Analysis of the window size effect for T-Wave Alternans detection through Machine Learning methods

Lidia Pascual-Sánchez[1], Rebeca Goya-Esteban[2], Fernando Cruz–Roldán [1], Antonio Hernández–Madrid [3], Manuel Blanco-Velasco[1]

[1] Universidad de Alcalá, Madrid, Spain
[2] Universidad Rey Juan Carlos, Madrid, Spain
[3] Hospital Ramón y Cajal, Madrid, Spain

Abstract

T-wave alternans (TWA) is a phenomenon observed in the electrocardiogram (ECG), characterized by a consistent fluctuation of the ventricular repolarization segment. These episodes are regarded as a marker of high risk of ventricular vulnerability and sudden cardiac death. Numerous analysis strategies have been introduced to detect TWA in the ECG. However, detection of TWA on ambulatory recordings remains an open issue, so this work addresses the problem using a set of machine learning (ML) methods. Decision Trees (DT), Random Forest (RF) and K-Nearest-Neighbors (KNN) are fed with features extracted from three representative TWA analysis methods, namely the Spectral Method, the Modified Moving Average and the Time Method. Since ambulatory ECG exhibits high variability, this work investigates the impact that the analysis window size has in the performance, so, short-term frames of heartbeats (hb) with different sizes are considered. An ensemble dataset of 750 instances made of real ECG signals with added TWA episodes was utilized. Longer windows showed better performance in terms of F1-score. A non-parametric statistical test demonstrated significant performance increase for windows of 40 hb and 30 hb compared to 20 hb, but not for windows of 40 hb compared to 30 hb, possibly indicating an upper limit for the window length. All three ML models yield comparable scores and can learn from signal excerpts of varying lengths to identify alternant waves of 35 μV.

1. Introduction

T-wave alternans (TWA) refers to a beat-to-beat disparity in the amplitude, duration, or waveform of the ST-T complex in the electrocardiogram (ECG). Several studies have linked the presence of TWA to elevated cardiac risk [1, 2], hence, it has been suggested as a potential marker of sudden cardiac death risk [3, 4].

A variety of methods have been proposed for TWA detection and estimation [5, 6]. The Spectral Method [1], the Complex Demodulation method [7], the Modified Moving Average [8] are among the most common methods. Other alternative methods have been devised, employing different signal processing techniques, such as the Laplacian Likelihood Ratio [5, 9]; Matched filter [10]; the Wavelet Transform [11] and adaptive time–frequency analysis [12], among many others. Moreover, various signal processing techniques are employed to preprocess the signal before TWA detection, such as Principal Component Analysis [13], Empirical Mode Decomposition [14, 15], and Bootstrap resampling [16]. In spite of the number of proposals, the validation and confrontation of the algorithms are troublesome due to the lack of definition of a clinical gold standard. Due to the absence of annotated databases, testing frameworks often resort to using synthetic signals, usually generated combining an ECG segment with an alternant wave, and noise [5, 6, 9, 11–13]. This approach enables TWA detection due to the actual knowledge of the alternans characteristics. ECG, alternans and noise signals may be real or simulated. More realistic strategies are closer to being able to replicate the nonstationary nature of the real phenomenon.

Machine learning (ML) and deep learning (DL) techniques have been extensively applied to ECG problems in recent years. Thorough reviews can be found in [17] for DL and in [18] for ML. The authors conclude that ML approaches help to improve data-driven decision making in the diagnosis of heart diseases. However, they also found that researchers mainly focus on model's performance rather than in interpretability and explainability. Additionally, there have been scarce attempts to tackle the challenge of TWA detection by means of ML and DL approaches. In [19] various ML classifiers were tested using the T-Wave alternans Database from PhysioNet website. However, this particular database lacks explicit labels, instead is ranked according to the level of T–wave al-

ternans present in the ECG. Consequently, the authors had to set a threshold on this rank in order to assign classification labels (+ TWA and - TWA). They also augmented the database modeling synthetic cardiac cycles and alternans based on parameters from real signals.

In this work we use a realistic database [20], appropriate to test the presence of TWA alternans using ML algorithms. Alternans–free ECG segments were collected from public databases [21], using the widely accepted Spectral Method (SM) as a gold standard. Afterward, TWA was introduced in approximately half of the control segments, adding a real alternant wave [22]. The use of actual ECGs allows to better capture the physiological cardiac dynamics. A set of ML methods, namely, Decision Trees (DT), Random Forest (RF) and K-Nearest-Neighbors (KNN) are used to address the TWA detection problem. Special attention is paid to the training, validation, and test procedures, carefully designing a methodology that prevents overfitting. Furthermore, since ambulatory ECG exhibits high variability, this work aims to evaluate the impact that the analysis window size has in the performance, so, short-term frames of heartbeats (hb) with different sizes were considered.

