Available online at www.sciencedirect.com
ScienceDirect
AIM
Artificial Intelligence in Medicine 00 (2024) 1–11

Active Learning and Margin Strategies for Arrhythmia Classification in Implantable Devices

José-María Lillo-Castellano[a], Inmaculada Mora-Jiménez[a], María Martín-Méndez[b], Laia Cerdá[b], Arcadi García-Alberola[c], José Luis Rojo-Álvarez[a,d,*], Devis Tuia[e]

[a] *Universidad Rey Juan Carlos, Department of Signal Theory and Communications, Telematics and Computing, Cam. del Molino, 5. 28942 Fuenlabrada (Madrid,Spain*
[b] *Medtronic Ibérica ® S.A, Dep. Cardiac Rhythm and Heart Failure. C/ María de Portugal 9, 28050 (Madrid, Spain).*
[c] *Hospital CU Virgen de la Arrixaca. Arrhythmia Unit, Ctra. Madrid-Cartagena, s/n. 30120-El Palmar (Murcia, Spain)*
[d] *DIlemma Ltd startup, Cam. del Molino, 5. 28942-Fuenlabrada (Madrid, Spain).*
[e] *Ecole Polytechnique Fédérale de Lausanne, Environmental Computational Science and Earth Observation Laboratory. 1950-Sion (Switzerland).*

Abstract

Background and Objectives. The massive storage of cardiac arrhythmic episodes from Implantable Cardioverter Defibrillators (ICD) and the availability of new artificial intelligence algorithms are opening up new opportunities for electrophysiological knowledge extraction. However, expert cardiologists must provide accurate and reliable episode labeling, which is manual, expensive, and time-consuming. *Methods.* In this work, we propose using Active Learning (AL) to design classification models that streamline the manual process of labeling cardiac arrhythmic episodes. When AL is used, relevant episodes for classification are selected and then presented to the human expert to be labeled so that the burden associated with the labeling process can be dramatically reduced. *Results.* We adapted four large-margin-based AL strategies to a previously proposed classification methodology. We benchmarked them in problems involving 3 and 8 arrhythmia types in 9908 episodes extracted from a massive national repository of ICD data. Specifically, the relevance of the episode-patient diversity for classification was evaluated. Results showed that the gold standard performance (obtained using all episodes) was achieved when approximately 20% (50%) episodes from 60% (85%) of patients were included in the 3-class (8-class) model design. *Conclusions.* We can conclude that AL techniques are beneficial for designing classification models and can streamline the human labeling process of massive ICD datasets.

Keywords: Implantable Cardioverter Defibrillator, Episode of Cardiac Arrhythmia, Statistical Classification, Expert Labeling.

1. Introduction

Predictive data analytics is assuming a leading role in clinical practice by providing healthcare professionals with support for their daily decision-making. Advances in analytics technologies have been significant,

*Corresponding author
Email address: joseluis.rojo@urjc.es (Devis Tuia)

enabling the development of new models to address a wide range of clinical classification problems. Data-driven classification models are typically derived from Machine Learning (ML) techniques and often rely on sample similarity measures. Cardiovascular disease has been an intense area of activity in recent years, with many reviews in the past few years focusing on the clinical applications and basic cardiovascular research of ML techniques [1]. A review of ECG detection and classification based on learning systems was conducted [2], concluding that the primary focus is often accuracy and noting the increasing interest in wearable technologies for applying ML and Deep Learning (DL) approaches.

But in all cases, when designing models with massive datasets, opting for a curated subset comprising relevant and representative samples becomes advantageous. This is due to the exponential increase in computational memory and processing demands as more samples are integrated into the design of data-driven models. Within the ML literature, this approach of selecting pertinent samples is known as Active Learning (AL). A fundamental concept in AL is that a classification model trained with a thoughtfully chosen reduced set of samples can generalize as effectively as one trained with a larger set of randomly selected samples [3]. To achieve this, AL is characterized by an iterative exchange between humans and the model. In this process, the model identifies samples with classification uncertainty and presents them to the human for labeling. The human then annotates these uncertain samples, contributing to the continuous improvement of the model generalization capabilities. To date, the effectiveness of AL has been demonstrated in a large number of different nature applications, such as remote sensing [4, 5, 6, 7], humanitarian geolocalization [8], industrial robotics [9], or biomedical applications [10, 11, 12, 13], among many others.

