PAPER • **OPEN ACCESS**

From 12 to 1 ECG lead: multiple cardiac condition detection mixing a hybrid machine learning approach with a one-versus-rest classification strategy

To cite this article: Santiago Jiménez-Serrano *et al* 2022 *Physiol. Meas.* **43** 064003

IOP Publishing Physiol. Meas. 43 (2022) 064003 https://doi.org/10.1088/1361-6579/ac72f5

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
8 January 2022

REVISED
16 May 2022

ACCEPTED FOR PUBLICATION
24 May 2022

PUBLISHED
28 June 2022

PAPER

From 12 to 1 ECG lead: multiple cardiac condition detection mixing a hybrid machine learning approach with a one-versus-rest classification strategy

Santiago Jiménez-Serrano[1,*], **Miguel Rodrigo**[2], **Conrado J Calvo**[3], **José Millet**[1] **and Francisco Castells**[1]

[1] Instituto ITACA, Universitat Politècnica de València, València, Spain
[2] CoMMLab, Universitat de València, València, Spain
[3] Universitat Politècnica de València, València, Spain
[*] Author to whom any correspondence should be addressed.

E-mail: sanjiser@upv.es

Keywords: ECG, signal processing, feature extraction, feature selection, machine learning, classification, cardiac conditions detection

Abstract

Objective. Detecting different cardiac diseases using a single or reduced number of leads is still challenging. This work aims to provide and validate an automated method able to classify ECG recordings. Performance using complete 12-lead systems, reduced lead sets, and single-lead ECGs is evaluated and compared. *Approach*. Seven different databases with 12-lead ECGs were provided during the *PhysioNet/Computing in Cardiology Challenge* 2021, where 88 253 annotated samples associated with none, one, or several cardiac conditions among 26 different classes were released for training, whereas 42 896 hidden samples were used for testing. After signal preprocessing, 81 features per ECG-lead were extracted, mainly based on heart rate variability, QRST patterns and spectral domain. Next, a One-versus-Rest classification approach made of independent binary classifiers for each cardiac condition was trained. This strategy allowed each ECG to be classified as belonging to none, one or several classes. For each class, a classification model among two binary supervised classifiers and one hybrid unsupervised-supervised classification system was selected. Finally, we performed a 3-fold cross-validation to assess the system's performance. *Main results*. Our classifiers received scores of 0.39, 0.38, 0.39, 0.38, and 0.37 for the 12, 6, 4, 3 and 2-lead versions of the hidden test set with the Challenge evaluation metric (*CM*). Also, we obtained a mean *G*-score of 0.80, 0.78, 0.79, 0.79, 0.77 and 0.74 for the 12, 6, 4, 3, 2 and 1-lead subsets with the public training set during our 3-fold cross-validation. *Significance*. We proposed and tested a machine learning approach focused on flexibility for identifying multiple cardiac conditions using one or more ECG leads. Our minimal-lead approach may be beneficial for novel portable or wearable ECG devices used as screening tools, as it can also detect multiple and concurrent cardiac conditions.

1. Introduction

The clinical importance of cardiac diseases, commonly linked with population ageing, is rising along with their incidence and prevalence (Chow *et al* 2012), and as a consequence is becoming the foremost cause of death worldwide (Gaziano *et al* 2010, Virani *et al* 2021). In this sense, the 12-lead ECG is the primary technique in cardiac diagnosis (Kligfield 2002), being a cheap and widely available screening tool. However, it requires experienced clinicians to interpret ECG recordings, being this process a time-consuming task subject to inter-observer variability (Bickerton and Pooler 2019).

In order to help in this task, some previous works have addressed the problem of automatic ECG abnormalities detection in different ways. Early approaches were mainly based on time-frequency analysis and features (Alexakis *et al* 2003, Chazal *et al* 2004, Christov *et al* 2006, Mahmoud *et al* 2006, Kostka and Tkacz 2007), or wavelet and Fourier signal transforms (Martínez *et al* 2004, Mahmoodabadi *et al* 2005, Minami *et al* 1999,

Yang and Shen 2013, Aqil *et al* 2015). Next, fuzzy logic and machine learning techniques were tested in order to detect heart diseases (Vafaie *et al* 2014, Chen *et al* 2018). A comparison of machine learning methods for the classification of five different cardiovascular abnormalities can be found in Hagan *et al* (2021), where support vector machines, artificial neural networks, and ensembles of decision trees were tested. In Zheng *et al* (2020a), a Gradient Boosting Tree classifier achieved the best classification performance among four different cardiac rhythms using 12-leads ECG registers. In recent years, deep learning, mostly based on convolutional neural networks (CNN), is also gaining importance in classification tasks of ECG segments and has been used to detect Atrial Fibrillation in a single lead (Xiong *et al* 2018, Rahimeh *et al* 2021, Krasteva *et al* 2021), or to improve the annotating performance of P-QRS-T waves through a recurrent CNN model (Sodmann *et al* 2018). Other works aim to detect as many as nine cardiac conditions in large ECG datasets, mostly using CNNs (Jo *et al* 2021, Yang *et al* 2021, Hua *et al* 2021, Dai *et al* 2021, Zhang *et al* 2021). Besides this, recent works have reported results for the detection of 24–27 cardiac abnormalities in databases containing a maximum of 66 361 samples, using 12 leads and different deep learning frameworks (Zhaowei *et al* 2021, Giovanni *et al* 2021, Zhao *et al* 2022).

On the other hand, wearable devices are gaining great interest for the early detection of cardiac diseases in both research and clinical settings due to their potential for massive screening and monitoring of cardiac conditions in a wide variety of scenarios (Kumari *et al* 2017). Furthermore, these devices usually capture the ECG with a reduced number of leads. In fact, the ECG signal obtained from devices such as smartwatches that pick up the ECG from the differential potential of right and left hands/wrists, is equivalent to lead I.

Nevertheless, accurate diagnosis of cardiac diseases using a single or reduced number of leads in an automatic way is still challenging (Dunn *et al* 2018, Georgiou *et al* 2018). Despite the aforementioned advantages, a main drawback of systems with a reduced number of leads is the loss of morphologic features and patterns only visible in specific leads. For example, the characteristic atrial *f*-wave described during atrial fibrillation is mainly visible in lead V1, whereas it is barely appreciable in the lead I (Cheng *et al* 2013). Therefore, exploiting the information of the atrial signal is not likely to provide good discrimination from reduced lead systems where V1 is missing. Another example worth mentioning is the typical ST-segment elevation caused by transmural ischemia, although this is only visible in leads picked up by electrodes placed in the direction toward the ST vector of the affected region points (Deshpande and Birnbaum 2014). Thus, this feature is, in practice, exclusive of precordial leads. Another important diagnostic feature that could be missed in reduced lead systems that exclude precordial leads (i.e. using only limb leads) is the T-wave inversion caused by a variety of cardiac syndromes (Said *et al* 2015).

Despite these limitations, smart devices with reduced lead sets can still be useful as massive screening tools. The main challenge in the diagnostic with these devices is to find an appropriate balance between sensitivity (in order not to miss important cardiac problems) and specificity (to avoid unnecessarily collapsing health centers). Hence, accurate algorithms suitable for mobile or wearable devices able to detect cardiac conditions in ECG registers are highly desirable.

In this work, we aim to develop and validate a robust methodology able to identify a wide variety of cardiac conditions from five different ECG leads combinations (12, 6, 4, 3, and 2-leads) plus the single lead 'I' scenario. This aim wants to address the clinical need for novel wearables technologies with limited electrocardiographic information in which multiple cardiac conditions can be identified automatically. We also aim to evaluate the classification performance of our solution with short ECG segments, as it is likely to be obtained with smart devices.

2. Materials

To validate the performance of the proposed approach, our experiments were conducted and examined in the context of the *PhysioNet/Computing in Cardiology Challenge* 2021 (Perez Alday *et al* 2020, Reyna *et al* 2021), where seven databases with a total of 131 149 ECG traces were used. We used the 88 253 12-lead ECG recordings provided as public training s*et al*so containing the age and gender of the patient for each record. On the other hand, the challenge organizers assessed online the performance of the classification models with hidden validation and test databases, composed of 6 630 and 36 266 ECG recordings, respectively. Raw data have seven different sources, as shown in table 1. The first source is the China Physiological Signal Challenge 2018 database (CPSC) (Liu *et al* 2018). The second source is the St. Petersburg Institute of Cardiological Technics database (INCART) (Tihonenko *et al* 2008). The third source is the Physikalisch-Technische Bundesanstalt database (PTB) (Bousseljot *et al* 1995, Wagner *et al* 2020). The fourth source is the Georgia 12-lead ECG Challenge database (G12EC). The fifth database comes from an undisclosed American institution. The sixth source is the Chapman University, Shaoxing People's Hospital (Chapman-Shaoxing) (Zheng *et al* 2020a) and Ningbo First Hospital (Ningbo) (Zheng *et al* 2020b) databases (CHAP-SHX).

And finally, the seventh source is the UMich Database from the University of Michigan.

