Multiple Cardiac Disease Detection from distinct ECG Leads Sets Using a Hybrid Supervised and Unsupervised Machine Learning Approach

Author Name[1], Author Name[2] and Author Name[1,2]

[1] Department One, Institution One, City One, Country One
[2] Department Two, Institution Two, City Two, Country Two

E-mail: xxx@xxx.xx

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Standard 12-lead ECG is the primary technique in cardiac diagnostic. However, detecting different cardiac diseases using single or reduced number of leads is still challenging. The purpose of this work is to provide and validated a method able to classify ECG records using multiple or single ECG-lead combinations among 26 different cardiac conditions in the context of the 2021 PhysioNet/Computing in Cardiology Challenge.
Methods: We resampled and filtered the ECG signals, and extracted 81 features mostly based on Hearth Rhythm Variability (HRV), QRST-waves patterns and spectral domain. Then, we trained a One-vs-Rest classification approach made of 26 different binary classifiers. To do so, we selected for each class a set of features and then train and select a model among two binary Supervised Classifiers and one Hybrid Unsupervised-Supervised classification system, thus, allowing each ECG to be classified as belonging to none or more than one class. Finally, we performed a 3-fold cross validation to assess the system performance.
Results: Our classifiers received scores of ??, ??, ??, ??, and ?? (ranked ??th, ??th, ??th, ??th, and ??th out of 39 teams) for the 12, 6, 4, 3 and 2-lead versions of the hidden test set with the Challenge evaluation metric. Also, we obtained a G value ($G = \sqrt{Sensitivity * Specificity}$) of 0.80, 0.78, 0.79, 0.79, 0.77 and 0.74 for the 12, 6, 4, 3, 2 and 1-lead versions of the public training set.
Discussion: We proposed and validated a flexible and scalable machine learning approach to detect a large set of cardiac diseases using multiple or single ECG leads combinations. 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. Accuracy in detection can be improved adding more disease-specific features or rules to our system.
<Not more than 300 words>

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

0. Preprint considerations

Dear Physionet Challenge organizers,

Our entry code for the Special Issue is ready to be evaluated and it is available at https://github.com/sjimenezupv/itaca_upv.cinc2021.special_issue

The authors agree to submit our paper to the focus collection in Physiological Measurement.

The changes that we made with respect to the previous congress submission are as follows:

* After signal filtering stage, we used only 15 seconds of ECG record, as other participants did, improving the processing time and avoiding some unwanted collateral effects of long records.
* We no longer use features for long ecg records (stats over the long signal with sliding window, among others). In total we removed more than the previous 45 features.
* We added more specific qrs and t pattern features that improved our classification (still to be detailed in the correct format in this paper).
* We changed some threshold in the feature selection that improved the validation.
* We created a mixed approach, where each binary classifier will be no longer only a FFNN, but could be made of a FFNN, Naïve Bayes (NB), or Hybrid approach.
* The hybrid classification approach is based also in the unsupervised k-means algorithm, (k=3), where 3 cluster centers are looked for, and a FFNN or NB are associated to each of them.
* A model selection system for this binary classifiers was created for the training process
* Data preprocessing was also added in order to use only 26 classes, and avoid using samples that didn't belong to any of them for training.
* Fixed some bugs and imprecissions in the code
* Improved all the classification scores with respect the previous submission.

These are the big changes that we made (maybe I forgot some). We hope it will be enough to allow us to test our code in the online test set.

In the current version of the paper (preprint), you can see all the results, and methods are almost finished (all the data/tables/figures are currently in their final data version). However, some part of the text will be changed in order to do not coincide with the CinC text-version.

We know that the final scores are not the highest of the challenge, but our methodology was tested and described carefully. We also will thank any comment regarding the attached (unfinished) document content.

Sincerely yours,
Santiago Jiménez

1. Introduction

The clinical importance of cardiac arrhythmias is increasing along with their incidence and prevalence mostly associated with population aging [1]. Besides this, nowadays wearable devices are gaining great interest as monitoring devices in both research and clinical settings [2]. Although standard 12-lead ECG is the primary technique in cardiac diagnostic, detecting different cardiac diseases using single or reduced number of leads is still challenging [3].

The aim of this study is to is to provide and evaluate methods able to classify ECG records using minimal lead information in the context of the 2021 PhysioNet/Computing in Cardiology Challenge [4, 5], using also only a single-lead.

2. Materials

As database for this study we used the 88,253 12-lead ECG registers provided by the competition as training set containing also the age and gender of the patient for each record. Deeper explanation of the database can be found in [4, 5].

