 & \dots & P_i & \dots & P_C \end{array} \quad N_T \end{array}$$

For the ith class, n_{ii}^T is the number of correctly classified examples and n_{ij}^F is the number of examples of class i classified as class j; N_i is the total number of examples for class i, P_i is the number of examples classified as class i, and N_T is the total number of examples in the dataset

$$N_i = n_{ii}^T + \sum_{m \neq i} n_{im}^F$$

$$P_i = n_{ii}^T + \sum_{m \neq i} n_{mi}^F$$

$$N_T = \sum_{i=1}^{C} N_i = \sum_{i=1}^{C} P_i.$$

Then, S_i and P_i^+ for the ith class are defined as follows:

$$S_i = \frac{n_{ii}^T}{N_i} \tag{19}$$

$$P_i^+ = \frac{n_{ii}^T}{P_i} \tag{20}$$

and the global accuracy (A), sensitivity (S), and positive predictive value (P^+) are calculated as follows:

$$A = \frac{1}{N_T} \sum_{i=1}^{C} n_{ii}^T = \sum_{i=1}^{C} \frac{N_i}{N_T} S_i \tag{21}$$

$$S = \frac{1}{C} \sum_{i=1}^{C} S_i \tag{22}$$

$$P^+ = \frac{1}{C} \sum_{i=1}^{C} P_i^+. \tag{23}$$

From this equations, it is clear that any imbalance in the class representation directly impacts over the P^+, P_i^+, and A calculation, but not over the S and S_i.

Although the AAMI recommendation does not suggest any measure to deal with the strong class size imbalance (see Table I), we considered weighting the classes previous to the calculation of P_i^+ and A in order not to neglect the performance of the less represented classes. The balancing approach used in this paper consists in multiplying each row of the confusion matrix by a constant, such that the sum of each row N_i is equal for all classes, or $N_i = N_j \ \forall i \neq j$. This is equivalent to repeat examples of the less represented classes, in order to balance the class presence. We will refer to this as the balanced performance estimation method in Section III. We also use another way of showing the global performance referred as "by recording," which consists in averaging the performance estimates in a record-by-record (or subject) way.

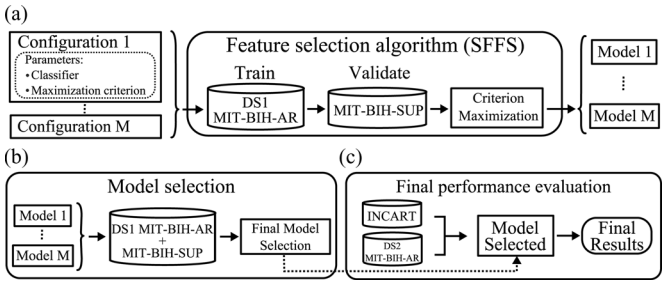

Fig. 4. Block diagram describing the experiments performed in this paper. In panel a, the feature selection algorithm is summarized, indicating the train and validation dataset division, as well as the different parameters of the algorithm. In panel b is shown the methodology to obtain the best performing model among the different searches performed. Finally, in panel c, the best performing model is selected for the final performance evaluation in the test datasets.

TABLE II
SUMMARY OF THE BEST PERFORMING MODELS FOUND WITH THE SFFS ALGORITHM SEPARATING ALL AAMI2 CLASSES

			Model Evaluation								
	Opt.		Normal		Suprav.		Ventr.		Total		
Classifier	Crit.	# Features	S	P^+	S	P^+	S	P^+	A	S	P^+
LDC-C	J_{P^+}	**8**	**93**	**98**	**78**	**40**	**68**	**70**	**91**	**80**	**70**
LDC	J_{P^+}	10	92	98	57	38	77	50	89	75	62
QDC	J_{P^+}	7	80	98	7	12	89	22	77	59	44
LDC-C	J_S	10	90	98	77	33	68	74	88	79	68
LDC	J_S	10	92	98	74	37	70	67	89	78	67
QDC	J_S	9	87	98	43	32	80	33	84	70	55
de Chazal et al. [4]		48	87	98	57	30	63	36	84	69	55

In the left, the parameters used for the SFFS to find the model, and in the right, the performance DS1 of the MIT-BIH-AR and the whole MIT-BIH-SUP database (with cross validation of k=10 recordings). The best performing model (in bold) is selected for the final performance evaluation. The results are expressed in percentages.