The remainder of this paper is organized as follows. Section 2 briefly describes the database and the signal model, additionally it reviews the signal processing and ML methods used in this work for TWA detection. The results are shown in Section 3. Finally, some conclusions are derived in Section 4.

2. Methods

2.1. Signal model and database

The absence of annotated databases presents a significant challenge when benchmarking TWA methods. Various approaches have been explored to acquire such signals, including the use of synthetic ECGs and the subsequent introduction of artificial alternans [15]. To maintain authenticity, our prior work [20] utilized real signals from ambulatory recordings obtained from Physionet to guarantee reproducibility. Subsequent works have also adopted this approach of introducing artificial TWAs into authentic ECGs.

The SM serves as the gold standard for detecting potential alternans in these authentic ECGs, allowing for the removal of these segments and the compilation of a TWA-free database. Candidate signals are extracted from three distinct databases: the MIT-BIH Arrhythmia Database (mitdb), the European ST-T Database (edb), and the MIT-BIH Normal Sinus Rhythm Database (nsrdb). Consequently, we obtained 575 signal segments from 30 patients, with an uneven distribution from the three datasets.

Subsequently, these raw control signals are organized into a patient-balanced dataset, with each patient contributing 25 frames, randomly selected. This approach is designed for tracking short-duration TWA episodes, resulting in a total of 750 instances, which constitute the definitive database. The number of heartbeats comprising each frame is referred to as the window size. While our previous work considered this size to be optimal at 32 heartbeats per frame, this study explores variations, including longer and shorter windows, such as 40 and 20 heartbeats, respectively, as well as a middle point of 30 heartbeats. Additionally, a subsequent preprocessing step is carried out on the signals, involving signal resampling, baseline wander removal, lowpass filtering, and other necessary adjustments. For more detailed information, please refer to [20].

Real alternant waves of 35 μV where added to roughly half of the 25 frames obtained from each patient, specifically to 13, whereas the remaining 12 frames were alternans-free.

2.2. Signal Processing and Machine Learning methods

While the SM is widely regarded as the gold standard and has gained general acceptance for benchmarking alternans detection, it necessitates specific conditions for its application, such as quasi-stationary signal conditions and the requirement for stress tests, which are incompatible with ambulatory recordings. This is why the utilization of ML aims to integrate some of the most well-established methods to enhance predictions in TWA detection, rather than relying solely on a single technique. In our prior work, three of these methods were selected, each of which has demonstrated its performance through various clinical studies: TM, MMA, and the previously mentioned SM.

To accomplish this objective, a feature extraction step is employed, during which the three obtained features are input into the classification algorithms. For TM and MMA, the TWA amplitude is estimated, while the SM is evaluated using the K score, with a threshold of 3 or higher indicating the presence of TWA in a given segment. Consequently, each of the 750 collected instances corresponds to a set of three features.

The ML methods comprising the classification framework in this study facilitate the categorization of each instance into either 0 (representing no TWA) or 1 (indicating the presence of TWA).

KNN, as an instance-based learning algorithm, classifies a test instance by identifying the K nearest neighbors from the training dataset based on the Euclidean distance metric. The value of the hyperparameter K, which represents the number of neighbors considered, is determined through a grid search procedure.

In contrast, the DT algorithm relies on a tree-like struc-

ture composed of decision rules. This algorithm optimizes two key hyperparameters: the minimum leaf size and the maximum number of splits. These hyperparameters significantly influence the structure and complexity of the decision tree and are also optimized using a grid search scheme.

Lastly, RF is an ensemble learning method that harnesses the power of multiple decision trees to enhance prediction accuracy and mitigate overfitting. In addition to fine-tuning the two aforementioned hyperparameters (minimum leaf size and maximum number of splits) for individual decision trees, RF also refines the number of trees employed within the ensemble.

To assess the performance of the models, a five-fold cross-validation (CV) step is initially implemented to fine-tune the different hyperparameters of the models. This is conducted while ensuring the avoidance of potential intra-patient overfitting effects, meaning that patients in the training set are not present in the test set. Additionally, to evaluate the robustness of the ML algorithm, a permutation procedure of the test set is also carried out, as comprehensively detailed in [20]. In this procedure, the set is initially divided into six groups, with the first five being analyzed through the 5-fold CV, while the sixth group constitutes the test set. In each permutation, the test set transitions to one of the other groups, and the model's performance is reassessed. This iterative process continues until every group has served as the test set, ultimately yielding a mean and standard deviation score for the model's performance.

