Supervised Classification of Brugada Syndrome Patients by ECG-derived Markers

Alba Isabel-Roquero[1], Pedro Gomis[2], Luis Tortosa[2], Alvaro Leva[3], Flavio Palmieri[2,3], Elena Arbelo[1,4,5]

[1] Institut d' Investigació August Pi y Sunyer (IDIBAPS), Barcelona, Spain
[2] ESAII Dept, EEBE, CREB, Universitat Politècnica de Catalunya, Barcelona, Spain
[3] Fundació de Recerca Sant Joan de Déu, Esplugues-Barcelona, Spain
[4] Arrhythmias section, Hospital Clínic, Barcelona,Spain
[5] European Reference Network for rare, low prevalence complex diseases of the heart - ERN GUARD-Heart

Abstract

Brugada syndrome (BrS) has been associated with risk of ventricular fibrillation and sudden cardiac death (SCD). Its risk stratification remains challenging as the only accepted factor is the presence of resuscitated cardiac arrest or arrhythmogenic syncope and the majority of patients are diagnosed in the asymptomatic phase. Moreover, the only treatment available to prevent SCD is the implantation of a cardiac defibrillator, which can lead to adverse events such as inappropriate shocks. In this study, we present Machine Learning (ML)/supervised classification tools for BrS risk stratification based on the automatic analysis of long-term high-resolution electrocardiographic information. For this purpose, 12-lead ECG 24h Holter and clinical variables from 64 Brugada subjects were used. ECG signals were preprocessed with a signal-averaging algorithm to reduce noise and obtain individual ECG beats for delineation, resulting in 11 ECG biomarkers. Subsequently, 4 different ML/supervised algorithms based on Decision Tree, XGBoost, K-Nearest Neighbors and support vector machine algorithms were tested. AUC results were around 90%, however sensitivity results were around 50%. The results do not efficiently predict BrS symptomatic patients at risk of SCD, which is mainly caused by the reduced number of symptomatic patients. Further studies with additional subjects and variables may improve this prognosis.

1. Introduction

Brugada syndrome (BrS) is a rare inherited disease showing a distinctive electrocardiographic (ECG) pattern and is associated with a high incidence of sudden and unexpected arrhythmic accidents. It is believed to cause 4–12% of all sudden cardiac deaths (SCD) and up to 20% among patients without identifiable structural abnormalities [1, 2]. It can be classified within a group of genetic disorders known as channelopathies that affects the genes in charge of the function of the cardiac ionic channels.

The BrS is characterized by a typical electrocardiographic pattern described as an ST elevation followed by either a coved-shaped (type 1) or saddle-shaped (type 2) slope, in the absence of other structural abnormalities [1, 2]. These patterns can be variable and sometimes remain hidden, affecting the diagnosis of this disease.

The risk stratification and management of patients remains challenging as the only accepted risk is the presence of resuscitated cardiac arrest or arrhythmogenic syncope and the vast majority of patients are diagnosed in the pre-clinical (asymptomatic) phase with an annual incidence of SCD of 0.5%-1% [1, 3]. Moreover, the only treatment available to prevent SCD is the implantation of a cardiac defibrillator, which can cause adverse events such as inappropriate shocks (up to 9%) or infections (20%)[4]. In cases where the patient would not experience SCD throughout their life, the use of these devices may worsen their cardiac health, as its implantation has associated risk of morbidity and mortality.

Despite the progress in understanding the mechanisms behind the manifestation of the Brugada phenotype, there are currently no reliable indicators available to accurately determine an individual's risk of SCD. Some researchers have developed predictive algorithm based on the use of 10-second ECG signal from control visit [5]. However, these may be not sufficient to illustrate BrS, as the majority of patients do not show the electrocardiographic pattern in such a reduced period of time.

For this reason, we proposed a supervised classification tool for BrS risk stratification based on the automatic anal-

ysis of long-term high-resolution electrocardiographic information and other clinical variables.