H. Model Selection and Dimensionality Reduction

It is well known that low-dimensional models generalize better to examples not presented during the training phase, resulting in a more robust and realistic classifier [18]. In order to obtain a small and well-performing model, a sequential floating feature selection algorithm (SFFS) was used [22]. The SFFS algorithm can be briefly explained as the combination of two simpler steps, a sequential forward selection (SFS) algorithm followed by a sequential backward selection (SBS) algorithm. The SFFS iterates for all model sizes, starting from a single feature model, and registering the best performances found for each model size. Each iteration starts with an SFS step, and from a model size greater than two features after each SFS step, an SBS step is repeated until the performance of the model found is not greater than the registered for this smaller model size. This way the algorithm goes forward and backward (like floating) searching at each step for the path of maximum performance.

TABLE III
PERFORMANCE COMPARISON BETWEEN THE MODEL SELECTED IN TABLE II AND THE REFERENCE CLASSIFIER [4] SEPARATING ALL AAMI2 CLASSES IN DS2 OF MIT-BIH-AR

de Chazal et al. [4]

Truth \ Algorithm	n	s	v'	Total
N	40718	1863	1677	44258
S	307	1361	169	1837
V'	235	845	2529	3609
Total	41260	4069	4375	49704

Model selected in Table II

Truth \ Algorithm	n	s	v'	Total
N	41950	2002	236	44188
S	216	1422	197	1835
V'	473	222	2911	3606
Total	42639	3646	3344	49629

Performance calculation mode	Classifier	Automatic	# Features	Normal S	Normal P^+	Suprav. S	Suprav. P^+	Ventr. S	Ventr. P^+	Total A	Total S	Total P^+
Imbalanced	This work	yes	8	95	98	77	39	81	87	93	84	75
	de Chazal et al. [4]	yes	48	92	99	74	33	70	58	90	79	63
	de Chazal et al. [7]	no	48	94	99	88	47	95	82	94	92	76
Balanced	This work	yes	8	95	79	77	88	81	88	84	84	85
	de Chazal et al. [4]	yes	48	92	80	74	73	70	84	79	79	79

Both models were trained in DS1 of the same database. First, the confusion matrices for both models are shown, and then, the class and total performances are summarized. The performances are expressed in percentages for both, balanced and imbalanced class presence in the dataset

The algorithm ends when the specified greater model size is reached. The result of the algorithm is the model found with maximum performance. The interested reader is referred to [22] for a detailed description and to [19] for an implementation of the SFFS algorithm. The performance metrics used by the feature selection algorithm were a weighted class S and P^+ calculated as follows:

$$J_S = \frac{\sum_{i=1}^{C} \pi_i S_i}{\sum_{i=1}^{C} \pi_i} \tag{24}$$

$$J_{P^+} = \frac{\sum_{i=1}^{C} \pi_i P_i^+}{\sum_{i=1}^{C} \pi_i} \tag{25}$$

with C classes and being S_i and P_i^+ the class sensitivities and positive predictive defined in the previous section. The class weights π_i allow the possibility of directing the search to specific class performances.

I. Experiment Setup

In this paper, we are interested in finding a reduced dimension, well performing and generalizing model in a multidatabase context. The experiment can be divided in three steps.