Database	#ECG Recordings	#ECG categories
CPSC	6,877	9
CPSC2	3,453	72
INCART	74	37
PTB	516	17
PTB-XL	21,837	50
Georgia	10,344	67
Chapman-Shaoxing	10,247	
Ningbo	34,905	
Total	88,253	

Table 1. Basic information of the seven different databases that forms the training dataset.

ECG category	Abbreviation	#samples 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+1281
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	28971
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	18918
Sinus Tachycardia	STach	9657
T-wave Abnormal	TAb	11716
T-wave Inversion	TInv	3989

Table 2. Cardiac conditions to be detected, abbreviation used in this work and number of samples labelled with each in the public training dataset. Note than one ECG-record could be labelled with more than one of this classes. Some classes contains

more than one cardiac disease since share the same penalty its misclassification values dit not penalize during the Challenge Scoring.

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 2.

We first introduce the signal preprocessing process used to clean the ECG records of undesirable noise and artifacts. Then we present the feature extraction applied to the previous cleaned signals that give us as a result the input dataset to be used during the models training process. Next, we describe the preprocessing of the dataset in order to remove outliers and missing values and the scoring rules for the models that will be trained. Then, we also describe the feature selection process used in order to reduce the input dataset dimensionality and thus decreasing the machine learning models training time and improving the accuracy if a right subset is chosen. Finally we present our hybrid supervised and unsupervised binary classification models setups and the One-vs-Rest classification approach used in order to provide a multi-class classification system able to detect new ECG records among the distinct 26 classes mentioned previously. The binary models selection and validation processes during the whole system training is also described.

All these steps were performed using MATLAB (R2021a, The MathWorks) and all the code used both to train and validate the models can be found in *https://github.com/sjimenezupv/itaca_upv.cinc2021.special_issue* .

In addition to the official leads sets of the *2021 PhysioNet/Computing in Cardiology Challenge* (12, 6, 4, 3 and 2-leads), we also report this methodology results using only the single lead 'I'.

#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

Table 1. Number and sets of ECG leads evaluated in this work.

3.1 Labelling Preprocessing

<hablar de los cambios de etiquetas empleados> First, all ECG signals were resampled to 500 Hz if necessary in order to homogenize the sampling frequency to the most common value in the databases used.

3.1 Signal Preprocessing

First, all ECG signals were resampled to 500 Hz if necessary in order to homogenize the sampling frequency to the most common value in the databases used.

Next, a 50 Hz second-order IIR notch filter plus a Butterworth band-pass filter between 0.5 Hz and 40 Hz were applied, thus removing baseline wander artifacts and high frequency noise such as powerline interference. Then, we removed the first and last second of each signal in order to leave out the filtering stabilization stage.

Lastly we removed the signal aberrant artifacts. To do so, we used a 0.5 seconds sliding window in order to calculate anomalous maximum and minimum values, and sections surrounded by outliers were set to zero in the corresponding ECG lead. To consider that a signal value is aberrant, we used the next upper and lower thresholds

$$upper_{th} = median(max_{values}) + std(max_{values}) * tol_1$$

$$lower_{th} = median(min_{values}) - std(min_{values}) * tol_2$$

where max_{values} and min_{values} contains the minimun and maximun values for each 0.5 seconds signal window, plus tol_1 and tol_2 takes a value of 5 if the dominant wave in the QRS complex is R in the case of the upper limit, or S in the case of the lower limit, and a value of 4 otherwise.

Finally, in order to avoid big differences in the lasting time among the ECG records, in the next stages of this work we only used the first 15 seconds of signal (if available), removing the remaining signal seconds from the ECG registers as was previously done in [REF REF].

3.2 Feature Extraction

We automatically extracted 81 signal features from each ECG-lead, mostly derived from ventricular activity, most of them previously used in [6, 7]. To carry out this task, initially, we extracted the RR sequence using a QRS detector based on the first derivative of the ECG. Then we filtered the outliers from the RR sequence, and obtained the first and second derivatives of that sequence (RRd1, RRd2). Also, we created a T-wave detector in order to obtain the QT interval and other related features.

Finally, we got both the QRS and T wave patterns for each lead using a ±100ms window over all the QRS and T wave detections.

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

Using the above information, the extracted signal features for each lead can be grouped as follows:

Group 1. Basic statistics over the R and T waves voltages (mean, standard deviation). 4 features.

Group 2. Basic statistics over the QT interval in milliseconds (mean, standard deviation). 2 features.