A field of relevance in cardiac arrhythmia monitoring is implantable devices. In recent years, an increasing number of individuals facing a heightened risk of sudden cardiac events have received treatment by implanting an Implantable Cardioverter Defibrillator (ICD). This small, battery-powered device is surgically placed beneath the skin and operates with constrained memory and computational capabilities. It is fitted with slender wires, known as leads, positioned within the right ventricular apex, sometimes in the septum, and the right atrium or the left ventricular epicardium. These leads enable the ICD to record intracardiac electrical signals, called electrograms or EGMs. These EGMs are then used for automatically detecting arrhythmic episodes, triggering the application of various therapies such as pacing, cardioversion, or defibrillation when necessary [14, 15]. Present-day ICDs retain the corresponding EGM data when an arrhythmic episode is detected. Subsequently, this data can be uploaded into a centralized database during the patient's visit to the cardiologist or even through remote follow-up transmissions. This approach empowers cardiologists to meticulously review each arrhythmic episode and assess, during subsequent follow-ups, whether the detection and treatment were executed effectively or require changes [14].

In our prior research [16], we demonstrated that innovative approaches utilizing compression similarity measures can effectively classify ICD arrhythmic episodes with high accuracy. The dataset we utilized exemplifies a common challenge in analyzing biomedical signals: it requires expert revision for accurate labeling,

which can pose challenges when employing ML methods. Despite those challenges, only a few approaches have been proposed for leveraging AL in the context of large cardiac signal repositories [17, 18]. In our effort to assess how AL can enhance the performance of automatic arrhythmic episode classification methodologies, we compare the results of applying four distinct AL approaches based on large-margin techniques. It is important to emphasize that these AL algorithms are particularly well-suited to ML classification schemes, such as Support Vector Machines (SVMs) used in the proposed methodology. Moreover, implantable devices need to classify arrhythmia episodes while being power and memory-constrained, so they need to work with algorithms with as few parameters as possible. In this setting, SVM classifiers only need to compare samples to training ones, so these training samples need to be wisely chosen, which is a strength of AL techniques. Additionally, we investigate the importance of patient diversity in the classification process, which is relevant in the context of ECG-based ML systems.

2. Active Learning

Let us consider a labeled dataset $\mathcal{X} = \{(\mathbf{x}_i, l_i)\}_{i=1}^{I}$ used to train a classification model, where samples $\mathbf{x}_i \in \mathbb{R}^d$ (d is the dimension of the feature space) and labels $l_i \in \{C_1, \ldots, C_L\}$, with L the number of different classes and these labels are available for a limited number of samples I. Let us also consider an unlabeled dataset $\mathcal{U} = \{\mathbf{x}_j\}_{j=1}^{J}$, known as the *pool of candidates*, with samples $\mathbf{x}_j \in \mathbb{R}^d$ and cardinality $J \gg I$. An AL algorithm aims to iteratively select those samples from $\mathcal{U}$, which could improve the classification performance of the model designed with $\mathcal{X}$. In AL, the selection criterion is a *heuristic* [3] usually based on the discovery of *uncertain* or *low confidence* regions of the feature space. If the model solved such regions, the classification performance would greatly improve. In short, the general learning procedure of an AL algorithm can be described as follows [3]: (1) for a given iteration ϵ, the set of C_ϵ candidates, $\mathcal{S}^\epsilon = \{\mathbf{x}_j\}_{j=1}^{C_\epsilon}$, is selected from $\mathcal{U}^\epsilon$ following a predefined heuristic; (2) the selected candidates are labeled by the human, i.e., $\{l_j\}_{j=1}^{C_\epsilon}$ are determined; (3) the labeled set $\mathcal{S}'^\epsilon = \{(\mathbf{x}_j, l_j)\}_{j=1}^{C_\epsilon}$ is added to $\mathcal{X}^\epsilon$, i.e., $\mathcal{X}^{(\epsilon+1)} = \mathcal{X}^\epsilon \cup \mathcal{S}'^\epsilon$, and $\mathcal{S}^\epsilon$ is removed from $\mathcal{U}^\epsilon$; (4) the model is retrained; (5) the process is repeated until a stopping criterion is met.