1) In the first step, we search for the best performing model, from the 39 available features, in the training (DS1 of MIT-BIH-AR) and validation (MIT-BIH-SUP) sets [see Fig. 4(a)]. In each iteration of the SFFS algorithm, the current model was trained in DS1 of MIT-BIH-AR and its performance was evaluated in the MIT-BIH-SUP database. As the data divisions in both databases do not share any recording, the features selected should retain the generalization properties. Several parameter configurations were studied for the SFFS algorithm, like the effect of the classifier (LDC, LDC-C, and QDC) and the optimization criterion (J_S or J_{P^+}) for the search. The weight compensation used in the experiments for the LDC classifier is $w_N = 1$, $w_S = 10$, and $w_V = 10$. The same weights were also studied for the J_{P^+} and J_S criterion $\pi_N = 1$, $\pi_S = 10$, and $\pi_V = 10$. At the end of this step, we have an optimal feature set for each parameter configuration.

TABLE IV
FEATURES USED IN THE MODEL SELECTED IN TABLE II FOR THE FINAL PERFORMANCE EVALUATION

Feature	Description
$\ln(RR[i])$	Current RR interval
$\ln(RR[i+1])$	Next RR interval
$\ln(RR_1)$	Average RR interval in the last minute
$\ln(RR_{20})$	Average RR interval in the last 20 minutes
k_Z^x	Zero-cross position of the WT autocorrelation signal in lead 1
k_Z^y	Zero-cross position of the WT autocorrelation signal in lead 2
k_M^x	Maximum position of the WT autocorrelation signal in lead 1
k_M^y	Maximum position of the WT autocorrelation signal in lead 2

2) The second step [see Fig. 4(b)] is the selection of the best performing model, among the best models obtained in the previous step for each parameter configurations. For this purpose, we compare the global results (A, S, and P^+) obtained in the union set of DS1 of MIT-BIH-AR dataset and the MIT-BIH-SUP database, using a recording-based k-fold cross validation with $k = 10$ recordings.

3) Finally, the performance of the selected model is evaluated in DS2 for comparison with [4], as shown in Fig. 4(c). Additionally, the performance in the INCART database is compared to that obtained in DS2 to asses how the model behaves in completely different databases.

The results presented in this paper are compared to the classifier developed in [4] (reference classifier in the rest of this paper), being this, to our knowledge, the best performing fully automatic multiclass classifier (AAMI compliant) reviewed in the literature. In order to perform a fair comparison, some methodological aspects were maintained as similar as possible. The implementation of the classifier suggested in [4] was contrasted with the reported results obtaining comparable results. With this implementation, we could evaluate the generalization capability of the reference classifier in the MIT-BIH-SUP database, since this experiment was not performed in [4]. In these situations, where the experiments were already performed in [4], the reported results were used.

All experiments described in this paper will focus to achieve automatic classification between the three AAMI2 classes (N, S, and V'), since the fusion class is poorly represented in

TABLE V
DETAILED RESULTS GROUPING BY RECORDING (OR SUBJECT), FOR THE MODEL SELECTED IN TABLE II SEPARATING ALL AAMI2 CLASSES IN DS2 OF MIT-BIH-AR, FOLLOWING AAMI RECOMMENDED PERFORMANCE MEASURES