Table 1. The number of ECG recordings in the public training database, as well as in the official hidden validation and test databases used in the *PhysioNet/Computer in Cardiology Challenge* (2021).

Database	Recordings in official public training set	Recordings in official hidden validation set	Recordings in official hidden test set	Total recordings
CPSC	10 330	1463	1463	13 256
INCART	74	—	—	74
PTB	22 353	—	—	22 353
G12EC	10 344	5167	5161	20 672
Undisclosed	—	—	10 000	10 000
CHAP-SHX	45 152	—	—	45 152
UMich	—	—	19 642	19 642
Total	88 253	6630	36 266	131 149

Table 2. Cardiac conditions to be detected (plus Normal Sinus Rhythm), abbreviation and number of samples in the public training dataset.

Cardiac conditions to be detected (classes)	Abbreviation	Number of samples for training
Atrial fibrillation	AF	5255
Atrial flutter	AFL	8374
Bundle branch block	BBB	522
Bradycardia	Brady	295
Complete left bundle branch block|left bundle branch Block	CLBBB|LBBB	213|1,281
Complete right bundle branch block|right bundle branch Block	CRBBB|RBBB	1779|3051
1st degree AV block	IAVB	3534
Incomplete right bundle branch block	IRBBB	1857
Left axis deviation	LAD	7631
Left anterior fascicular block	LAnFB	2186
Prolonged PR interval	LPR	392
Low QRS voltages	LQRSV	1599
Prolonged QT interval	LQT	1907
Nonspecific intraventricular conduction disorder	NSIVCB	1768
Normal sinus rhythm	NSR	28 971
Premature atrial contraction|supraventricular premature beats	PAC|SVPB	3041|224
Pacing rhythm	PR	1481
Poor R-wave progression	PRWP	638
Premature ventricular contractions|ventricular premature beats	PVC|VPB	1279|659
Q-wave abnormal	QAb	2076
Right axis deviation	RAD	1280
Sinus arrhythmia	SA	3790
Sinus bradycardia	SB	18 918
Sinus tachycardia	STach	9657
T-wave abnormal	TAb	11 716
T-wave inversion	TInv	3989

Public training recordings used in this work came from five of the mentioned databases: CPSC, INCART, PTB, G12EC and CHAP-SHX. Deeper explanations of the databases (sampling frequency, samples lasting, etc) can be found in Perez Alday *et al* (2020), Reyna *et al* (2021).

Table 2 shows the cardiac conditions to be detected and the number of samples labelled with such labels in the public training dataset. In this dataset, one ECG-record could be labelled with more than one class, and some classes contain two cardiac diseases since they scored as the same diagnosis.

We relabeled the next four classes in the dataset: LBBB as CLBBB, RBBB as CRBBB, SVPB as PAC and finally VPB as PVC. This way, we reduced the number of classes to be detected from 30 to 26. The main reason to do this is that each relabelled class shares the main rhythm and features with its final label and also scored as the same diagnosis in the context of the *PhysioNet/Computer in Cardiology Challenge 2021* (Reyna *et al* 2021).

On the other hand, 6 287 ECG records that did not belong to any of the 26 classes to be detected were removed from the training dataset since no binary classifier in this work might take them as positive samples.

In addition to the official leads sets of the *PhysioNet/Computing in Cardiology Challenge 2021* (12, 6, 4, 3 and 2-leads), we also report the results using only lead I. We consider this important in order to assess the suitability of our automatic classification methods for ECG signals in devices that record only lead I. Table 3 shows the number and sets of ECG leads evaluated in this work, used to train and validate the classification models.

Table 3. Number and sets of ECG leads used to train and validate our classification models.

#Leads	Leads Sets
12	I, II, III, aVR, aVL, aVF, V1-V6
6	I, II, III, aVR, aVL, aVF
4	I, II, III, V2
3	I, II, V2
2	I, II
1	I

3. Methods

In this section, we describe the methodology proposed to identify 25 ECG abnormalities plus the Normal Sinus Rhythm, summing up a total of 26 different classes to be detected, using the different combinations of ECG leads shown in table 3.

We first introduce the models' validation scheme and scoring rules applied. Next, we present noise reduction, signal processing and feature extraction. Subsequently, feature selection for dimensionality reduction is described. Then, we present training and validation processes for the proposed binary classification models setups. To conclude, we describe the One-versus-Rest classification approach used in order to provide a multi-class classification system intended to detect the distinct 26 cardiac categories previously mentioned.

The framework where we performed all these steps was MATLAB (R2021a, The MathWorks); the developed code, used both to train and validate the models, is available at https://github.com/sjimenezupv/itaca_upv.cinc2021.special_issue. On the other hand, the hardware used during this work was a computer with an Intel(R) Core(TM) i9-10900K CPU @ 3.70 GHz processor, 64 GB of RAM memory, an SSD drive with 1TB of available memory and an NVIDIA GeForce RTX 3080 GPU CUDA capable. This computer had Windows 10 Pro (64 bits) as the operating system installed. In this work, we will present our computational costs in terms of time used during the main three processes (feature extraction, and models training and testing) plus the performance of our system in terms of samples processed per second.

3.1. Validation scheme

We used 3-fold cross-validation with the available 88 253 training samples since the number of samples in the database was large enough and the training time could become unnecessarily high for a larger number of folds. Moreover, we selected the samples for each fold with no bias among all the distinct databases available, thus having the same number of samples coming from each database in each fold. This way, for each cross-validation fold, we used 58 835 (66.6%) samples for training and 29 418 (33.3%) for offline testing in order to assess the performance of the classification models, using the combinations of 12, 6, 4, 3, 2 and also 1 lead.

On the other hand, we report the results in the official test and validation datasets (also named online test and validation datasets) containing 36 266 and 6 630 12-lead ECG recordings, respectively. These results were provided by the challenge organizers using the official combinations of 12, 6, 4, 3 and 2 leads, not including any single-lead result like ours in the cross-validation. These online datasets should not be confused with the training dataset used during the offline cross-validation. The official validation and test datasets were unavailable to train nor validate any classification model in this work and were used only to report our models' scoring metrics described next.

3.2. Scoring

In this work, we report the *PhysioNet/Computing in Cardiology Challenge* 2021 scoring rule described in Reyna *et al* (2021). This scoring rule, named Challenge Metric (*CM*), uses a collection $C = [c_j]$ as a list of positive or negative diagnoses, computing a multi-class confusion matrix $A = [a_{ij}]$, where a_{ij} is the number of recordings that were classified as belonging to class c_i but actually belong to class c_j. Then, a matrix $W = [w_{ij}]$ weighted the confusion matrix based on the similarity of treatments or differences in risks. This way, an unnormalized score S is obtained using the next expression

$$S = \sum_{ij} w_{ij} * a_{ij}. \tag{1}$$

The score S is then normalized in order to give a value of 1 for a classifier that always outputs the true class or classes and a value of 0 for an inactive classifier that always outputs negative predictions for each class. This normalization gives as a result the final *CM* score value, achieved using the following ratio

$$CM = \frac{S - S_{inactive}}{S_{true} - S_{inactive}}, \tag{2}$$

where $S_{inactive}$ is the score for the inactive classifier and S_{true} is the ground-truth classifier score. This CM score could also give negative values if the ratio of false positives is high, thus showing a lower score than a classifier that returns only negative labels.

On the other hand, we also report the geometric mean among sensitivity and specificity both for each individual binary classifier and the global mean $\pm$ standard deviation for each lead combination. This score, named here G-metric, was used in order to select the binary classifiers with the best performance during the training and validation in our experimentation and corresponded to the next expression

$$G = \sqrt{Sensitivity * Specificity}. \tag{3}$$

We also report the Area Under the ROC Curve (AUROC) and F-measure values despite the fact that we did not use them to validate nor select any hyper-parameter or binary model during the training stage.

3.3. Signal preprocessing

The duration of the signals in the dataset was heterogeneous, with a mean lasting from 10 to 1800 s, whereas their sampling frequency was mostly 500 Hz (Perez Alday *et al* 2020, Reyna *et al* 2021, Zheng *et al* 2020a, 2020b). The signal preprocessing described below was applied to all the recordings regardless of their duration.

First, all ECG signals were resampled to 500 Hz if necessary, using the *resample* Matlab R2021a function, in order to homogenize the sampling frequency to the most common value in the employed database.

Next, we applied a Butterworth band-pass filter between 0.5 and 40 Hz, thus removing baseline wander artifacts and high-frequency noise such as powerline interference. Following this, we removed the first and last second of each signal in order to leave out the filtering stabilization stage.

Then, we removed aberrant signal artifacts mainly caused by sudden patients' movements, which appeared as abrupt changes in the signal level. To detect them, we used a 0.5 s sliding window and calculated their maximum and minimum values. If the difference among these values in neighboring windows exceeded a certain threshold, we considered them anomalous, and the signal segment within this window was set to zero.

Finally, we removed the signal segments beyond the first 15 s if they were available. Hence, we used short ECG segments and avoided big differences in the lasting time of the dataset samples, as previously done in Xiaoyu *et al* (2021), Wickramasinghe and Athif (2021), Aublin *et al* (2021).