Hence, an AL algorithm requires the interaction between humans and models. The human provides knowledge of the task through the labeling, whereas the model provides computational capacity. The selection strategy defining a heuristic is crucial for the AL algorithm and sets the success of the learning process. Different AL algorithms have been proposed, and at least four AL heuristic families are differentiated [3], those based on committees, large margins (specific to SVMs), posterior probability, and clusters. Due to the nature of our classification task for arrhythmic episodes and mainly to the classification methodology used, we focus on the AL algorithm based on large-margin heuristics in this work. An AL algorithm based on large-margin is characterized by quantifying the *uncertainty* depending on the distance to the separating

classification hyperplane. Classifiers such as SVMs [19], used in our previous proposed methodology, are naturally tailored for this type of AL philosophy. In short, the absolute value of the decision function (before taking its sign) can be used to quantify our confidence in the classification of a sample.

Let us consider a binary classification problem (L=2). The distance of a sample $\mathbf{x}_*$ to the SVM hyperplane is given by:

$$d(\mathbf{x}_*) = \sum_{i=1}^{I} \alpha_i l_i k(\mathbf{x}_i, \mathbf{x}_*) + b \tag{1}$$

where $k(\mathbf{x}_i, \mathbf{x}_*)$ is a kernel function defining the similarity between the unlabeled candidate $\mathbf{x}_*$ and the I samples $\{\mathbf{x}_i\}_{i=1}^{I}$ used for designing the SVM classifier; $l_i \in -1, 1$; and α_i values denote the coefficients associated with each training sample. Samples with an associated nonzero α coefficient are known as *support vectors* and contribute to defining the classification model. More detail on SVM description and kernel functions can be found in the SVM basic literature [19, 20].

Conceptually, SVMs are based on a sparse representation of the training set. In this setting, relevant samples are those with a high or nonzero associated coefficient. Hence, samples lying within the model margin potentially contain relevant information, which would improve the classification performance by shifting the position of the classification hyperplane. Different large-margin-based heuristics can be defined depending on the classification task [3]. The following describes the most common heuristics used in this work.

Margin Sampling (MS). This heuristic is also known as *most ambiguous* [21, 3] and is characterized by taking the SVM geometrical properties into account. Specifically, it considers support vectors [22], and candidates are selected by minimizing the following cost function:

$$\mathcal{S}_{MS}^{\epsilon} = \underset{\mathbf{x}_j \in \mathcal{U}}{argmin} \left\{ \max_{c} \{d_c(\mathbf{x}_j)\} \right\} \tag{2}$$

where $d_c(\mathbf{x}_j)$ is the distance of the sample $\mathbf{x}_j$ to the hyperplane defined for the $c \in \{C_1, \ldots, C_L\}$ class in a one-against-all multi-class scenario.

Breaking Ties (BT). This heuristic was proposed by Luo *et al.* [23] and characterizes uncertainty as the difference between the distance to the hyperplane and the nearest data points on each side of it, i.e., the following cost function is minimized:

$$\mathcal{S}_{BT}^{\epsilon} = \underset{\mathbf{x}_j \in \mathcal{U}}{argmin} \{d_{C_1}(\mathbf{x}_j) - d_{C_2}(\mathbf{x}_j)\} \tag{3}$$

where

$$d_{C_1}(\mathbf{x}_j) = \max_{c} \{d_c(\mathbf{x}_j)\}$$

$$d_{C_2}(\mathbf{x}_j) = \max_{c \setminus C_1} \{d_c(\mathbf{x}_j)\}$$

and C_1 is the class associated to d_{C_1}.