Group 3. Features based on the QRS and T patterns: Percentage of amplitude of T wave respect the R wave, 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. 11 features.

Group 4. Spectral features: Dominant frequency (fdom) using the Welch spectral density estimation method, percentage of the area in fdom±0.5Hz in the periodogram normalized in the range [0, 1] and the sum of the normalized periodogram in steps of 2Hz in the range [0, 30] Hz. 17 features.

Group 5. Basic statistics over the RR, RRd1 and RRd2 sequences (mean, standard deviation, kurtosis, skewness). 9 features.

Group 6. Features based on RRd1: RMSSD, pNN25, pNN50, pNN75, where pNNxx [8] denotes the percentage of intervals between normal beats exceeding xx ms. 4 features.

Group 7. Poincaré plot-based features using 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. 7 features.

Group 8. Lorenz plot-based features using RRd2: Angular variability, dispersion of the distance between points to origin, and differences between 2 and 3 consecutive beats. 8 features.

Group 10. Other features: Shannon entropy of the RR sequence, Lempel-Ziv complexity of the RR time series after binarization using the median as threshold, and ratio between the number of different QRS patterns found and the total number of waves detected. 3 features.

3.3 Feature Dataset Preprocessing

First, for each feature, outliers exceeding 3 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 for a given sample, we replaced that value for the median value in the dataset for such feature. According to this rule, and taking into account 974 features in the whole dataset using 12 leads (81 features for each lead plus age and sex of the patient), finally the 0.61% of values were replaced for the corresponding median.

Lastly, we performed a z-score using the training set to rescale the whole dataset.

<indicar que aquellas muestras que no tienen ninguna muestra correspondiente a nuestras clases son eliminadas>

<indicar que aquellas clases con menos de 150 muestras positivas también serían eliminadas del análisis, aunque en principio no ha hecho falta???>

3.4 Scoring

2021 PhysioNet/Computing in Cardiology Challenge scoring rules are described in [5], where only 26 classes are taken into account. Also, we report the G metric (G=sqrt(Sensitivity*Specificity)) in this work since it was used in order to select the binary classifiers with best performance during the training and validation stage.

3.5 Feature Selection

We performed a feature selection for each of the distinct 26 ECG categories in this work, previously to the training of each binary classifier mentioned below, using both supervised and unsupervised statistical filtering methods. Furthermore, we also applied the same feature selection methodology to each of the data clusters used to train the hybrid classification models that will be further presented in the next section.

Age and sex always were used in order to avoid an empty set of features. Next, we perform 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.

Finally, we get the correlation coefficient among the lasting features for each pair of features, and we removed the last feature of the pair where their correlation coefficient was greater or equal than 0.95.

The remaining features were used as inputs for the corresponding binary classifier. Thus, each further binary classifier presented had associated a subset of features specifically selected for its own training and validation.

3.6 Binary classification strategy. Hybrid Supervised and Unsupervised Classification Model

In this work, for each cardiac contition to be detected (Table xx), a binary classifier is selected to give the corresponding model response, as shown in Figure 1. This approach is also known as a One-vs-Rest classification model, where the samples of the dataset are labeled as positive or negative during the training of each binary classifier depending on if they belong or not to such class. Finally, the whole One-vs-Rest classification model will give a binary response indicating if an unseen sample belongs or not to each

of the classes previously used during the training process. Thus, each binary model solves an independent classification problem in the whole classification system, been possible to assign a new sample to none or more than one class. This is specially usefull in this work since some samples could not belong to any cardiac category previously seen, or, on the contrary, belong to more than one cardiac condition.

Decir que en el híbrido, para cada cluster, si no hay suficientes muestras, no se entrena.

Mencionar también qué sucede en los mismos casos en caso de las redes neuronales,

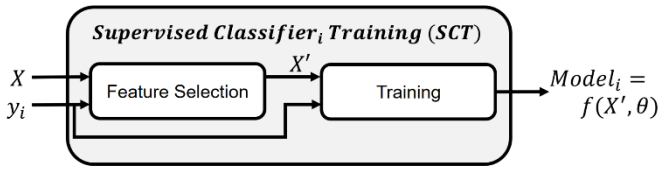

Figure 1. Summary diagram for the Supervised Classifier Training (SCT) process used in this work. X denotes the input dataset whereas y_i the real binary response. As a result, a binary classification system f with parameters θ that takes as inputs the selected features X' is given.

Furthermore, previously to the training of each binary classifier, we performed the feature selection described previously, in order to use as inputs an estimate of the features that best fits their own binary classification problem.