	Number of beats			Normal		Supraventr.		Ventricular		Totals		
Rec	N	S	V'	S	P^+	S	P^+	S	P^+	A	S	P^+
100	2235	33	1	100%	77%	70%	100%	100%	100%	100%	90%	92%
103	2079	2	0	99%	50%	0%	0%	–	–	99%	50%	25%
105	2524	0	41	97%	100%	–	–	51%	94%	96%	74%	97%
111	2120	0	1	99%	100%	–	–	100%	99%	99%	100%	100%
113	1785	6	0	99%	100%	100%	99%	–	–	99%	100%	100%
117	1531	1	0	100%	100%	100%	100%	–	–	100%	100%	100%
121	1858	1	1	99%	100%	100%	99%	100%	100%	99%	100%	100%
123	1512	0	3	100%	100%	–	–	0%	0%	100%	50%	50%
200	1733	30	826	96%	58%	27%	81%	92%	91%	94%	72%	77%
202	2059	55	20	67%	87%	93%	56%	50%	87%	68%	70%	77%
210	2419	22	205	94%	86%	91%	81%	69%	87%	92%	85%	85%
212	2745	0	0	100%	100%	–	–	–	–	100%	100%	100%
213	2637	28	582	100%	63%	46%	100%	44%	47%	89%	63%	70%
214	2000	0	257	97%	100%	–	–	94%	98%	97%	96%	99%
219	2080	7	65	86%	46%	0%	0%	82%	100%	86%	56%	49%
221	2029	0	396	93%	99%	–	–	99%	100%	94%	96%	100%
222	2271	208	0	72%	92%	89%	76%	–	–	73%	81%	84%
228	1685	3	362	100%	60%	33%	84%	93%	100%	99%	75%	81%
231	1565	1	2	98%	49%	0%	0%	50%	100%	97%	49%	50%
232	397	1381	0	100%	90%	78%	100%	–	–	83%	89%	95%
233	2227	7	841	100%	92%	71%	90%	83%	74%	95%	85%	85%
234	2697	50	3	100%	78%	72%	100%	100%	100%	99%	91%	93%
Total	44188	1835	3606	95%	83%	61%	73%	75%	86%	94%	80%	82%

TABLE VI
PERFORMANCE COMPARISON BETWEEN THE MODEL SELECTED IN TABLE II AND THE REFERENCE CLASSIFIER [4] SEPARATING ALL AAMI CLASSES IN DS2 OF MIT-BIH-AR

de Chazal et al. [4]

Truth \ Algorithm	f	n	s	v	Total
F	347	33	1	7	388
N	3509	38444	1904	303	44160
S	16	173	1395	252	1836
V	176	117	321	2504	3118
Total	4048	38767	3621	3066	49502

Model selected in Table II

Truth \ Algorithm	f	n	s	v	Total
F	370	11	2	5	388
N	8031	34270	1807	80	44188
S	28	124	1403	280	1835
V	321	46	182	2669	3218
Total	8750	34451	3394	3034	49629

Performance calculation mode	Classifier	Fusion S	Fusion P^+	Normal S	Normal P^+	Suprav. S	Suprav. P^+	Ventr. S	Ventr. P^+	Total A	Total S	Total P^+
Imbalanced	This work	95	4	78	99	76	41	83	88	78	83	58
	de Chazal et al. [4]	89	9	87	99	75	39	80	81	86	83	57
Balanced	This work	95	76	78	88	76	88	83	83	83	83	84
	de Chazal et al. [4]	89	86	87	80	76	84	80	83	83	83	83

Both models were trained in DS1 of the same database. First, the confusion matrices for both models are shown, and then, the classes and total performances are summarized. The performances are expressed in percentages for both, balanced and imbalanced class presence in the dataset.

the databases used. The restrictions imposed by the recording-oriented division of the data, and the fact that only a few recordings concentrate the majority of the examples of the fusion heartbeats, makes unfeasible to perform the feature selection using the original AAMI labeling. Despite this limitation for the model selection, the model obtained for the three AAMI2 classes was also retrained and evaluated, classifying the four AAMI classes to show its utility.

III. RESULTS

The main results for the experiments described in the previous section are summarized in Tables II and III. Table II shows the results of the best models obtained for the different parameter configurations during the model selection. The best performing of these models was an eight feature model trained in the DS1 of the MIT-BIH-AR. The eight features that the model comprehends are listed in Table IV. The classifier used was an LDC-C, using equal prior probabilities $P(\omega_i)$. The optimization criterion used in the SFFS was J_{P^+} with equal weights π_i.