Level Uncertainty with Diversity. In AL algorithms, the sample selection is performed by batches, i.e., more than one sample is selected per iteration. In this case, one wants to avoid selecting similar high-uncertainty samples. The concept of data *diversity* essentially concerns the ability to select samples as different as possible from each other and Depending on each problem, data diversity has been widely studied in the AL field. For our real clinical scenario of arrhythmic episodes, the diversity can be defined using the patient information, i.e., the patient who suffered from the arrhythmic episode. Thus, aiming to analyze the effect of this diversity, in this work, we defined two heuristics with patient diversity (PD) from the two previous ones, i.e., the MS-PD and the BT-PD. For both heuristics, we denote the pool of candidates as $\mathcal{U} = \{(\mathbf{x}_j, p_j)\}_{j=1}^{J}$, with patient indicators $p_j \in \{1, \ldots, P\}$ and P as the number of different patient. Then, MS-PD and BT-PD heuristics candidates are selected by minimizing cost functions in Eqs. (2) and (3) , respectively, but subject to the following constraint:

$$p_j \neq p'_j \quad \forall \mathbf{x}_j, \mathbf{x}'_j \in \mathcal{S}^{\epsilon} \tag{4}$$

Note that Eq. (4) ensures that patient diversity takes priority over episode uncertainty.

3. Database of arrhythmic episodes

A clinical classification scenario requiring AL techniques is the Big-Data Scientific COOperation Platform (SCOOP) project of cardiology developed in Spain by Medtronic® [24]. The classification of cardiac arrhythmic episodes recorded by ICDs was manually carried out in SCOOP by a committee of expert cardiologists, by labeling each arrhythmic episode into eight different categories (see [16] for further details). Moreover, SCOOP was included in an observational research study called UMBRELLA and informed consent was obtained as detailed in [25, 26], ensuring the legal regulations for scientific data exploitation and patient privacy.

For this work, a set of 9908 arrhythmic episodes recorded by ICDs from 840 patients in 44 different Spanish hospitals were extracted from SCOOP. Each episode consists of two EGMs sampled at 128Hz and 1 byte of amplitude resolution and an associated set of ICD time notes (known as *markers*). Episodes were recorded either using the ICD configuration pair [*CanToHVB*, *VtipToVring*] or [*AtipToAring*, *VtipToVring*] (4051 and 5857 episodes from 324 and 517 patients, respectively). More details about these configurations can be found in [16].

Each arrhythmic episode was evaluated, manually analyzed, and labeled through a systematic clinical process. A scientific committee of 6 expert cardiologists (reviewers) with deep ICD backgrounds defined a classification guide (detailed in [16]). The guide considers eight arrhythmic categories based on: (1) the arrhythmia origin; (2) the heart activation events; and (3) the signal EGM waveform. Arrhythmic episodes had lengths of 23.9±15.6 s (mean±std), median length of 19.7 s, and interquartile range of 13.2 s. The

Table 1. Relative occurrence of the arrhythmic categories in the SCOOP dataset used in this work.

Category		**Number of Episodes**		**Occurrence**	
8-class	**3-class**	**8-class**	**3-class**	**8-class**	**3-class**
ST	Atrial	1419	3196	14.32%	32.26%
AF		851		8.59%	
SVT		545		5.50%	
UST		381		3.84%	
SMVT	Ventricular	6132	6525	61.89%	65.85%
VF		393		3.97%	
TWO	Other	121	187	1.22%	1.89%
NS		66		0.67%	

number of episodes per patient was 11.8±22.8, with a median of 4 and an interquartile range of 11. The relative occurrence of each arrhythmic category (8 classes) in this work is shown in Table 1, as well as their grouping into three major sets (3 classes, broadly used in clinical practice) according to the arrhythmia origin, namely, atrial, ventricular (which are the more clinically urgent), and other episodes. Atrial rhythms included Sinus Tachycardia (ST), Atrial Fibrillation (AF), Supraventricular Tachycardia or Flutter (SVT), and Uncertain Supraventricular Tachycardia (UST). Ventricular rhythms included Sustained Monomorphic Ventricular Tachycardia (SMVT) and Sustained Polymorphic or Ventricular Fibrillation (VF). Other rhythms included T-wave oversensing (TWO) and Noise Sensing (NS). See [27] for a detailed description of the grouping criteria.

4. Experimental Setup

This section describes the figures of merit and the validation strategy used to assess the performance of AL heuristics. Likewise, our experimental setup and kernel functions are also detailed.