3.4. Feature extraction

We automatically extracted 81 signal features from each of the available 12 ECG-leads in the training set, plus the age and sex from the recordings metadata, getting a total of 974 feature values from each sample. These extracted features were derived from the ventricular activity and mostly based on heart rhythm variability, QRS and T-waves patterns, plus the lower part of the spectral domain, being previously used in Zheng *et al* (2020a), Jiménez-Serrano *et al* (2017, 2021). Whereas part of the variables extracted were dependent of the heart rhythm variability, and this should give similar metrics for all ECG leads, they were extracted independently for each ECG channel. This allowed to take into account differences in heart rhythm detection for each ECG lead, as well as to maintain the structure of the rest of the lead-dependent variables.

To accomplish this task, firstly, for each lead, we extracted the RR sequence using a QRS detector based on the Pan and Tompkins algorithm and the first derivative of the ECG (Pan and Tompkins 1985). Then we filtered the outliers from the *RR* sequence using the mean $\pm$ three standard deviations as thresholds and obtained the first and second derivatives of that sequence, named here *RRd1* and *RRd2*. Moreover, we created a T-wave detector using a 300 ms window and an offset of 100 ms from the R-waves previously identified, getting the index of the maximum absolute value in this window as the location of the T-wave. With this information, we obtained the QT interval and other related features.

Next, we got both the QRS and T wave patterns for each lead using a $\pm$100 ms window over all the QRS and T wave detections.

On the other hand, we got Welch's power spectral density estimation for each lead in order to obtain some frequency-based features.

Using the above QRS and T waves marks, *RR* sequences, wave patterns and spectral information, we grouped in ten different categories the 81 signal features extracted from each lead detailed in table 4. The performance of both QRS and T-wave detectors is closely related to the final classification performance since 72 out of these 81 features depend on such marks.

Table 4. Detail of the ten feature categories extracted from each ECG lead, the number of features in each category, and the features description. Furthermore, age and sex were obtained from the recordings metadata.

Feature categories	Number of features	Features description
1—Waves voltages	4	Basic statistics over the R and T waves voltages (mean, standard deviation)
2—QT time	2	Basic statistics over the QT interval in milliseconds (mean, standard deviation)
3—QTc-time	3	Mean QT corrected duration with the formulas of Bazett (1920), Fridericia (1920) and Framingham (Sagie *et al* 1992)
4—QRS and T Pattern	25	Features based on the QRS and T patterns: Percentage of the amplitude of T wave respect the R wave, the sign of the R and T waves (positive or negatives), percentage of waves discards and RMSE during the R and T pattern definition, and maximum values for first and second derivatives of both patterns We also split the patterns into three parts: before the absolute maximum value, after the absolute maximum value and before the absolute minimum value, and after the minimum value. For each part, we got the maximum and minimum derivative values. Furthermore, we got a value indicating if the absolute maximum or minimum value was dominant in each pattern
5—*RR* stats	9	Basic statistics over the *RR*, *RRd1* and *RRd2* sequences selecting only those stats that presented significant differences among positive and negative samples: mean (*RR*, *RRd1*), standard deviation (*RR*, *RRd1*, *RRd2*), kurtosis (*RR*, *RRd1*, *RRd2*) and skewness (*RR*)
6—*RRd1* based	4	Features based on the *RRd1* sequence: RMSSD, pNN25, pNN50, pNN75, where pNNxx (Mietus *et al* 2002) denotes the percentage of intervals between normal beats exceeding 25, 50 and 75 ms
7—Poincaré's plot	7	Poincaré plot-based features using sequence *RRd1*: Maximum, minimum, mean, standard deviation, kurtosis and skewness of the distances among all the points plus the absolute difference between the maximum and minimum distance values
8—Lorenz's plot	8	Lorenz plot-based features using sequence *RRd2*: Angular variability, dispersion of the distance between points to the origin, and differences between 2 and 3 consecutive beats
9—Spectral features	17	Dominant frequency (*fdom*) using the Welch spectral density estimation method; spectral concentration (*SC*) in *fdom*$\pm$ 0.5 Hz in the periodogram normalized in the range [0, 1] and the sum of the normalized periodogram in steps of 2 Hz in the range [0, 28] Hz
10—Other features	2	Lempel-Ziv complexity of the *RR* time series after binarization using the median as threshold and Shannon entropy of the *RR* sequence
Total features/lead	81	

3.5. Dataset preprocessing

First, for each feature in the training fold, outliers exceeding three times the standard deviation above or below the median were replaced by these same limits. Next, if some sample contained a *NaN* value, due to a feature extraction error or the impossibility of obtaining such value, it was replaced by the median in the dataset for such feature. According to the previous two rules, 1.24% of the values were outliers, and 0.31% *NaN* values, been replaced for their corresponding limit or median value.

Lastly, for each training fold during the cross-validation, we performed a *z*-score in order to rescale the dataset, saving the mean (μ) and standard deviation (σ) of each feature by means of the following expression

$$z = \frac{x - \mu}{\sigma}, \tag{4}$$

where x denotes the original feature value, and z is the rescaled one. During the validation stage of our cross-validation, we used the saved μ and σ values in order to rescale in the same way the input feature values. Also, in the validation stage, we replaced the possible *NaN* values with the median values of the training fold since the binary classification models cannot deal with *NaN* values. Outlier values in validation were not replaced.

3.6. Feature selection

We applied a feature selection process using both supervised and unsupervised statistical filtering methods. To do this, we used as input the features corresponding to the available leads (see table 3). In addition, age and sex always were selected in order to avoid an empty set of features. Next, we performed a two-sample *Student's t-test* with an alpha value of 0.05 for each feature, taking into account if the sample belongs or not to the specified class, and all the features that did not pass the significance test were removed.

Student's t-test on large datasets are very sensitive to differences between measures and allow to identify variables with potential predictive value. This approach had the limitation of including many variables with correlated information. Therefore, for each possible pair of the remaining features, we removed one of them if their Pearson's correlation coefficient was greater or equal to 0.95.

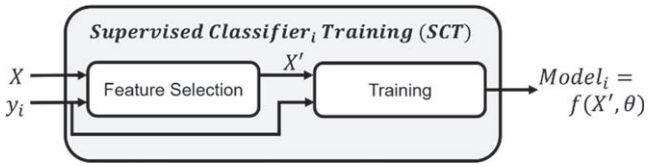

Figure 1. Summary diagram for the supervised classifier training (SCT) process.

3.7. Binary classification strategy

In this work, for each cardiac condition to be detected (table 2), a set of binary classifiers were trained and validated to finally select the one with the best performance. The selected model was included in the final multi-classification system, described in section 3.8, in order to give the corresponding label for a given class.

We differentiate two different types of binary classifiers trained and validated during this stage: supervised classifiers (SC) and hybrid classifiers (HC). The first ones are classic supervised models trained and validated with the whole training dataset. The last proposed hybrid model mixes an unsupervised k-means algorithm with the previous SC approach. Once all the binary classification models for a given class were validated, we selected the one that presented the higher G score.

A feature selection was performed for each of the distinct 26 ECG categories previously to the training of each binary classifier. Thus, each binary classifier presented next had associated a subset of features specifically selected for its own training and validation process.

The next sections detail the training processes for the binary SC and HC models, plus the selection process of the binary model (SC or HC) with better performance for each cardiac condition.

3.7.1. Supervised classifiers training (SCT)

The training for a given supervised binary classifier is depicted in figure 1. There, we use as input the dataset X and the ground truth labels vector y_i for a given class. Next, we performed the feature selection getting the subset of features X'.

Lastly, we performed a training and validation process for distinct types of classification models: feed forward neural networks (FFNN) and Naïve Bayes (NB), in our setup; then, we selected the one with the higher G score. As a result, we get a binary classification function f that uses θ as the model parameters obtained during the training process and the X' selected features as inputs needed.

Regarding the Naïve Bayes model, we used X' in order to train a binary classifier for each class using the *fitcnb* Matlab R2021a function. By default, this function models the predictor distribution within each class using a Gaussian distribution having some mean and standard deviation; and the prior class probability distribution as the relative frequency distribution of the classes in the data set. In this sense, the prior class probability distribution did not modify the results for this model, so finally, we used the default setup containing prior probabilities.

On the other hand, we trained two different FFNNs, both with an architecture of one hidden layer and using *tansig* as activation function: the first one made of 18 units and the last one with 32 units in the hidden layer. Furthermore, the output layer used *mapminmax* as activation function mapping the output in the range $[-1, 1]$. Apart from this, all the FFNN were trained with the default objects and parameters in the Matlab R2021a Deep Learning Toolbox, using *trainscg* (Scaled Conjugate Gradient) as the training function; the *useGPU* flag was switched on in order to use the available Graphics Processing Units to speed up the training, and the *showResources* flag was switched off.