Finally, we create a system for training, validation and selection of the best binary classifiers among three different types of machine learning models for each class. The first machine learning model is a Naïve Bayes (NB) classifier, the second one is Feed Forward Neural Networks (FFNN), and the third one is an Hybryd classifier based on unsupervised and supervised machine learning methods that will be described further.

This binary classifier can we tested we evaluate three different approaches.

First,

Lastly, as unsupervised classification approach

The Hybrid classification approach is based on a k-means algorithm (k=3). Thus, for classifying a new unseen sample, we first get the euclidean distance from this sample to each of the three cluster centers; then, we choose the the cluster than minimize this distance, and use the supervised classification model associated to such cluster to finally give the binary response, i.e. use the Naïve Bayes or FFNN trained previously only with the data belonging to this cluster.

$$g(hybrid) = \sum_{i=1}^{3} g_i * \left(\frac{n_i}{n}\right)$$

Where g_i denotes the g value for the classifier corresponding to the cluster I, n_i the number of samples belonging to the cluster i, and n the total number of samples used to train the binary model.

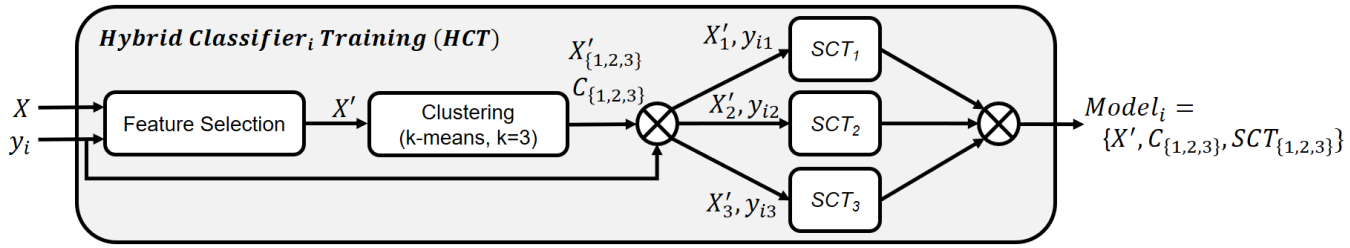

Figure 2. Summary diagram for the Hybrid Classifier Training (HCT) process for a binary classifier used in this work.

3.6 One-vs-Rest Classification Approach

Each binary classifier uses its selected set of features as inputs that best fits its own classification problem.

'47665007' (right axis deviation) => siempre se usa nbayes

Next, each binary classifier corresponds to a FFNN made of 18 or 32 hidden units, and a threshold for the output to give the binary response. All the FFNN where trained with the default objects and parameters in the Matlab R2020b Deep Learning Toolbox, using the *trainscg* (Scaled Conjugate Gradient) as training function, the *useGPU* flag switched on in order to use the available GPUs to speed up the training and the *showResources* flag switched off.

Using as inputs the selected features for a given class, the 75% of training data was used to train the FFNN and the resting 25% to select the output threshold in the range [-1, 1] that achieves a higher G value, both with 18 and 32 hidden

units. Finally, among the two trained models, we choose the one that presented a higher G value to be used in the whole One-vs-Rest classification model.

3.7 Model Validation

Since the number of samples in the database is large enough and the training time could become unnecessarily high for a cross-validation with a large number of folds, we used a 3-fold cross validation with the 88,253 training samples. Furthermore, we selected the samples for each fold with no bias among all the distinct databases available.