Figures of Merit. The Accuracy Rate (AR) [28] is the most common merit figure used to quantify performance in classification scenarios. However, when imbalance is present in the data (i.e., classes are not equally represented), AR may lead to wrong conclusions since majority classes can mask failures by saturating classifiers and ignoring the minority classes. Thus, when imbalance is present, better performance interpretations can be achieved by considering other merit figures such as the specificity, the sensitivity, the area under the ROC curve (AUC), the F-Score, or the Cohen's Kappa coefficient, among others [28]. For this work, the Cohen's Kappa coefficient (κ) [28] has been used to assess the classifier performance of the 3- and 8-class imbalanced scenarios aiming at complementing the AR merit figure. The Kappa coefficient

expresses how reliable a classifier performance is, considering imbalance and quantifying whether the level of agreement between the target and the model output could be due to random chance [28].

Performance Assessment. In this work, a validation strategy based on random resampling without replacement was used to assess the performance using AL algorithms. In this strategy, the original dataset is 1000 times (runs) randomly divided into two different subsets: (1) a training set containing episodes from the 70% of patients of the dataset; and (2) a test set with episodes of the remaining 30% of patients. The consideration of runs for using different training and test sets allowed us to get an empirical confidence interval of the merit figures, thus providing information about the approach stability. Likewise, this strategy estimates the model performance when no episodes of the same patient are simultaneously considered in training and test sets. The latter allows us to evaluate the generalization capability of the AL heuristic in real-world clinical scenarios as new patients are incorporated, consequently reducing the over-fitting risk.

Experimental Methodology. We have considered the methodology proposed by the authors in [16] to classify arrhythmic episodes and evaluate the AL heuristics. The methodology is mainly composed of three stages: (1) A minimal preprocessing of arrhythmic episodes; (2) The extraction of a data compression-based kernel among episodes (a similarity matrix); And (3) the classification using a free-parameters kernel-based SVM classifier. In AL algorithms, the labeled dataset $\mathcal{X}$ was initialized by random selection of 1% of the episodes from the training set, considering the remaining 99% of them as the unlabeled $\mathcal{U}$ dataset. For each heuristic and ϵ iteration, 1% of the episodes of the training set were selected from $\mathcal{U}$ as candidates, i.e., $\mathcal{S}^{\epsilon}$ is added to $\mathcal{X}$. Thus, $\epsilon = 100$ iterations were computed for each heuristic and run. Besides aiming to analyze the patients' diversity, the percentage of patients who included at least one episode in $\mathcal{X}$ was also evaluated.

Two bounds (upper and lower limit) on the classification performance were computed for comparison. The upper limit is obtained when learning with the whole training dataset, i.e., $\mathcal{X} \cup \mathcal{U}$. For estimating the lower limit in performance, we assessed a model such that $\mathcal{S}^{\epsilon}$ candidates were randomly selected in each iteration, getting the *Random Selection* (RS) heuristic.

Similarity Among Arrhythmic Episodes. The data compression-based kernels computed for our 8 and 3 class scenarios are represented in Fig. 1b and 1a, respectively. Through these graphical representations (built sorting episodes by class label and similarity degree), the effectiveness of a kernel can be readily evaluated. Green or light (blue or dark) colors are associated with high (low) similarity values in these panels. Thus, if these kernels are compared with their associated ideal case (Figs. 1d and 1c, i.e., when similarity among samples of the same class is either 1 or 0), it can be noted that a high degree of similarity among episodes of the same class is present. However, many episodes also present a high degree of similarity with other classes, suggesting potential inefficiencies in the kernel functionality for such instances or even raising the possibility of labeling errors.

5. Results

Performance and comparisons among AL heuristics are presented in Fig. 2. A total of 5 heuristics (4 based on large-margin and one associated with RS) were evaluated in both 8-class and 3-class schemes. In Figs. 2 (a) to (d), the median value of performance (AR and κ) for the lower (RS heuristic) and upper (gold standard) bounds are represented by employing a red curve and a horizontal black line, respectively. The four curves between both bounds are associated with the median values of performance achieved by the four large-margin-based AL heuristics: MS (yellow), BT (purple), MS-PD (blue), and BT-PD (green).