The performance of the selected model in the test set (DS2) is compared with the reported by de Chazal *et al.* [4] in Table III. The model found in this paper achieves better performance for the three classes. Table V presents the performance by recording in the test set, following the recommendations of the AAMI [12] for result presentation.

The performance of the selected model with the four AAMI classes (N, S, V, and F) is reported in Table VI. The model found achieves a performance slightly lower than the reference, but it must be noted that the selected model was optimized for the three AAMI2 classes (N, S, and V').

TABLE VII
CONFUSION MATRIX AS A RESULT OF SEPARATING ALL AAMI2 CLASSES IN THE INCART DATABASE

		Algorithm			
		n	s	v'	Total
Truth	N	140983	10576	1958	153517
	S	84	1660	214	1958
	V'	644	3007	16559	20210
	Total	141711	15243	18731	175685

Performance calculation mode	Dataset	Normal		Suprav.		Ventr.		Total		
		S	P^+	S	P^+	S	P^+	A	S	P^+
Imbalanced	$DS2$ MIT-BIH-AR	95	98	77	39	81	87	93	84	75
	INCART	92	99	85	11	82	88	91	86	66
Balanced	$DS2$ MIT-BIH-AR	95	79	77	88	81	88	84	84	85
	INCART	92	92	85	80	82	87	86	86	86
By recording	$DS2$ MIT-BIH-AR	95	83	61	73	75	86	94	80	82
	INCART	93	90	64	66	71	86	91	79	85

Model used is the selected from Table II, trained in DS1 of the MIT-BIH-AR database. The performances are expressed in percentages, and grouped by performance calculation mode for easy comparison with the results obtained from DS2 of MIT-BIH-AR (see Table III).

Finally, the performance of the model found in the INCART database is presented in Table VII. The performance obtained in this database is comparable for all classes with that obtained in DS2.

IV. DISCUSSION AND CONCLUSION

In this paper, we have presented a methodology to develop a simple and robust heartbeat classification system, and we evaluated it focusing in the generalization capability. In order to do this, we take into consideration the MIT-BIH-SUP [14] and the INCART databases in addition to the widely used MIT-BIH-AR, all freely available in Physionet [15]. Although these databases are bigger than the original MIT-BIH-AR, the fusion class defined in the AAMI standard [12] is not so well represented as the other classes. This limitation is overcome by adopting the alternative labeling AAMI2 proposed in this paper. The AAMI2 labeling make sense from a physiological point of view, since the AAMI fusion class comprehends these heartbeats, which results from the simultaneous occurrence of normal and ventricular heartbeats.

From the results obtained for the model selection presented in Table II, several models that outperform the reference classifier [4] were achieved. The best model found consists of eight features: $\ln(\text{RR}[i])$, $\ln(\text{RR}[i+1])$, $\ln(\text{RR}_1)$, $\ln(\text{RR}_{20})$, k_Z^x, k_Z^y, k_M^x, and k_M^y, which are described in Table IV. As can be noted, the selected features are computed without exception from time-interval measurements. This could be explained, given that the used databases do not always include the same pair of ECG leads in each recording. Therefore, the classification performances of features, which are calculated from amplitudes are heavily degraded. The directional features (like the VCG_ϕ) were also probably affected by this fact, even if the clinical importance of this kind of features is well known by cardiologists [1]. In contrast, intervals seem to retain the classification ability with independence of the pair of leads chosen. The first four features in the model are clearly connected to the evolution of heart rhythm, while the other four can be understood as surrogate measurements of the QRS width, and therefore, the QRS morphology. As a result, the model found has the evident advantage of a lower size, which results in a computational saving and lower error in the parameter estimation during the training phase. In addition, it only relies on the QRS fiducial point detection, making the classifier model robust to degraded signals, where the delineation of the ECG waves is not reliable.