Using as input X', 75% of training data was used to train the FFNNs. With the resting 25%, we selected for each FFNN the output threshold in the range $[-1, 1]$ that maximized the G metric score using steps of 0.005 in that range. This threshold was used as the cut-off point to classify the samples as positives or negatives. This way, we dealt with the problem of imbalanced classes in the dataset where the ratio of positive samples always was lower than negatives for each cardiac condition.

Finally, among the three SC trained during the SCT process (one NB plus two different FFNN), we chose the one that presented a higher G score value, named here G_{SCT}.

3.7.2. Hybrid classifiers training (HCT)

The Hybrid binary classification approach is depicted in figure 2. First, we perform a feature selection from the input dataset X and the ground truth labels for a given class y_i in order to get the subset of features X'. Then, the core of our proposal is based on an unsupervised machine learning technique (k-means) that clusters in three different groups the input training set X' in order to get three different cluster centroids that we named $C_{\{1,2,3\}}$,

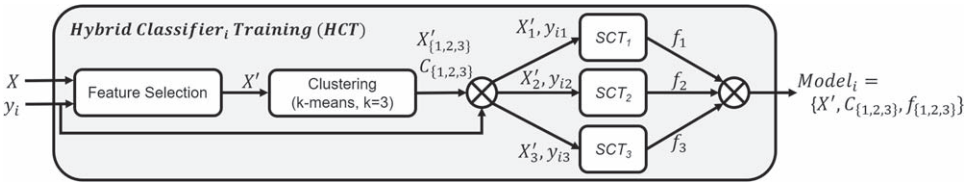

Figure 2. Summary diagram for the hybrid classifier training (HCT) process.

and three subsets of samples associated to each centroid named $X'_{\{1,2,3\}}$. Finally, for each subset of samples in $X'_{\{1,2,3\}}$, we performed the SCT process described previously in order to get an SC with the best classification performance possible for each data cluster. Finally, as the output of the HCT process, we saved the selected feature indexes $X'_{\{1,2,3\}}$, the cluster centroids $C_{\{1,2,3\}}$, and the SC binary classifiers corresponding to each cluster, named here $f_{\{1,2,3\}}$.

Thus, for classifying a new unseen sample x, we first remove the unselected features from the tuple of values giving as results the new tuple x'; then we get the euclidean distance from this sample to each of the three cluster centroids in $C_{\{1,2,3\}}$ and choose the cluster that minimizes this distance; finally, we use the SC model in $f_{\{1,2,3\}}$ associated to such cluster in order to give the binary response, i.e. use the NB or FFNN trained previously only with the data belonging to this cluster. This classification process is expressed as follows, where $\hat{y}_i$ is the label predicted by our hybrid system

$$\hat{y}_i = f_{\{c_{x'}\}}(x') | c_{x'} = argmin\ (distance(C_{\{1,2,3\}},\ x')). \tag{5}$$

We chose the arbitrary value of $k = 3$ in order to obtain a compromise between training time and accuracy. Regarding the training time, it has to be noted that k multiplies the total number (#) of SC models to be trained and evaluated during the HCT proposed approach following the next formula

$$HCT(\#SC\ Trained\ Models) = k * \#classes. \tag{6}$$

The accuracy of the supervised binary classifiers associated with each cluster also depends on the number of input samples provided during the training process. Since some classes had a lower rate of positive samples, a k value greater than 3 resulted in clusters with few positive samples available as inputs, being 3 the optimal value tested.

Once the number of clusters was selected, we used as k-means training parameters 200 iterations as maximum and 10 different replicates in order to try to achieve the maximum separation among clusters and their corresponding centroids.

On the other hand, to assess the accuracy of the model, we took into account the G metric of each supervised classifier associated with one cluster, weighted to the number of samples in such cluster with the next expression

$$G_{HCT} = \sum_{i=1}^{3} G(f_i) * \left(\frac{n_i}{n}\right), \tag{7}$$

where $G(f_i)$ and n_i denote the G value and the number of samples of the classifier belonging to cluster i, and n is the total number of samples in the training fold. Thereby, large clusters had greater importance in the G_{HCT} score of the model, while the reverse is true for small clusters. Furthermore, if some cluster had less than 40 samples, or less than 40 positive samples, we did not train nor validate the corresponding f_i SC, giving a $G(f_i)$ value of zero to this cluster since we cannot assess a correct binary classification with this data.

3.7.3. Model selection

Once we trained and validated both the SC and the HC classification models approaches, we selected only one binary classifier in order to be used in the whole multi-classification system. To do so, we chose the one with the highest G value, following the next expression

$$Classifier_i = Classifier(argmax(G_{SCT},\ G_{HCT})), \tag{8}$$

where G_{SCT} and G_{HCT} are the G score values corresponding to the selected binary models during the SCT and our proposed HCT hybrid approach.

3.8. One-versus-rest classification approach

Once we selected the best trained and validated binary classifier for each of the 26 cardiac categories to be detected, we built a final multi-classification system depicted in figure 3. This approach is also known as a One-versus-Rest classification model, where the samples of the dataset are labelled as positive or negative using each

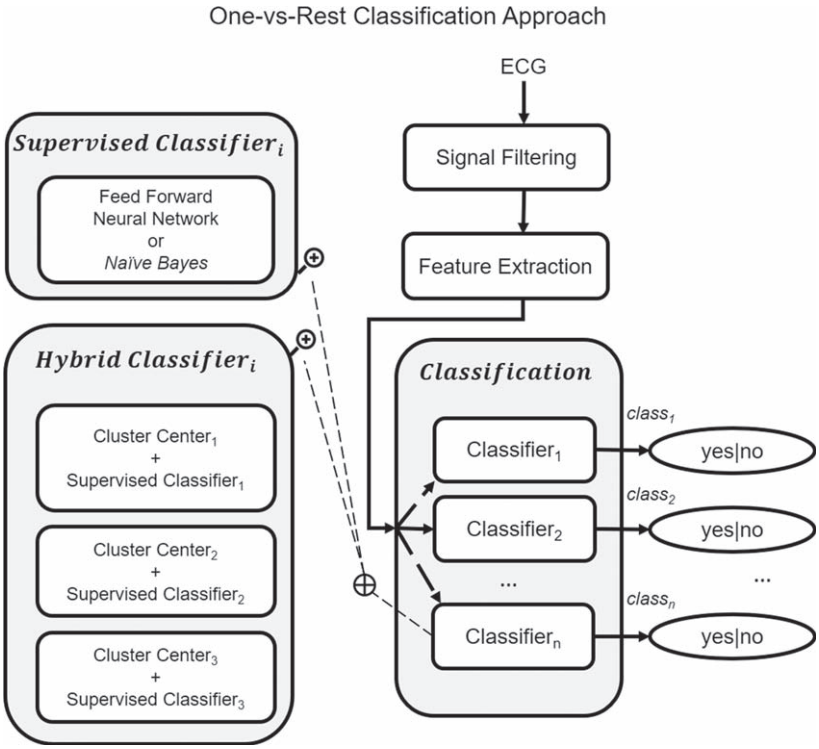

Figure 3. Summary diagram of the One-versus-Rest classification approach.

binary classifier for each known class. Thus, the whole One-versus-Rest classification model will give a binary response indicating if an unseen sample belongs or not to each of the 26 classes previously used during the training process. Furthermore, as detailed previously, each binary classifier uses the selected set of features that best fits its own classification problem. Consequently, each binary model solves an independent classification problem in the whole classification system, being possible to assign a new sample to none, one or more than one class. This is especially useful in this work since some samples could not belong to any cardiac category previously trained, or, on the contrary, belong to more than one cardiac condition.

Finally, we made an exception during the training of the right axis deviation (RAD) class, where we always used an NB binary classifier. The main reason for this was that the SCT and HCT classification approaches gave us lower and inconsistent G scores among distinct validation folds using both FFNN or the hybrid binary classifier for RAD. On the other hand, NB presented a consistent and higher G score for RAD during the validation process, not overfitting the input dataset.

4. Results

4.1. Computational costs

Each of the three folds used during the cross-validation contained ~58 800 samples for training and ~29 400 samples for testing. With this sampling size as context, we timed three main processes of our experimentation set: signal processing and feature extraction, model training (feature selection included), and model testing (time taken in classifying samples unseen by the models).

The time needed to perform the signal processing and feature extraction for the 12-lead ECG records was 566.79 $\pm$ 24.95 s (9.45 $\pm$ 0.42 min), giving us a performance of 103.74 $\pm$ 4.16 samples per second using the hardware described previously.

Next, we measured the time needed to perform the training, validation and selection of the corresponding One-versus-Rest multi-classifiers for each leads set combination. Finally, we performed the same operation for timing the testing of the unseen samples by the models. Figure 4 shows the trend of the computational time needed for both processes being the training process the more time consuming, whereas the testing process presents a linear behavior with a small slope with respect to the leads number.

Finally, figure 5 shows the performance during the training and testing of the models in terms of the number of samples processed per second. Best performance corresponds to the single-lead classifier, with 21.20 $\pm$ 0.35 samples s^{-1} during training and 18.02 $\pm$ 0.18 samples s^{-1} during testing. As expected, lower performance

Figure 4. Trend of mean time for training and test of each ECG multi-classifier during the 3-fold cross-validation on the public training set.