2. Methods

2.1. Materials

The study included 12-lead ECG holter signals of 24 hours of 64 subjects with BrS, including 54 asymptomatic patients and 10 symptomatic patients who had suffer from SCD or syncope. Moreover, clinical data of the patients was added including some variables such as diagnosis age, sex, documented history of SCD in the family, documented history of ventricular fibrillation, spontaneous Type 1 ECG pattern, among others.

2.2. Preprocessing

In order to use this data for the classification, the signals were pre-processed with MATLAB. For this purpose, firstly, the signals were splited in segments of 3 minutes to perform the delineation with the ECG-kit [6]. This step was necessary to extract the ECG fiducial points and obtain all the beats of each segment. After this, the signals were prepared to be filtered for which a signal averaging technique (SAV) was applied, as shown in Fig. 1. This technique permitted to reduce the noise level while preserving the high-frequency component of the signal. This algorithm consisted of aligning all the beats of each 3-minute segment to obtain a signal template by computing the mean of the aligned beats. This template was compared to each individual beat to obtain the similarity between them. The similarity was evaluated performing the cross-correlation and a value of the shift is obtained. The signals with shift values lower than the QRS duration are corrected and the rest are rejected.

After noise reduction in each individual lead, all 1-second SAV beats were delineated to extract the fiducial points, including the peak, onset and offset of P wave, QRS complex and T wave. Then, 10 biomarkers were computed: ST deviation at J-point and J-point + 60 ms, ST slope, PR interval, QT interval, corrected QT interval, average power of QRS (Pavg), absolute value of QRS area (areaQRSabs) and late potentials (LP) variables such as the duration of the (QRSd) and the RMS signal amplitude in the last 40ms (RMS40) of filtered QRS complex.

2.3. Supervised classification methods

After the preprocessing, a statistical analysis of the biomarkers, based on the application of Student's t-test, was made to study the statistical differences between asymptomatic and symptomatic patients. Moreover, the obtained biomarkers and the clinical data of each patient were used to feed several classification models, developed in Python, in order to optimise a model for the prediction of symptomatic and asymptomatic patients. For this purpose, 4 models were tested including Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest (RF) and Extreme Gradient Boosting (XGBoost). In order to feed the model, seven different datasets with the optimal variables obtained with different algorithms such as Pearson correlation, decision Tree approach or ANOVA test are tested to find the best combination of ECG-derived biomarkers and clinical data. These subsets were tested on each of the 4 models.

To evaluate the performance of each model on the 7 subsets, stratified cross validation was applied. After finding the best combination of variables, the dataset was divided into train (75%) and test (25%) and the model was optimized using hyperparameter tuning, more specifically Grid Search. Besides, other techniques were applied in order to optimise the model such as weighted models, which was applied to the SVM and the XGBoost.

3. Results

Concerning the preprocessing, the obtained results showed a considerable reduction of noise after the application of the SAV algorithm. Moreover, these results were compared to previous algorithms applied based in conducting ensemble averaging without shift correction and rejection conditions. In the case of noisy segments, the SAV algorithm showed improved results by better preserving the amplitude of the original signal, as shown in in Fig. 1.

After the evaluation of the different datasets, results showed that the best dataset was dataset 5, obtained with the ANOVA test algorithm, containing 30 variables. Specifically, the 20 most significant biomarkers were selected as they shown the best results on the classification. This dataset included 4 clinical variables such as recovered SCD, family history of SCD before the age of 55 years, the presence of previous Atrial Fibrillation (AF) events and the proband condition. On the other hand, the dataset include 16 ECG biomarkers, being related almost half of the biomarkers to ST elevation. The set of most significant electrocardiographic variables is shown in Table 1.

The statistical results showed that some ECG variables such as the averaged power of the QRS complex or the absolute value of QRS area were significantly higher in symptomatic patients in some precordial leads such as V1 or V2 ($p < 0.001$). Some results are shown in Table 2.