3. Results

Table 1 shows the results for each ML method and for each analysis window length. F1-score is presented since it combines both precision and recall in one metric. The results are shown as the mean $\pm$ the standard deviation of the permutation procedure presented in section 2.2. All three ML methods exhibit comparable performance. Test results show an increasing F1-score trend as the analysis window length increases. In order to assess if the improvement for longer windows is statistically significant, a nonparametric hypothesis test, based on Bootstrap resampling, is conducted. Let $F1_{w1}$ and $F1_{w2}$ denote the F1-scores for two different window lengths and let $\Delta F1 = F1_{w2} - F1_{w1}$ be the difference. The hypothesis test will contrast the null hypothesis, $H_0 : \Delta F1 = 0$, that both window lengths have the same performance, against the alternative hypothesis, $H_1 : \Delta F1 \neq 0$, that they have different performance. In order to approximate the probability density function (*pdf*) of $\Delta F1$, we use Bootstrap resampling. In each resampling iteration the 750 predictions of the test sets for both $w1$ and $w2$ are resampled with replacement and paired to the true labels, and F1-scores $F1^*_{w1}(b)$ and $F1^*_{w2}(b)$ are computed.

An estimation of the confidence interval (CI) for $\Delta F1$, can be readily obtained from the ordered statistics $F1^*_{w1}(b)$ and $F1^*_{w2}(b)$ obtained in each resampling. The differences between the two methods are statistically relevant in terms of statistic $F1$ when the 95% IC of $\Delta F1$ does not overlap the zero value [23].

Table 1. F1-scores of ML models for three analysis window sizes. Results are shown as the mean $\pm$ the standard deviation.

		20 hb	**30 hb**	**40 hb**
DT	Train	0.86 $\pm$0.01	0.89 $\pm$0.01	0.91 $\pm$0.00
	Test	0.81 $\pm$0.05	0.87 $\pm$0.05	0.89 $\pm$0.03
KNN	Train	0.87 $\pm$0.01	0.89 $\pm$0.01	0.91 $\pm$0.01
	Test	0.84 $\pm$0.03	0.87 $\pm$0.05	0.90 $\pm$0.02
RF	Train	0.87 $\pm$0.01	0.89 $\pm$0.01	0.92 $\pm$0.00
	Test	0.82 $\pm$0.05	0.87 $\pm$0.05	0.89 $\pm$0.02

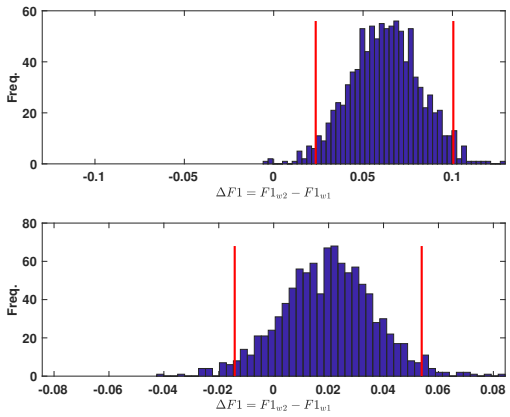

Figure 1. Estimated pdfs for $\Delta F1$. The 95% CI is represented with bars. Top panel for DT, and 20 vs 30 hb windows. Bottom panel for DT, and 40 vs 30 hb windows.

Nine hypothesis test were conducted to compare the F1-scores, within each ML method for the three windows lengths. Windows of 40 hb significantly outperformed 20 hb windows. Comparing closer window lengths, statistically significant improvement was found for windows of 30 hb compared to 20 hb (DT and RF), but not for windows of 40 hb compared to 30 hb. Figure 1 shows an example of the estimated pdfs for $\Delta F1$ represented as an histogram. The top panel, for the comparison with DT, and 20 vs 30 hb windows, shows significant improvement (since the IC does not contain the zero) for the 30 hb window (since the IC lays on the right side). The bottom panel shows no significant difference for the comparison with DT, and 40 vs 30 hb windows.

4. Conclusions

We conclude that all three ML models yield comparable scores and can learn from signal excerpts of varying

lengths to identify alternant waves of 35 μV. Also longer analysis windows showed better performance in terms of F1-score, but the statistical analysis possibly indicates an upper limit for the window length in terms of performance improvement.

Acknowledgments

This work has been partially supported under research project grants EPU-INV/2020/002 from Community of Madrid and PID2022-140786NB-C32 from Spanish Ministry of Science and Innovation.

References

[1] Smith JM, Clancy EA, Valeri CR, Ruskin JN, Cohen RJ. Electrical alternans and cardiac electrical instability. Circulation January 1988;77(1):110–121.