Figure 5. Mean performance for each of the ECG multi-classifier models during the 3-fold cross-validation on the public training set, in terms of the number of samples processed per second.

results take place using the 12-leads ECG classifier, with 4.86 $\pm$ 0.02 samples s^{-1} for training and 7.46 $\pm$ 0.07 samples s^{-1} during testing. It has to be noted that each time a sample is tested, the challenge version of the code must read the sample file from the disk, adding an overhead that does not exist during the training process.

4.2. Feature selection analysis

First, table 5 shows the percentage of features selected from the whole dataset in order to train the binary classifiers that finally compose the One-versus-Rest multi-classifier models for each leads set combination. The method employed selected a mean of 60 $\pm$ 10% of available features (removing the remaining for the corresponding binary classifier) with no significant differences among the leads combinations used.

Second and last, in order to know the performance and behavior of our feature selection method among the different ECG-leads, we got the percentage of the selected features for each binary classifier and grouped them by the lead from which they were obtained; these results are shown in table 6. During the training of the models using 12-leads, only features extracted from lead V1 presented a slightly bigger selection percentage with a value of 8.81 $\pm$ 0.76%. In the case of the models trained with 4 and 3-leads, features extracted from lead V2 presented a lower selection percentage with values of 24.15 $\pm$ 2.28% and 32.01 $\pm$ 2.72% each. Other leads combinations

Table 5. Percentage (mean ±standard deviation) of the selected features for each binary classifier during the 3-fold cross-validation on the public training set.

Leads	Selected features [%]
12	59.62 ± 9.99
6	61.01 ± 10.73
4	60.24 ± 10.15
3	60.35 ± 9.88
2	61.43 ± 10.89
1	61.62 ± 10.48

Table 6. Percentage (mean ± standard deviation) of the selected features for the binary classifiers used during the 3-fold cross-validation, grouped by ECG-leads. Age and sex were always selected.

Selected features by lead [%]	12-leads models	6-leads models	4-leads models	3-leads models	2-leads models
I	8.48 ± 0.67	16.55 ± 1.25	25.05 ± 1.79	33.22 ± 2.18	48.86 ± 2.61
II	8.50 ± 0.57	16.60 ± 0.90	25.15 ± 1.68	33.37 ± 2.33	49.07 ± 2.75
III	8.32 ± 0.72	16.25 ± 1.34	24.59 ± 1.86	—	—
aVR	8.57 ± 0.61	16.74 ± 1.03	—	—	—
aVL	8.69 ± 0.72	16.96 ± 1.20	—	—	—
aVF	8.30 ± 0.81	16.20 ± 1.44	—	—	—
V1	8.81 ± 0.76	—	—	—	—
V2	8.06 ± 0.73	—	24.15 ± 2.28	32.01 ± 2.72	—
V2	7.98 ± 0.62	—	—	—	—
V4	8.06 ± 0.50	—	—	—	—
V5	7.86 ± 0.44	—	—	—	—
V6	8.01 ± 0.42	—	—	—	—

Table 7. Percentage of the distinct types of models selected for the binary classifiers depending on the leads combination used for training during the 3-fold cross-validation on the public training set.

#Leads	Hybrid [%]	FFNN [%]	NB [%]
12	38.46	57.69	3.85
6	34.62	55.13	10.26
4	33.33	58.97	7.69
3	30.77	62.82	6.41
2	28.21	64.10	7.69
1	32.05	58.97	8.97

did not present significant differences in the percentages among features corresponding to distinct leads, being this ratio proportional to the number of leads employed.

In summary, the results presented in tables 5 and 6 show that a mean of 40 ± 10% of the extracted features was filtered using our selection method and that the ratio of these features was balanced among the available leads in each binary classification model.

4.3. Model selection analysis

Percentages of the distinct types of models selected for the binary classifiers depending on the leads combinations used for training are shown in table 7. More than 50% of the time, an FFNN model was selected, whereas, in a range between 28% and 38% of the time, our proposed HC improved the performance during the validation process, and thus, was selected as the final binary classifier for a given cardiac condition. Lastly, NB was selected less than 11% of the time, mainly where the other models did not perform well.

Moreover, we detail the percentage of the distinct type of models selected for each binary classifier in table 8. For ten cardiac conditions, only FFNN was selected, whereas, in the other categories, a mix of models was

Table 8. Percentage of the distinct type of models selected for each binary classifier during the 3-fold cross-validation on the public training set with all the ECG-lead combinations.

Class	Hybrid [%]	FFNN [%]	NB [%]
AF	—	100	—
AFL	—	100	—
BBB	55.56	—	44.44
Brady	—	88.89	11.11
CLBBB|LBBB	83.33	—	16.67
CRBBB|RBBB	—	100	—
IAVB	33.33	66.67	—
IRBBB	94.44	5.56	—
LAD	50.00	50.00	—
LAnFB	11.11	88.89	—
LPR	11.11	72.22	16.67
LQRSV	83.33	16.67	—
LQT	100	—	—
NSIVCB	33.33	66.67	—
NSR	66.67	33.33	—
PAC|SVPB	—	100	—
PR	94.44	5.56	—
PRWP	94.44	—	5.56
PVC|VPB	—	100	—
QAb	44.44	55.56	—
RAD	—	—	100
SA	—	100	—
SB	—	100	—
STach	—	100	—
TAb	—	100	—
TInv	—	100	—

Table 9. Challenge scores metric (*CM*) for our final multi-lead classification models using 3-fold cross-validation on the public training set, one-time scoring on the official and hidden validation set, and one-time scoring on each official and hidden test sets as well as on the entire official hidden test set. The competition organizers did not evaluate the single lead configuration.

#Leads	Training set (cross-validation) (*CM*)	Validation set (*CM*)	CPSC test set (*CM*)	G12EC test set (*CM*)	Undisclosed test set (*CM*)	UMich test set (*CM*)	Entire test set (*CM*)
12	0.435 ± 0.009	0.440	0.301	0.465	0.284	0.418	0.388
6	0.402 ± 0.003	0.431	0.281	0.457	0.262	0.410	0.376
4	0.421 ± 0.001	0.435	0.279	0.457	0.286	0.418	0.387
3	0.420 ± 0.004	0.432	0.278	0.457	0.282	0.415	0.384
2	0.414 ± 0.005	0.428	0.268	0.459	0.252	0.407	0.373
1	0.388 ± 0.005	—	—	—	—	—	—

selected depending on the fold and the number of ECG-leads used for training. NB was used 100% of the time only once in RAD since we used a rule to do so defined previously.

4.4. Model scoring analysis

Best results using the *CM* in the entire official hidden test set had a value of 0.388 using 12 leads. However, the lower score obtained in the same entire test set had a *CM* value of 0.373 using only two leads, with a slight difference of 0.015 with respect to the first one. On the other hand, using the 3-fold cross-validation in the public training set, we got a *CM* value of 0.435 ± 0.009 using 12 leads and 0.388 ± 0.005 using a single lead. Nonetheless, the differences of the 6, 4, 3 and 2 leads configuration were also lower comparing their *CM* values with the 12 leads combination, being 0.033 the maximum. Table 9 shows the whole results set using the *CM* score during our cross-validation and in the different hidden test datasets. The AUROC mean values in the test dataset were 0.86 in CPSC, 0.81 in G12EC, 0.84 in the undisclosed database and 0.82 in the UMich test set, with no significant differences among the distinct lead combinations.

Table 10 shows the mean and standard deviation of different performance metrics in the classification of the 26 scored classes in the challenge during the cross-validation in the public training set, where a higher *G* value of 0.80 was achieved using 12 leads, followed by a *G* value of 0.79 using both 4 and 3 leads, and 0.78 using both 6

Table 10. Mean±standard deviation of other performance metrics among the classification of the 26 scored classes for our final selected multi-lead classification models using 3-fold cross-validation on the public training set: Area Under the ROC Curve, *F*-measure, Sensitivity, Specificity and *G* metric.

#Leads	AUROC	*F*-measure	Sensitivity	Specificity	*G*
12	0.817 ± 0.008	0.286 ± 0.001	0.817 ± 0.003	0.795 ± 0.007	0.804 ± 0.004
6	0.811 ± 0.017	0.261 ± 0.002	0.784 ± 0.005	0.779 ± 0.005	0.777 ± 0.001
4	0.828 ± 0.014	0.270 ± 0.002	0.800 ± 0.005	0.787 ± 0.002	0.788 ± 0.002
3	0.837 ± 0.009	0.269 ± 0.002	0.801 ± 0.002	0.784 ± 0.004	0.786 ± 0.002
2	0.823 ± 0.011	0.262 ± 0.004	0.787 ± 0.004	0.779 ± 0.004	0.775 ± 0.002
1	0.784 ± 0.007	0.240 ± 0.001	0.751 ± 0.006	0.757 ± 0.006	0.744 ± 0.005

Table 11. *G* metric values for the single binary classifiers, using 3-fold cross-validation on the public training set.