It is worth noting than the performance achieved by the reference classifier [4] in the union of train and validation dataset (see Table II) is lower for all classes than the obtained in the final performance reported in Table III. The same phenomenon happens with the suggested model in a smaller degree, with the exception of the supraventricular performance. This phenomenon was also reported in [4], obtaining better performance in the test set than in the training set. These results suggest that DS2 dataset may not be a good data sample to measure the actual performance of a classifier. To avoid this bias in the actual performance, it may be convenient in future works that the final performance estimation would be performed applying other methodologies or redefining the test dataset. One reason that could be biasing the results in DS2 is the different amount of examples by recording for the supraventricular class. As can be seen in Table V, recordings 232 and 222 concentrate the majority of the examples for the supraventricular class, which means that failing in these recordings impacts considerably to the S class performance. For this reason, the average performances presented in Table V could also be of importance, since each recording or subject is equally weighted in the average.

The results presented in [7], where the automatic classifier of [4] is assisted by a LE to improve its performance, are also compared in Table III. This suggests that a similar approach of combining the knowledge of a LE with our model, could also lead to a comparable improvement in the baseline performance.

An additional assessment of the suggested model classifying the four AAMI (N, S, V, and F) classes is presented in Table VI. The results verify the validity of the model achieving slightly lower performance than the results presented in [4]. It must be noted that the model presented in this paper was optimized for the AAMI2 labeling (N, S, and V'), and the classifier is mainly misclassifying normal heartbeats as fusion, as shown in Table VI.

The results in Table VII suggest that the selected features have good generalization capability when evaluating the performance in heartbeats not considered during the development phase, like the ones from the INCART database. The imbalanced performance is comparable for all classes except the supraventricular, where a decrease in the P^+ occurred. This could be explained by an increased class imbalance in the INCART database, which is about 75-to-1, while in MIT-BIH-AR is 22-to-1 approximately. This is confirmed by the balanced results (equivalent to a class imbalance of 1-to-1) in the same table, where the performance figures are very similar. The validity of the generalization capability of the proposed model, is somehow restricted to the available data, and should be corroborated in future works by including new databases in the analysis or other methodologies. Despite this limitation, the degree of generalization of the suggested model is expected to be better than models obtained, considering only the MIT-BIH-AR database.

One limitation of the presented approach is the Gaussian assumption of the data imposed by the classifier, since many features were observed not to fulfill this requirement. Despite this evident limitation, the linear decision regions in the feature space defined by the LDC-C allowed us to select those features, which inherently provide better classification performance. Considering the proposed classifier and feature model as a reference for future improvements, the effect of the lack of Gaussianity can be mitigated using more complex classifiers, like ANN's or mixture of Gaussians. These classifiers allow more complex decision regions in the feature space, retaining details of the training data, which may improve the classification performance.

Despite the improved results presented in this paper, there is still room for improvement in the field, since the S and P^+ for the supraventricular class are of 77% and 39%, and for the ventricular class (though better) are of 81% and 87%. This results suggest that other features, classifiers or meta-classifier strategies (like LE assistance) can be developed in order to improve the performance, specially in the supraventricular class.

ACKNOWLEDGMENT

The authors would like to thank to Dr. P. de Chazal for the help in the implementation of the methodology described in [4].

REFERENCES

[1] G. J. Taylor, *150 Practice ECGs: Interpretation and Review*. Oxford, U.K.: Blackwell, 2002.

[2] Y. H. Hu, S. Palreddy, and W. Tompkins, "A patient-adaptable ecg beat classifier using mixture of experts approach," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 9, pp. 891–899, Sep. 1997.

[3] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sörnmo, "Clustering ECG complexes using hermite functions and self-organizing maps," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 838–848, Jul. 2000.

[4] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004.

[5] O. Inan, L. Giovangrandi, and G. Kovacs, "Robust neural-netwok-based classification of premature ventricular contractions using wavelet transform and timing interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2507–2515, Dec. 2006.