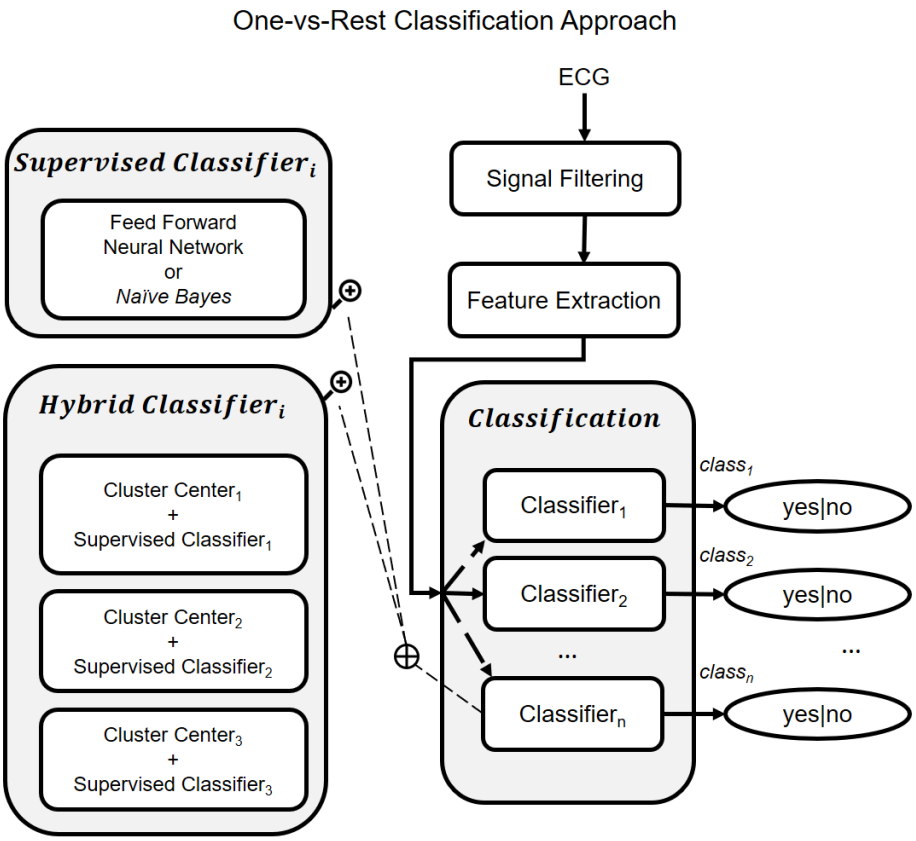

Figure 3. Summary diagram of the One-vs-Rest classification approach used during this work. The classification for a given unseen ECG record needs of three main stages: a signal filtering, a feature extraction process and a labelling system that uses a binary classifier for each of classes to be detected. Each of the binary classifiers correspond to a Supervised or Hybrid Classifier, where in the first case contains only a FFNN or NB model, and in the second case contains 3 cluster centers associated to 3 Supervised models. For an Hybryd model, a new sample is labelled using the Supervised Classifier corresponding to its nearest cluster center.

4. Results

This section presents the detailed results achieved in this work. First, we present the computational costs both for training and testing the classificacion models. Second, we perform an analysis of the feature selection results. Third, we present an analysis of the selected models during the training and validation process. Finally, we detail the score metrics achieved during the testing process of each of the 6 One-vs-Rest multi-classifier models and also we report detailed metrics for each binary classifier giving an overview of the results for each cardiac disease using the different combination of leads.

4.1 Computational costs

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

The time needed to perform the signal processing and feature extraction for the 12-lead ECG records was of 590 seconds (9.83 minutes). This gave us a performance of 99.72 samples per second for signal feature extraction using the the whole leads set employing the hardware described previously.

Next, we measured the time needed to perform the feature extraction, training, validation and selection of each of the 6 One-vs-Rest multi-classifiers. 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, been the training process the more time consuming whereas the testing process presents a linear behaviour 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 number of samples processed per second. Best performance corresponds to the single-lead classifier, with 21.4 samples/second during training and 17.4 samples/second during testing. As expected, lower performance results take place using the 12-leads ECG classifier, with 4.8 samples/second for training and 7.3 samples/second 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 disk, adding an overhead that does not exist during the training process.

4.2 Feature Selection Analysis

First, in Table tt is shown the percentage of features selected from the whole dataset in order to train each binary classifier that finally compose the 6 One-Vs-Rest multi-classifier models. The method employed selected a mean of 60±10% of available features (removing the remaining for the corresponding binary classifier) with no significant differences among the leads combination used.

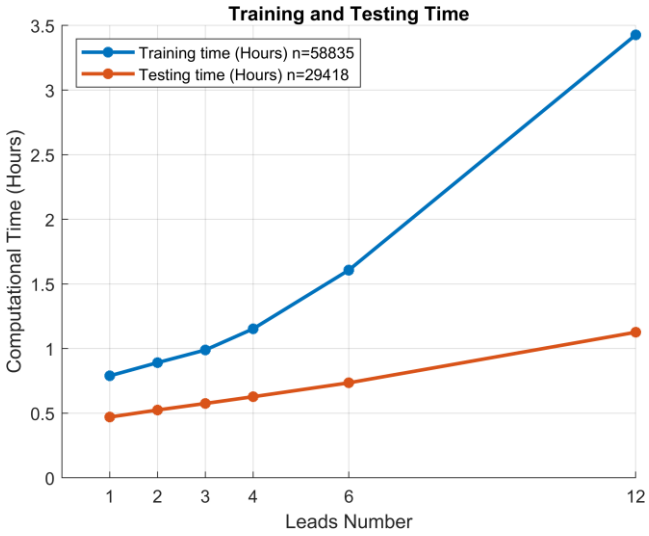

Figure 4. Trend of training and test time for each of the ECG multi-classifier models in one fold of our experimentation.