Class	*G* (12-leads)	*G* (6-leads)	*G* (4-leads)	*G* (3-leads)	*G* (2-leads)	*G* (1-leads)
AF	0.86 ± 0.00	0.87 ± 0.01	0.86 ± 0.00	0.87 ± 0.00	0.87 ± 0.01	0.87 ± 0.00
AFL	0.87 ± 0.01	0.86 ± 0.01	0.87 ± 0.00	0.86 ± 0.00	0.85 ± 0.01	0.84 ± 0.00
BBB	0.76 ± 0.01	0.73 ± 0.04	0.75 ± 0.03	0.72 ± 0.00	0.72 ± 0.02	0.68 ± 0.01
Brady	0.75 ± 0.04	0.73 ± 0.03	0.78 ± 0.01	0.78 ± 0.02	0.79 ± 0.05	0.73 ± 0.03
CLBBB|LBBB	0.91 ± 0.01	0.90 ± 0.00	0.90 ± 0.01	0.89 ± 0.01	0.88 ± 0.01	0.87 ± 0.02
CRBBB|RBBB	0.91 ± 0.01	0.85 ± 0.00	0.88 ± 0.01	0.88 ± 0.01	0.86 ± 0.01	0.85 ± 0.00
IAVB	0.75 ± 0.03	0.78 ± 0.01	0.76 ± 0.00	0.77 ± 0.01	0.78 ± 0.00	0.75 ± 0.00
IRBBB	0.80 ± 0.01	0.69 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.68 ± 0.01	0.66 ± 0.01
LAD	0.87 ± 0.00	0.86 ± 0.00	0.83 ± 0.02	0.83 ± 0.00	0.83 ± 0.00	0.66 ± 0.00
LAnFB	0.90 ± 0.01	0.91 ± 0.00	0.90 ± 0.00	0.90 ± 0.00	0.91 ± 0.01	0.66 ± 0.02
LPR	0.66 ± 0.03	0.66 ± 0.02	0.71 ± 0.02	0.70 ± 0.00	0.69 ± 0.02	0.69 ± 0.03
LQRSV	0.79 ± 0.01	0.76 ± 0.00	0.76 ± 0.01	0.78 ± 0.01	0.77 ± 0.01	0.70 ± 0.01
LQT	0.76 ± 0.01	0.74 ± 0.02	0.75 ± 0.03	0.76 ± 0.00	0.74 ± 0.00	0.73 ± 0.01
NSIVCB	0.69 ± 0.01	0.66 ± 0.01	0.68 ± 0.00	0.69 ± 0.00	0.69 ± 0.00	0.68 ± 0.01
NSR	0.80 ± 0.01	0.79 ± 0.00	0.79 ± 0.01	0.79 ± 0.01	0.80 ± 0.00	0.79 ± 0.01
PAC|SVPB	0.80 ± 0.01	0.80 ± 0.00	0.79 ± 0.01	0.79 ± 0.01	0.80 ± 0.00	0.78 ± 0.00
PR	0.89 ± 0.01	0.86 ± 0.00	0.88 ± 0.02	0.88 ± 0.00	0.87 ± 0.01	0.84 ± 0.00
PRWP	0.77 ± 0.03	0.61 ± 0.04	0.74 ± 0.00	0.76 ± 0.01	0.67 ± 0.00	0.64 ± 0.00
PVC|VPB	0.80 ± 0.03	0.78 ± 0.00	0.79 ± 0.01	0.79 ± 0.02	0.78 ± 0.02	0.76 ± 0.01
QAb	0.69 ± 0.01	0.68 ± 0.01	0.68 ± 0.01	0.69 ± 0.01	0.63 ± 0.02	0.63 ± 0.01
RAD	0.66 ± 0.01	0.56 ± 0.00	0.46 ± 0.01	0.44 ± 0.02	0.39 ± 0.02	0.30 ± 0.02
SA	0.84 ± 0.01	0.83 ± 0.01	0.84 ± 0.00	0.85 ± 0.00	0.84 ± 0.01	0.86 ± 0.01
SB	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00
STach	0.94 ± 0.00	0.94 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.94 ± 0.00
TAb	0.75 ± 0.00	0.72 ± 0.00	0.72 ± 0.00	0.72 ± 0.01	0.72 ± 0.00	0.69 ± 0.00
TInv	0.76 ± 0.00	0.72 ± 0.01	0.74 ± 0.00	0.72 ± 0.01	0.72 ± 0.01	0.70 ± 0.01

and 2 leads. The single lead classification model got a *G* score of 0.74 with a balanced ratio among sensitivity and specificity of 0.75 and 0.76.

In table 11, we show the results achieved for the individual binary classifiers during the cross-validation in the public training set, where the next ten different cardiac conditions had *G* scores equal to or greater than 0.85 in some lead combinations: AF, AFL, CLBBB|LBBB, CRBBB|RBBB, LAD, LAnFB, PR, SA, SB and STach.

Finally, figure 6 shows the boxplots of the *G* score values for each of the cardiac conditions during the cross-validation in the public training set, using the values obtained with all the leads combinations. There, we observe that AF, AFL, CLBBB|LBBB, PR, SB and STach present *G* score values higher than 0.85 in all the leads combinations with no significant differences among them.

5. Discussion

As expected, the best performance was achieved with 12 leads, which suggests that the standard 12-lead system should not be replaced by reduced lead systems in clinical practice. However, performance decrease using fewer leads was low, according to the reported results of *CM* and *G* scores. These results show a great potential of reduced lead sets out of the clinical environment, e.g. intended for massive screenings and monitoring for early detection. It should also be mentioned that classification using four and three leads outperformed the ones using six leads. Although at a first glance this seems surprising, these results can be well explained by the fact that,

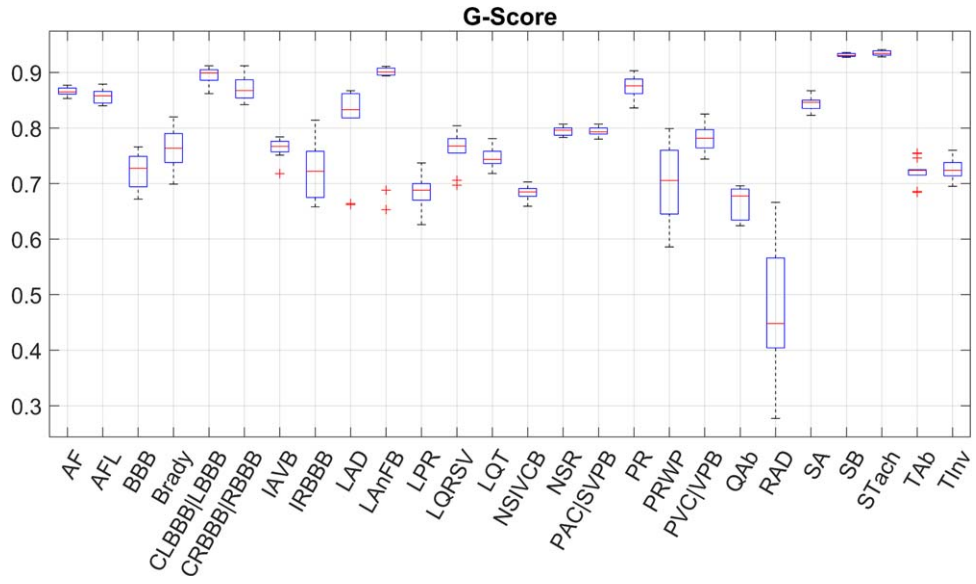

Figure 6. Boxplot of the *G* metric values for each single binary classifier, using the values obtained with all the lead combinations during the 3-fold cross-validation on the public training set.

among the limb leads that contained the 6-lead subset, only two of them are independent; indeed, the 6-lead limb system is recorded using only three electrodes. On the other hand, the 3- and 4-lead subsets add the precordial lead V2, which is independent of the limb leads and hence, adds clinically valuable information. According to this, performance using the 6-lead subset should be contrasted with the 2-lead subset for a more fair comparison. In fact, models with 2 leads had the closest scoring results to the ones obtained with 6 leads (we got *CM* values of 0.373 and 0.376 in the hidden test set, and mean *G* values of 0.775 and 0.777 during the 3-fold cross-validation in the public training set, both using 2 and 6 leads, respectively). These results also suggest that more important than increasing the number of leads is to include leads that contain complementary information.

5.1. Results with one lead and suitability for smart devices

Focusing on the outcomes using only lead I, during the cross-validation, we got mean *CM* and *G* values of 0.39 and 0.74 each using a single lead, whereas using the whole 12-leads set, the same scores were 0.43 and 0.80, respectively. Besides the fact that the performance decreased, these results are of high interest, as this allows the ECG recording and automated diagnosis in a number of daily situations where the standard 12-lead would not be possible (e.g. using a wearable or smartwatch, handling a smartphone or when driving with a smart steering wheel, just to mention a few). Furthermore, it should be highlighted that, for some cardiac conditions, classification performance using only lead I equaled or even improved classification with 12 leads. Specifically, *G* values for single binary classifiers during the 3-fold cross-validation in the training set were: 0.86 with 12-leads, 0.87 with one lead for AF; 0.75 with 12-leads, 0.75 with one lead for IAVB; 0.66 with 12-leads, 0.69 with one lead for LPR; 0.84 with 12-leads, 0.86 with one lead for SA; 0.93 with 12-leads, 0.93 with one lead for SB; and 0.94 with 12-leads, 0.94 with one lead for STach. Most of these single binary classification results even improve using the 2-leads classification model.