In relation to the classification models, the best results were obtained with the RF and the XGBoost. The confusion matrixes showed sensitivity results (True Positive Rate) of 46.6% and 47.07% and specificity results (True Negative Rate) of 99.85% and 46.6%. The results, displayed in Table 3, obtained after the cross validation ap-

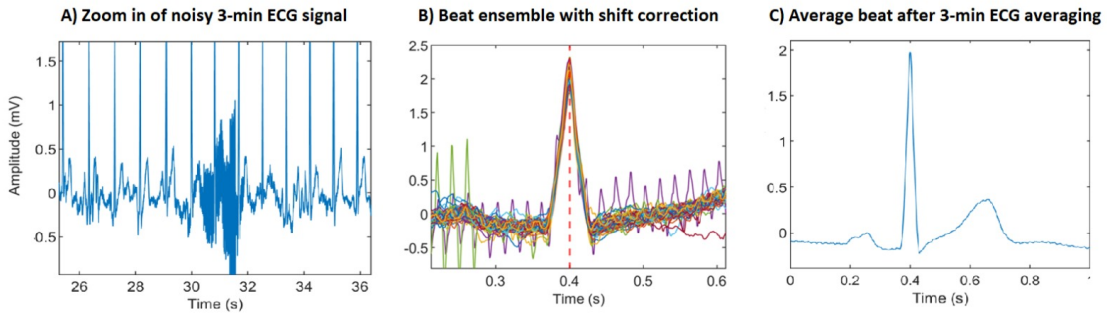

Figure 1. A) Interval of a noisy 3-min ECG segment. B) Beat ensemble without shift correction. C) Averaged beat after 3-min ECG averaging

Table 1. Best subset derived from the 20 most significant variables of ANOVA-test.

Variables	Description
Pavg aVL	Average power of the QRS complex in aVL
ST 0 I	J-point amplitude in I
AreaQRSabs aVL	Absolute area of the QRS complex in aVL
ST 60 V3	J-point +60 ms amplitude in V3
ST slope V5	Slope of the ST-segment in V5
ST 60 aVR	J-point +60 ms amplitude in aVR
ST 0 V5	J-point amplitude in V5
LP QRSd V3	QRS duration from Late Potentials in V3
AreaQRSabs III	Absolute area of the QRS complex in III
QTc V6	Corrected QRS duration in V6
AreaQRSabs V2	Absolute area of the QRS complex in V2
LP QRSd V1	QRS duration from Late Potential in V1
LP QRSd V4	QRS duration from Late Potential in V4
ST slope II	Slope of the ST-segment in II
ST 60 V1	J-point +60 ms amplitude in V1

Table 2. Statistical analysis of the Asymptomatic (n= 22215 beats) and Symptomatic patients (n= 4089 beats).

Variables	Asymptomatic	Symptomatic	p-value
Pavg V2 (mV^2)	0.119 ± 0.126	0.143 ± 0.146	$p < 0.001$
AreaQRSabs V2 ($\mu V \cdot s$)	24.68 ± 13.36	32.78 ± 24.28	$p < 0.001$
ST slope V1 ($\mu V/ms$)	-1.41 ± 1.81	-1.92 ± 1.89	$p < 0.001$
PR V1 (s)	0.185 ± 0.048	0.205 ± 0.065	$p < 0.001$
LP QRSd V1 (ms)	115.2 ± 37.27	135.1 ± 45.2	$p < 0.001$

plied to the RF and the XGBoost showed an AUC of 0.907 and 0.888 and a sensitivity of 54.30% and 58.38%, respectively. The confusion matrix of the RF can be shown in Figure 2.

4. Discussion

It has been evaluated whether a supervised classification model trained with ECG-derived biomarkers and clinical data is able to discriminate between symptomatic and asymptomatic condition with BrS. Regarding the preprocessing of the data, the SAV algorithm developed was more robust than a basic ensemble learning. Through the implementation of shift correction, the amplitude of the original signal was better preserved and the loss of information in the QRS complex was reduced.

The best models were the RF and the XGBoost with an AUC of 0.907 and 0.888 respectively. Additionally, the models shown optimal capabilities to detect asymptomatic patients (specificity >80% in all the cases) but a poor identification of symptomatic patients (sensitivity approximately of 50%), which is mainly due to the imbal-

Table 3. Final performance results for the best tunned version of each model.