Figure 5. Performance of the ECG multi-classifer models during the training and test of one fold of our experimentation, in terms of number of samples processed per second.

#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 tt. Mean±Standard Deviation of the percentage of selected features for each binary classifier during the 3-fold cross validation on the public training set.

Second and last, we obtained for each binary classifier the percentage of selected features corresponding to each of the leads used for training the model. Training the models with 12-leads only V1 features presented a slightly higger selection higger selection with a value of 8.81±0.76%. In the case of the models trained with 4 and 3-leads, V2 presented lower selection percentage with values of 24.15±2.28 32.01±2.72 each. In the other leads combinations, the selection percentage among features corresponding to distinct leads does not presented significant differences, been this ratio proportional to the number of leads employed.

% 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 pp. Mean±Standard Deviation of the percentage of features selected for a given binary classifier corresponding to each of the leads used during the 3-fold cross validation on the public training set.

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 uu. More than the 50% of the times, a FFNN model were selected whereas in a range between 28% and 38% of the times our proposed Hybrid classifier improved the performance during the validation process, and thus, were selected to finally perform the binary classification for a given category. Finally, NB was selected less than the 11 % of the times, mainly where the other models did not performed well.

#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%

Table uu. 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.

Moreover, we detail the percentage of the distinct type of models selected for each binary classifier in Table vv. For ten cardiac conditions, only FFNN are selected, whereas in the other categories a mix o models where 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 in the hidden test set using the Challenge score have a value of <???> using 12, 6 and 2 leads indistinctly. Table 1 shows the whole results set and ranking using the Challenge score.

Table 2 shows the mean of different performance metrics in the classification of the 26 scored classes in the challenge on the public training set, where higher G value of 0.76 was achieved using 12 leads, followed by a G value of 0.74 using both 6 and 2 leads.

Finally, Table 3 shows the results achieved for individual binary classifiers where G metric is greater than 0.8 in some of the lead combinations during the validation of the first training fold, where 10 different cardiac conditions reach this classification performance threshold.

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	-
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 vv. Percentage of the distinct type of models selected for each binary classifiers during the 3-fold cross validation on the public training set.

#Leads	Training	Validation	Test	Ranking
12	0.435±0.009	??	??	??
6	0.402±0.003	??	??	??
4	0.421±0.001	??	??	??
3	0.420±0.004	??	??	??
2	0.414±0.005	??	??	??
1	0.388±0.005	-	-	-

Table 1. Challenge scores for our final selected entry (team itaca-UPV) using 3-fold cross validation on the public training set, one time scoring on the hidden validation set, and one time scoring on the hidden test set as well as the ranking on the hidden test set.

#Leads	AUROC	F-measure	Sens.	Spec.	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 2. Mean of other performance metrics among the classification of the 26 scored classes for our final selected entry using 3-fold cross validation on the public training set: Area Under the ROC Curve, Sensitivity, Specificity and G metric; mean of the standard deviation was ±0.007.