However, this work also shows that some cardiac conditions need information from precordial leads to be better detected, such as BBB, LAD, LAnFB, LQRSV, PRWP and RAD, where the 12 leads configuration outperformed clearly with respect to other configurations. This suggests that this approach should be used cautiously when intended to detect cardiac conditions with low performance in our results.

Despite the benefits of automated ECG monitoring with wearables and smart devices, there are still some inaccuracies in the diagnosis. Therefore, special care should be taken when presenting the result to the user in a non-clinical scenario. On the one hand, false negatives could lead to a false sense of security. On the other hand, false positives could trigger an unnecessary visit to the doctor that, when handled massively, could lead absurdly to the collapse of health centers. Future optimization of these classifiers, e.g. by adding more disease-specific features and/or clinical rules, could prevent these inconveniences.

From a computational point of view, the average classification time with one lead halved the average classification time with 12 leads in the testing stage. Therefore, smart devices such as wearables would also benefit from lower computational costs and battery consumption.

5.2. Improvements of the work reported in computing in cardiology 2021

An initial version of this work was previously done and reported in the *Computing in Cardiology 2021* congress in the context of the *Physionet Challenge* (Jiménez-Serrano *et al* 2021). However, the current work presents important changes that lead us to improvements in the performance of all the scoring metrics. Regarding the *CM* score, in Jiménez-Serrano *et al* (2021) we got values of 0.34, 0.34, 0.27, 0.30 and 0.34 using the combinations of 12, 6, 4, 3 and 2 leads in the official hidden test set; whereas in this work, we improved the same scores with values of 0.39, 0.38, 0.39, 0.38 and 0.37, respectively. The same happened with the *G* score during the 3-fold cross-validation in the public training set, obtaining in Jiménez-Serrano *et al* (2021) values of 0.76, 0.74, 0.69, 0.69, 0.74 and 0.71 using combinations of 12, 6, 4, 3, 2 and 1 lead, whereas this work improved the previous ones with values of 0.80, 0.78, 0.79, 0.79, 0.78 and 0.74, respectively. The changes we made to achieve these improvements with respect to the previous work in Jiménez-Serrano *et al* (2021) are as follows:

After the signal filtering stage, we used only 15 s of ECG recording, as other participants did (Xiaoyu *et al* 2021; Wickramasinghe1 and Athif 2021, Aublin *et al* 2021), improving the processing time and avoiding some unwanted collateral effects of long records. The main problem in this sense is that a long recording scores the same as a short one, and in the database used, short recordings are the majority. Another problem that we found with long recordings was that the probability of founding more cardiac conditions increases, but the fixed labelling of the records could be wrong in different parts of the records, e.g. paroxysmal atrial fibrillation could appear and disappear in the same long ECG recording. Thus, splitting the signal into different parts with the same labelling or using a long sliding window during the feature extraction could become a no sense operation for training the classifiers.

This way, the authors think that the automatic classification of short ECG recordings no longer than 30 s have to be addressed in a different way than the classification of longer ECG recordings, as well as the labelling of each segment of signal for long recordings. Furthermore, since one of the aims of this work is to achieve classification models suitable for smart devices and able to deal with short ECG registers, this work assesses our classification scores in recordings lasting no more than 15 s. Moreover, one more difference from Jiménez-Serrano *et al* (2021) is that in this work, we no longer used features useful for long ECG recordings, such as stats over long signals applying a sliding window. Thus, in total, we removed 46 previous features of this type. On the other hand, we added more specific QRS and T pattern features that improved our classification scoring.

Next, we added data preprocessing in order to use only 26 classes mixing the cardiac conditions that weighted the same, and also, we avoided using samples that did not belong to any of the 26 cardiac categories for training. We also changed the threshold value in the second stage of the feature selection respect (Jiménez-Serrano *et al* 2021), from 0.90 to 0.95 in the correlation coefficient among two different features, slightly improving the validation results.

Finally, we added a mixed approach using supervised and unsupervised machine learning techniques, where each binary classifier could be made of an FFNN, NB, or HC, whereas in Jiménez-Serrano *et al* (2021), we only used FFNN as binary classifiers. The proposed hybrid classification system is based on an unsupervised k-means algorithm, where 3 cluster centroids are looked for, and an FFNN or NB is associated with each of them. Moreover, a model selection system for these binary classifiers was created for the training process.

To conclude the description of changes that we made regarding (Jiménez-Serrano *et al* 2021), we fixed some minor bugs and inaccuracies, releasing the latest version of the code in https://github.com/sjimenezupv/itaca_upv.cinc2021.special_issue. As a result, we improved all the classification scores with respect to the previous work.

5.3. Comparison with other works

As far as the authors' knowledge, other works were proposed in order to address this multi-leads classification problem with the same database during the *Computing in Cardiology 2021*, mostly based on deep learning. As observed in the comparative table in Reyna *et al* (2021), other competitors in the Challenge obtained substantially superior scores than those presented in this manuscript. These works presented different approaches to those used here.

In Xiaoyu *et al* (2021), a model based on SE-ResNet was built incorporating peak detection as a self-supervised auxiliary task. The *CM* scores were 0.55, 0.58, 0.58, 0.57 and 0.57 using the combinations of 12, 6, 4, 3 and 2 leads in the official hidden test set.

In Wickramasinghe and Athif (2021), two separated deep CNNs were trained using ECG segments of 20 s and their Fast Fourier Transform. The *CM* scores were 0.55, 0.51, 0.56, 0.55 and 0.56 using the combinations of 12, 6, 4, 3 and 2 leads in the official hidden test set.

In Aublin *et al* (2021), a voting system was designed, where ECG segments of 10 s feed a large deep CNN for each available lead. The *CM* scores were 0.48, 0.47, 0.47, 0.47 and 0.46 using the combinations of 12, 6, 4, 3 and 2 leads in the official hidden test set.

The mentioned works had promising results, but none of them evaluated the performance in a single lead with the challenge database as we did. Furthermore, our set of extracted features could be evaluated in an easy way by expert physicians in order to understand our models' results. Finally, we also highlight that our system was designed in a modular way, allowing future upgrades (or even downgrades), such as adding new classes, features and classification models to the training and validation system in an easy and clean way.

6. Conclusion

We presented and evaluated a methodology for multiple cardiac disease detection through ECG registers that combines feature extraction and selection, and a One-versus-Rest classification approach using FFNN, NB and a novel Hybrid approach as binary classifiers. Interestingly, after a systematic analysis, the classification results using only one or two leads were not far from the results with twelve leads, showing lower computational costs and being more suitable for wearables. Furthermore, for some individual cardiac conditions, using one or two ECG leads showed equal or better score values than the others leads setups. Improving the identification of some cardiac rhythms by incorporating more specific features or clinical rules for those cases where the performance was low should be an interesting direction to explore in future works.

Acknowledgments

This work was supported by PID2019-109547RB-I00 (National Research Program, Ministerio de Ciencia e Innovación, Spanish Government) and CIBERCV CB16/11/00486 (Instituto de Salud Carlos III).

ORCID iDs

Santiago Jiménez-Serrano https://orcid.org/0000-0003-2917-6053

References

Alexakis C, Nyongesa H, Saatchi R, Harris N, Davies C, Emery C, Ireland R and Heller S 2003 Feature extraction and classification of electrocardiogram (ECG) signals related to hypoglycemia *Computing in Cardiology 2003* 30, 537–40

Aqil M, Jbari A and Bourouhou A 2015 Evaluation of time-frequency and wavelet analysis of ECG signals *Third World Conf. on Complex Systems (WCCS), 2015, pp 1–5* (https://doi.org/10.1109/ICoCS.2015.7483229)

Aublin P, Ben Ammar M, Achache N, FixBenahmed M, El Hichami A, Barret M, Fix J and Oster J 2021 Cardiac abnormality detection based on an ensemble voting of single-lead classifier predictions *Computing in Cardiology 2021* 48

Bazett J C 1920 An analysis of time relation of electrocardiograms *Heart* **7** 353–67

Bickerton M and Pooler A 2019 Misplaced ECG electrodes and the need for continuing training *Br. J. Cardiac Nursing* **14** 123–32

Bousseljot R, Kreiseler D and Schnabel A 1995 Nutzung der EKG-signaldatenbank cardiodat der PTB über das Internet *Biomed. Tech. Biomed. Eng.* **40** 317–8

Chazal P, O'Dwyer M and Reilly R B 2004 Automatic classification of heartbeats using ECG morphology and heartbeat interval features *IEEE Trans. Biomed. Eng.* **51** 1196–206