[2] Rosenbaum DS, Albrecht P, Smith JM, Garan H, Ruskin JN, Cohen RJ. Electrical alternans and vulnerability to ventricular arrhythmias. New England Journal of Medicine January 1994;330(4):235–241.

[3] Merchant FM, Sayadi O, Moazzami K, Puppala D, Armoundas AA. T-wave alternans as an arrhythmic risk stratifier: state of the art. Current Cardiology Reports 2013; 15(9):1–9.

[4] Gimeno-Blanes FJ, Blanco-Velasco M, Barquero-Pérez O, García-Alberola A, Rojo-Álvarez JL. Sudden cardiac risk stratification with electrocardiographic indices - a review on computational processing, technology transfer, and scientific evidence. Frontiers in Physiology 2016;7(82).

[5] Martínez JP, Olmos S. Methodological principles of T wave alternans analysis: a unified framework. IEEE Transactions on Biomedical Engineering April 2005;52(4):599–613.

[6] Burattini L, Bini S, Burattini R. Comparative analysis of methods for automatic detection and quantification of microvolt T-wave alternans. Medical Engineering Physics 2009;31(10):1290–1298. ISSN 1350-4533.

[7] Nearing BD, Huang AH, Verrier RL. Dynamic tracking of cardiac vulnerability by complex demodulation of the T wave. Science 1991;252(5004):437–440.

[8] Nearing BD, Verrier RL. Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy. Journal of Applied Physiology 2002; 92(2):541–549.

[9] Monasterio V, Clifford G, Laguna P, Martínez JP. A multilead scheme based on periodic component analysis for T-wave alternans analysis in the ECG. Annals of Biomedical Engineering 2010;38(8):2532–2541.

[10] Burattini L, Zareba W, Burattini R. Adaptive match filter based method for time vs. amplitude characterization of microvolt ECG T-wave alternans. Annals of Biomedical Engineering 2008;36(9):1558–1564.

[11] Romero I, Grubb N, Clegg G, Robertson C, Addison P, Watson J. T-wave alternans found in preventricular tachyarrhythmias in CCU patients using a wavelet transform-based methodology. IEEE Transactions on Biomedical Engineering Nov 2008;55(11):2658–2665. ISSN 0018-9294.

[12] Ghoraani B, Krishnan S, Selvaraj RJ, Chauhan VS. T wave alternans evaluation using adaptive time–frequency signal analysis and non-negative matrix factorization. Medical Engineering Physics 2011;33(6):700–711.

[13] Monasterio V, Laguna P, Martínez JP. Multilead analysis of T-wave alternans in the ECG using principal component analysis. IEEE Transactions on Biomedical Engineering 2009;56(7):1880–1890.

[14] Blanco-Velasco M, Cruz-Roldán F, Godino-Llorente JI, Barner KE. Nonlinear trend estimation of the ventricular repolarization segment for T–wave alternans detection. IEEE Transactions on Biomedical Engineering 2010; 57(10):2402–2412.

[15] Blanco-Velasco M, Goya-Esteban R, Cruz-Roldán F, García-Alberola A, Rojo-Álvarez JL. Benchmarking of a T–wave alternans detection method based on empirical mode decomposition. Computer Methods and Programs in Biomedicine 2017;145:147–155. ISSN 0169-2607.

[16] Goya-Esteban R, Barquero-Pérez O, Blanco-Velasco M, Caamaño-Fernández A, García-Alberola A, Rojo-Álvarez J. Nonparametric signal processing validation in T–wave alternans detection and estimation. IEEE Transactions on Biomedical Engineering April 2014;61(4):1328—1338. ISSN 0018-9294.

[17] Somani S, Russak AJ, Richter F, Zhao S, Vaid A, Chaudhry F, De Freitas JK, Naik N, Miotto R, Nadkarni GN, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. EP Europace 2021;23(8):1179–1191.

[18] Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine 2022;128:102289.

[19] Karnaukh O, Karplyuk Y. Application of machine learning methods for artificial ECG with T–wave alternans. In IEEE 40th International Conference on Electronics and Nanotechnology. 2020; 613–617.

[20] Fernández-Calvillo MG, Goya-Esteban R, Cruz-Roldán F, Hernández-Madrid A, Blanco-Velasco M. Machine learning approach for twa detection relying on ensemble data design. Heliyon 2023;9(1).

[21] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation June 2000;101(23):215–220.

[22] Martínez JP, Olmos S, Wagner G, Laguna P. Characterization of repolarization alternans during ischemia: time-course and spatial analysis. IEEE Transactions on Biomedical Engineering April 2006;53(4):701–711.

[23] Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press, 1994.

Address for correspondence:

Lidia Pascual Sánchez

Department of Teoría de la Señal y Comunicaciones, Universidad de Alcalá

lidia.pascual@uah.es