Class Score: G-Score 12-leads 6-leads 4-leads 3-leads 2-leads 1-leads

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

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

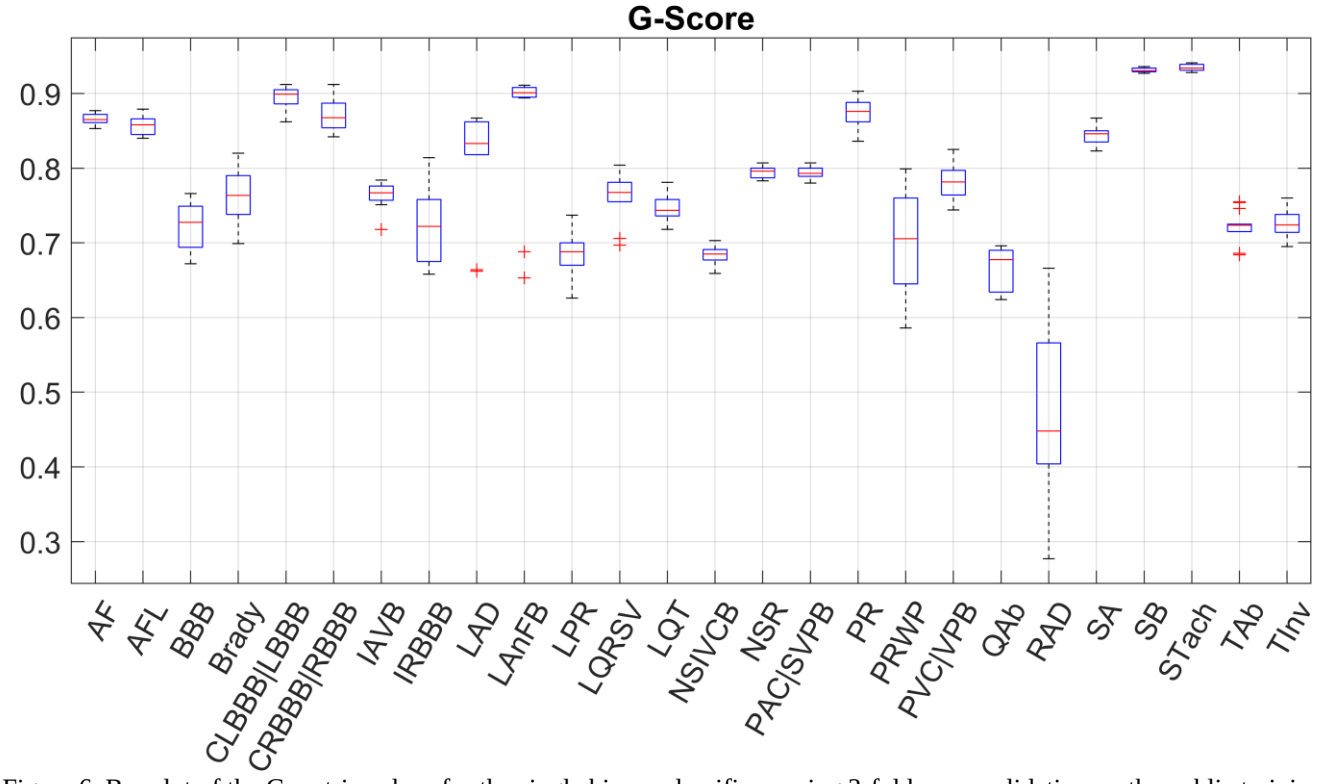

Figure 6. Boxplot of the G metric values for the single binary classifiers, using 3-fold cross validation on the public training set.