Model	Input data	Sensitivity	AUC
KNN	Dataset 5 (ANOVA)	31.60%	0.651
SVM	Dataset 1 (19 ECG-biomarkers)	27.22%	0.708
XGBoost	Subset 20 (ANOVA)	58.38%	0.888
RF	Subset 20 (ANOVA)	54.30%	0.907

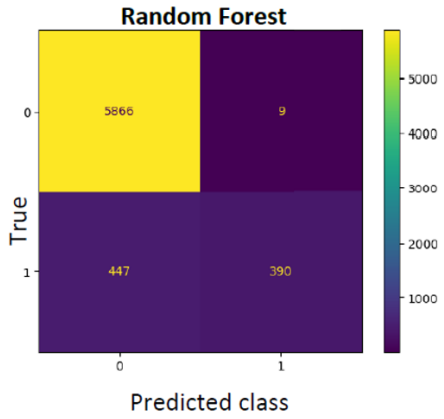

Figure 2. Confusion matrix of the RF model.

ance between classes. Even though the statistical analysis showed differences between asymptomatic and symptomatic patients in some variables, the classification models was not able to properly classify the high risk group.

The optimal dataset was Dataset 5, specifically the 20 most significant variables of this subset. The 4 clinical predictors reinforces some of the findings described in the literature such as the presence of previous AF events and recovered SCD as risk indicators. Moreover, the vast presence of ST variables among the predictors could be related to the BrS pattern, which is characterised by a descending elevation and a steepest slope of ST segment.

As a limitation, the use of averaged signals may mask short-term temporary changes in the ECG, such as the presence of the BrS pattern for a few minutes, and affect significant characteristics for the differentiation of both groups.

5. Conclusion

The main goal of the project was to evaluate the effectiveness of supervised classification models, trained with 12-lead ECG 24h-Holter signals and clinical data, in distinguishing between the symptomatic and asymptomatic conditions for a cohort of BrS subjects. For this purpose, a preprocessing step based on SAV techniques was applied obtaining successful results in reducing the noise and preserving the amplitude. However, the classification models obtained poor results in distinguish the symptomatic and the asymptomatic groups. Only the asymptomatic class was correctly classified, which was mainly due to the imbalance of data. The potential of predictive models should not be dismissed. Indeed, the clinical consistency of the variables giving the best performance encourage to further evaluate these models with large and balanced datasets. Further studies are being conducted, increasing the number of subjects with symptoms and preserving the electrocardiographic characteristics by using beat-to-beat data.

References

[1] Brugada J, *et al.* Present status of brugada syndrome. Journal of the American College of Cardiology 8 2018;72:1046–1059. ISSN 07351097.

[2] Brugada J, *et al.* Right bundle-branch block and st-segment elevation in leads v1 through v3. Circulation 2 1998;97:457–460. ISSN 0009-7322.

[3] Hernandez-Ojeda J, *et al.* The role of clinical assessment and electrophysiology study in brugada syndrome patients with syncope. American Heart Journal 2 2020;220:213–223. ISSN 00028703.

[4] Hernandez-Ojeda J, *et al.* Patients with brugada syndrome and implanted cardioverter-defibrillators. Journal of the American College of Cardiology 10 2017;70:1991–2002. ISSN 07351097.

[5] Liu CM, *et al.* A deep learning–enabled electrocardiogram model for the identification of a rare inherited arrhythmia: Brugada syndrome. Canadian Journal of Cardiology 2 2022; 38:152–159. ISSN 0828282X.

[6] Demski AJ, *et al.* ecg-kit: a matlab toolbox for cardiovascular signal processing. Journal of Open Research Software 4 2016;4:8. ISSN 2049-9647.

Acknowledgments

This work was supported by Fundació La Marató de TV3 (Projecte 245/U/2020), Barcelona, Spain.

Address for correspondence:

Pedro Gomis
Universitat Politècnica de Catalunya, Barcelona 08028 Spain
pedro.gomis@upc.edu