Chen Y, Wang X, Jung Y, Abedi V, Zand R, Bikak M and Adibuzzaman M 2018 Classification of short single-lead electrocardiograms (ECGs) for atrial fibrillation detection using piecewise linear spline and xgboost *Physiol. Meas.* **39** 104006

Cheng Z, Deng H, Cheng K, Chen T, Gao P, Yu M and Fang Q 2013 The amplitude of fibrillatory waves on leads aVF and V1 predicting the recurrence of persistent atrial fibrillation patients who underwent catheter ablation *Ann. Noninvasive Electrocardiol.* **18** 352–8

Chow G V, Marine J E and Fleg J L 2012 Epidemiology of arrhythmias and conduction disorders in older adults *Clin. Geriatric Med.* **28** 539–53

Christov I, Gómez-Herrero G, Krasteva V, Jekova I, Gotchev A and Egiazarian K 2006 Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification *Med. Eng. Phys.* **28** 876–87

Dai H, Hwang H G and Tseng V S 2021 Convolutional neural network based automatic screening tool for cardiovascular diseases using different intervals of ECG signals *Comput. Methods Programs Biomed.* **203** 106035

Deshpande A and Birnbaum Y 2014 ST-segment elevation: distinguishing ST elevation myocardial infarction from ST elevation secondary to nonischemic etiologies *World J. Cardiol.* **6** 1067–79

Dunn J, Runge R and Snyder M 2018 Wearables and the medical revolution *Personalized Med.* **15** 429–48

Fridericia L S 1920 Dir Systolendaeur in Elektrokardiogram bei normalen menchen und bei herzkranken *Acta Med. Scandinavica* **53** 469–86

Gaziano T A, Bitton A, Anand S, Abrahams-Gessel S and Murphy A 2010 Growing epidemic of coronary heart disease in low and middle income countries *Curr. Problems Cardiol.* **35** 72–115

Georgiou K, Larentzakis A V, Khamis N N, Alsuhaibani G I, Alaska Y A and Giallafos E J 2018 Can wearable devices accurately measure heart rate variability? a systematic review *Folia Med.* **60** 7–20

Bortolan G, Christov I and Simova I 2021 Potential of rule-based methods and deep learning architectures for ECG Diagnostics *Diagn.* **11** 1–13

Hagan R, Gillan C J and Mallett F 2021 Comparison of machine learning methods for the classification of cardiovascular disease *Inform. Med. Unlocked* **24** 100606

Zhang H, Liu C, Zhang Z, Xing Y, Liu X, Dong R, He Y, Xia L and Liu F 2021 Recurrence plot-based approach for cardiac arrhythmia classification using inception-resnet-v2 *Front. Physiol.* **12** 1–13

Jiménez-Serrano S, Rodrigo M, Calvo C J, Castells F and Millet J 2021 Multiple cardiac disease detection from minimal-lead ecg combining feedforward neural networks with a one-versus-rest approach *Computing in Cardiology 2021* 48, 2021
Jiménez-Serrano S, Yagüe-Mayans J, Simarro-Mondéjar E, Calvo C J, Castells F and Millet J 2017 Atrial fibrillation detection using feedforward neural networks and automatically extracted signal features *Computing in Cardiology 2017* 44, 131–4
Jo Y Y *et al* 2021 Detection and classification of arrhythmia using an explainable deep learning model *J. Electrocardiol.* **67** 124–32
Kligfield P 2002 The centennial of the einthoven electrocardiogram *J. Electrocardiol.* **35** 123–9
Kostka P and Tkacz E 2007 Feature extraction and selection algorithms in biomedical data classifiers based on time-frequency and principle component analysis *11th Mediterranean Conf. on Medical and Biomedical Engineering and Computing 2007. IFMBE Proc., p 16* (https://doi.org/10.1007/978-3-540-73044-6_19)
Krasteva V, Christov I, Naydenov S, Stoyanov T and Jekova I 2021 Application of dense neural networks for detection of atrial fibrillation and ranking of augmented ECG feature set *Sensors* **21** (6848) 1–35
Kumari P, Mathew L and Syal P 2017 Increasing trend of wearables and multimodal interface for human activity monitoring: a review *Biosens. Bioelectron.* **90** 298–307
Liu F *et al* 2018 An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection *J. Med. Imaging Health Inform.* **8** 1368–73
Mahmoodabadi S, Ahmadian A and Abolhasani M 2005 ECG feature extraction using daubechies wavelets *Proc. 5th IASTED Int. Conf. on Visualization, Imaging and Image Processing* pp 343–8
Mahmoud S S, Hussain Z M, Cosic I and Fang Q 2006 Time-frequency analysis of normal and abnormal biological signals *Biomed. Signal Process. Control* **1** 33–43
Martínez J P, Almeida R, Olmos S, Rocha A P and Laguna P 2004 A wavelet-based ecg delineator: evaluation on standard databases *IEEE Trans. Biomed. Eng.* **51** 570–81
Mietus J E, Peng C-K, Henry I, Goldsmith R L and Goldberger A L 2002 The pNNx files: re-examining a widely used heart rate variability measure *Heart* **88** 378–80
Minami K, Nakajima H and Toyoshima T 1999 Real-time discrimination of ventricular tachyarrhythmia with fourier-transform neural network *IEEE Trans. Biomed. Eng.* **46** 179–85
Pan J and Tompkins W J 1985 A real-time QRS detection algorithm *IEEE Trans. Biomed. Eng. BME* **32** 230–6
Perez Alday E A *et al* 2020 Classification of 12-lead ECGs: the physionet/computing in cardiology challenge 2020 *Physiol. Meas.* **41** 1-11
Rouhi R, Clausel M, Oster J and Lauer F 2021 An interpretable hand-crafted feature-based model for atrial fibrillation detection *Front. Physiol.* **12** 1–15
Reyna M A *et al* 2021 Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021 *Computing in Cardiololgy 2021* 48, 1–4
Sagie A, Larson M G, Goldberg R J, Bengtson J R and Levy D 1992 An improved method for adjusting the QT interval for heart rate (the framingham heart study) *Am. J. Cardiol.* **70** 797–801
Said S A, Bloo R, Nooijer R and Slootweg A 2015 Cardiac and non-cardiac causes of T-wave inversion in the precordial leads in adult subjects: a dutch case series and review of the literature *World J. Cardiol.* **7** 86–100
Sodmann P, Vollmer M, NathNand and Kaderali L 2018 A convolutional neural network for ECG annotation as the basis for classification of cardiac rhythms *Physiol. Meas.* **39** 104005
Tihonenko V, Khaustov A, Ivanov S, Rivin A and Yakushenko E 2008 St Petersburg INCART 12-lead arrhythmia database *PhysioBank PhysioToolkit PhysioNet* (https://doi.org/10.13026/C2V88N)
Vafaie M, Ataei M and Koofigar H 2014 Heart diseases prediction based on ECG signals' classification using a genetic-fuzzy system and dynamical model of ECG signals *Biomed. Signal Process. Control* **14** 291–6
Virani S S *et al* 2021 Heart disease and stroke statistics—2021 update: a report from the american heart association *Circulation* **143** 254-743
Wagner P, Strodthoff N, Bousseljot R D, Kreiseler D, Lunze F I, Samek W and Schaeffter T 2020 PTB-XL, a large publicly available electrocardiography dataset *Sci. Data* **7** 1–15
Wickramasinghe N L and Athif M 2021 Multi-label cardiac abnormality classification from electrocardiogram using deep convolutional Neural Networks *Computing in Cardiololgy 2021* 48, 2021
Xiaoyu L, Chen L, Xian X, Yuhua W, Jishang W, Yuyao S, Buyue Q and Xiao X 2021 Towards generalization of cardiac abnormality classification using ECG signal *Computing in Cardiology 2021* 48, 2021
Xiong Z, Nash M P, Cheng E, Fedorov V V, Stiles M K and Zhao J 2018 Ecg signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network *Physiol. Meas.* **39** 094006
Yang S and Shen H 2013 Heartbeat classification using discrete wavelet transform and kernel principal component analysis *IEEE 2013 Tencon - Spring* **2013** 34–8
Yang X, Zhang X, Yang M and Zhang L 2021 12-Lead ECG arrhythmia classification using cascaded convolutional neural network and expert feature *J. Electrocardiol.* **67** 56–62
Zhang D, Yang S, Yuan X and Zhang P 2021 Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram *Iscience* **24** 102373
Zhao Z *et al* 2022 Analysis of an adaptive lead weighted ResNet for multiclass classification of 12-lead ECGs *Physiol. Meas.* **43** 034001
Zhaowei Z *et al* 2021 Identification of 27 abnormalities from multi-lead ECG signals: an ensembled SE_ResNet framework with sign loss function *Physiol. Meas.* **42** 065008
Zheng J *et al* 2020a Optimal multi-stage arrhythmia classification approach *Sci. Rep.* **10** 1-17
Zheng J, Zhang J, Danioko S, Yao H, Guo H and Rakovski C 2020b A 12-lead electrocardiogram database for arrhythmia research covering more than 10 000 patients *Sci. Data* **7** 1–8