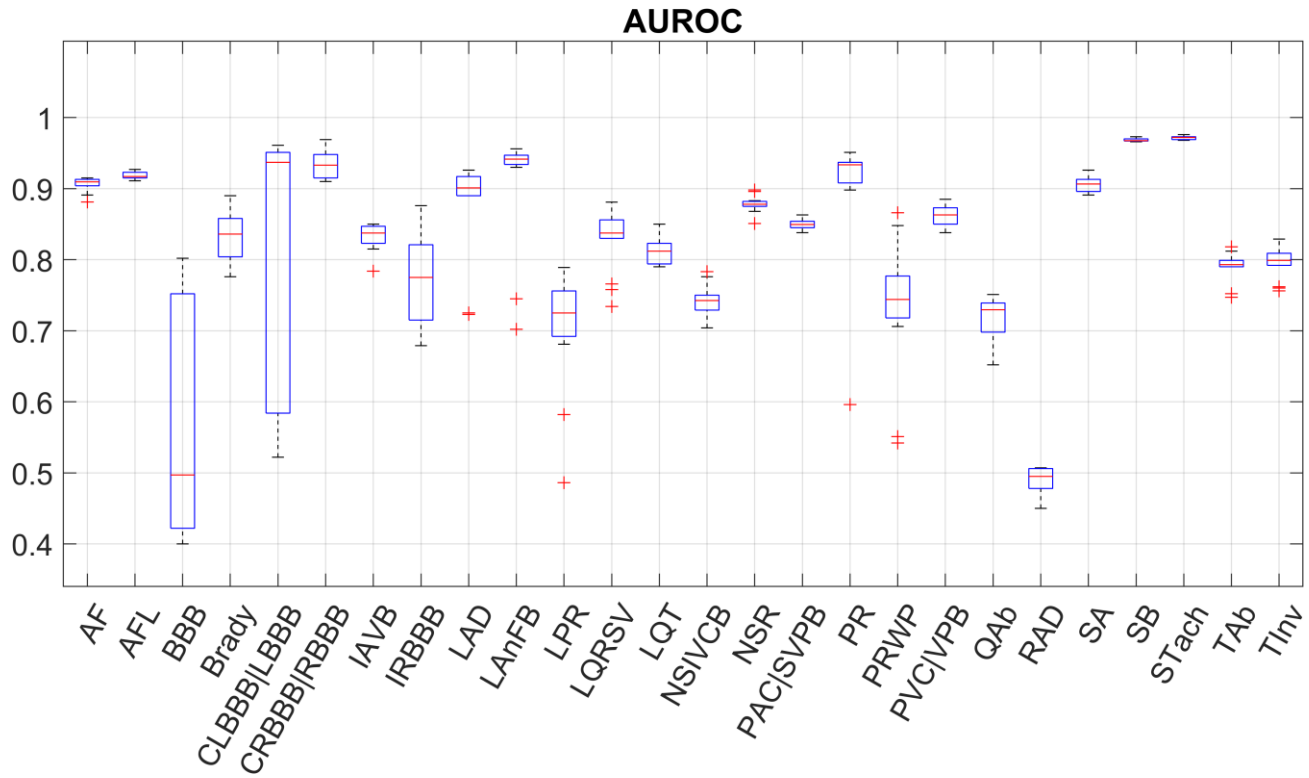

Figure 7. Boxplot of the Area Under the ROC curve (AUROC) values for the single binary classifiers, using 3-fold cross validation on the public training set.

5. Discussion

Results obtained in this work showed low differences among the results obtained in the G values among the classifiers that uses 12 leads and the ones that only uses minimal leads information. Classification using only one or two leads outperformed the ones using three and four leads, possibly due to leads V2 and III added some bias in their extracted features respect the ones in leads I and II. In this sense, further studies should get the classification performance for each single lead in order to know their individual performance.

On the other hand, since the signals from leads I and II share many characteristics with those offered by wearables devices, the resulting classification models may eventually be suitable for clinical use in wearable or automated control systems. This approach would benefit from low computational costs consumption during classification.

Nevertheless, poor results in the Challenge score metric could highlight that this approach should be used cautiously when detecting cardiac conditions with low performance in our results. Future optimization of these classifiers should improve by adding more disease-specific features and/or modifying the binary classification strategy.

<decir que a nuestro sistema es fácil añadir/quitar clases, características y modelos de machine learning, supervisados, no supervisados, híbridos e incluso de deep learning>

Se demuestra que existen clases que necesitan más de un lead para poder ser clasificadas correctamente, ver boxplot

6. Conclusion

We presented and evaluated a robust methodology for multiple cardiac disease detection through ECG registers that combines feature extraction and selection, and a One-vs-Rest classification approach using FFNN as binary classifiers. Interestingly, the classification results using only one or two leads outperformed the ones using three and four leads, and almost matched the ones with twelve leads, showing lower computational costs and been more suitable for wearable monitoring devices. Improving the identification of some cardiac rhythms by incorporating more specific features for those cases where the performance was low, should be an interesting direction to explore in the future.

References

[1] Chow GV, Marine JE, Fleg JL 2012. *Epidemiology of Arrhythmias and Conduction Disorders in Older Adults. Clinics in Geriatric Medicine, 28(4), 539–553*

[2] Kumari P, Mathew L, Syal P 2017. *Increasing Trend of Wearables and Multimodal Interface for Human Activity Monitoring: A review. Biosensors and Bioelectronics*, 90: 298-307.

[3] Kligfield P 2002. *The Centennial of the Einthoven Electrocardiogram. Journal of Electrocardiology*;35(4):123–129.

[4] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, et al. 2020. *Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiological Measurement*; 41.

[5] Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah A, Robichaux C, et al. 2021. *Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021*; 48:1–4.

[6] S Jiménez-Serrano, J Yagüe Mayans, E Simarro-Mondéjar, CJ Calvo, F Castells, J Millet 2017. *Atrial Fibrillation Detection Using Feedforward Neural Networks and Automatically Extracted Signal Features. Computing in Cardiology 2017*; 44:131–134.

[7] S Jiménez-Serrano, J Rodrigo, CJ Calvo, F Castells, J Millet 2021. *Multiple Cardiac Disease Detection from Minimal-Lead ECG Combining Feedforward Neural Networks with a One-vs-Rest Approach. Computing in Cardiology 2021*.

[8] Mietus JE, Peng C-K, Henry I, Goldsmith RL, Goldberger AL. 2002. *The pNNx Files: Re-examining a Widely Used Heart Rate Variability Measure. Heart*, 88: 378-380.