[6] I. Christov *et al.*, "Comparative study of morphological and time-frequency ecg descriptors for heartbeat classification," *Elsevier Med. Eng. Phys.*, vol. 28, pp. 876–887, 2006..

[7] P. de Chazal and R. B. Reilly, "A patient-adapting heartbeat classifier using ecg morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2535–2543, Dec. 2006.

[8] M. Llamedo and J. Martínez, "An ecg classification model based on multilead wavelet transform features," in *Computers in Cardiology 2007*, vol. 34, Washington, DC: IEEE Computer Society Press, 2007, pp. 105–108.

[9] W. Jiang and S. Kong, "Block-based neural networks for personalized ECG signal classification," *IEEE Trans. Biomed. Eng.*, vol. 18, no. 6, pp. 1750–1761, Nov. 2007.

[10] K. Park, B. Cho, D. Lee, S. Song, and J. Lee, "Hierarchical support vector machine," in *Computers in Cardiology 2008*, vol. 35, Washington, DC: IEEE Computer Society Press, 2008, pp. 229–232.

[11] T. Ince, S. Kiranyaz, and M. Gabbouj, "A generic and robust system for automated patient-specific classification of ecg signals," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 5, pp. 1415–1426, May 2009.

[12] *Testing and Reporting Performance Results of Cardiac Rhythm and ST-Segment Measurement Algorithms*, American National Standard, ANSI/AAMI/ISO EC57, 1998–(R)2008..

[13] R. Mark and G. Moody. (1997). "Mit-bih arrhythmia database," [Online]. Available: http://ecg.mit.edu/dbinfo.html.

[14] R. Mark, G. Moody, and S. Greenwald. (1990). "Mit-bih supraventricular arrhythmia database," [Online]. Available: http://www.physionet.org/physiobank/database/svdb/.

[15] A. L. Goldberger *et al*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[16] S. Greenwald, "Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information," Ph.D. dissertation, Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, 1990..

[17] J. P. Martínez, R. Almeida, S. Olmos, A. Rocha, and P. Laguna, "A wavelet-based ecg delineator: Evaluation on standard databases," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 570–581, Apr. 2004.

[18] F. van der Heijden, R. Duin, D. de Ridder, and D. Tax, *Classification, Parameter Estimation and State Estimation*. New York: Wiley, 2005.

[19] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. deRidder, D. Tax, and S. Verzakov. (2008). "Pr-tools, a matlab toolbox for pattern recognition," [Online]. Available: http://www.prtools.org.

[20] C. Bahlmann, "Directional features in online handwriting recognition," *Pattern Recognit.*, vol. 39, no. 1, pp. 115–125, 2006.

[21] P. Filzmoser, R. Maronna, and M. Werner, "Outlier identification in high dimensions," *Comput. Statist. Data Anal.*, vol. 52, pp. 1694–1711, 2008.

[22] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

Mariano Llamedo was born in Buenos Aires, Argentina, in 1979. He received the M.Sc. degree in electronic engineering from the Facultad Regional Buenos Aires, National Technological University (UTN-FRBA), Buenos Aires, in 2005.

From 2005 to 2008, he was with the Department of Electronic Engineering, UTN-FRBA, as an Assistant Professor and Research Fellow. Since 2008, he has been a Researcher with the Aragon Institute of Engineering Research, University of Zaragoza, Zaragoza, Spain. His current research interests include the field of biomedical signal processing, with main interest in signals of cardiovascular origin.

Juan Pablo Martínez was born in Zaragoza, Aragon, Spain, in 1976. He received the M.S. degree in telecommunication engineering and the Ph.D. degree in biomedical engineering from the University of Zaragoza (UZ), Zaragoza, in 1999 and 2005, respectively.

Since 2000, he has been an Assistant Professor at the Aragon Institute of Engineering Research, UZ, where he has been an Associate Professor since 2007. He is also with the Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, Zaragoza. His current research interests include biomedical signal processing, with main interest in signals of cardiovascular origin.