As expected, the performance of MS, BT, MS-PD, and BT-PD was significantly better than those obtained by RS. As samples were selected considering their uncertainty, the gold standard was quickly reached when approximately 50% (8 classes) or 20% (3 classes) of episodes in the training set were included. These results are relevant because they indicate that a high volume of episodes contains redundant or non-relevant information for the classification and that a smart selection strategy can significantly and positively impact the classifier performance. In general, BT heuristics performed better than MS ones because they reached high performance using fewer episodes. However, differences are not significant if the percentage of episodes included in the model is analyzed when the AL heuristics reach the gold standard. In Fig. 2, this percentage is represented by dashed vertical black lines. Specifically, the four heuristics reached the gold standard when 46% (8 classes) or 22% (3 classes) of episodes from the training set were included in the model. Predictably, with more categories, one needs more episodes with relevant information to solve the increase in complexity.

PD was analyzed by evaluating the percentage of patients included in the model with at least one episode per iteration. This evolution is represented in Fig. 3 (a) and (b) for both 8-class and 3-class schemes, respectively. Note that evolution is nonlinearly correlated with including episodes in the model for all heuristics (even for the RS one). The reason for this effect is twofold: (1) patients are not associated with the same number of episodes; and (2) the category imbalance. Moreover, there is a clear difference between the four large-margin heuristics and the RS one. As expected, the MS-PD and BT-PD heuristics include more patients in their classifiers because they have a PD constraint. However, differences with the other two heuristics (MS and BT) are insignificant (close to 1-2%), indicating that uncertainty is not mainly determined by the type of episode but by the patient. Specifically, when approximately 85% (8 class) and 60% (3 class) of patients from the training set are included in the model, gold-standard performance is reached. This result benefits systems such as SCOOP since the manual labeling could be optimized to search for episodes from different patients, and it even would avoid labeling non-representative episodes.

A pertinent aspect of the AL paradigm pertains to how performance within each category undergoes development throughout the training process. In Fig. 4, this evolution is shown for the BT-PD heuristic (similar results were obtained for the rest of heuristics excepting RS) in both 8-class and 3-class schemes.

Note that within the context of the 3-class scheme, the best learning performance for the "Atrial" and "Ventricular" categories is practically achieved when 5% of episodes are incorporated in the model design. However, for the "Other" category, there is an initial decrease followed by a sustained and progressive performance improvement until approximately 20% of episodes are included. This value matches the minimum number of episodes (dashed vertical line) required to reach the gold standard. This result suggests that: (1) the proportion of relevant episodes is lower than expected according to the information provided in Fig. 2. This is likely attributable to imbalance; (2) the minority category essentially conditions the learning process. This conclusion can also be extended to the 8-class scheme, wherein the imbalance is even more pronounced. Specifically, SPVT-VF and NS categories condition the learning, and to a lesser degree, SVT, UST, and TWO categories (which are not the most life-threatening critical ones to detect in ICDs). These results suggest that AL should be oriented towards labeling arrhythmic episodes belonging to minority categories, which would optimize the model learning process. Sampling strategies grounded in point density analysis may be explored [29] to further progress in this direction.

6. Discussion and Conclusion

In this study, we have explored the impact of four AL heuristics on the classification of ICD arrhythmic episodes. Specifically, we compared the performance of a previously proposed classification methodology when employing four large-margin-based AL heuristics in conjunction with a kernel-based SVM classifier. Additionally, we assessed the influence of patient diversity on performance and its effects on the learning process. Our findings highlight the potential advantages of AL techniques in tasks that demand expert knowledge. By fostering collaboration between computational models and domain experts, it is possible to create efficient training datasets containing relevant and representative arrhythmic episodes for classification. These datasets may undergo continual refinement to achieve enhanced performance. Furthermore, AL has the potential to support domain experts in their labeling tasks for arrhythmic episodes, reducing the economic burden associated with achieving high expert consensus.

As presented in this work, AL techniques can be used in current cardiac arrhythmia scenarios for improving classification methodologies in the context of Big Data systems for cardiac arrhythmic episodes. These advancements can potentially facilitate a transition in ICD patient treatment towards a more personalized strategy, reflecting the dynamic healthcare landscape. Their extension to DL approaches and ECG beat classification can also move forward the current state of the art in artificial intelligence for automated systems on arrhythmia discrimination.

Acknowledgment

Special thanks to the researchers and contributors in SCOOP platform and UMBRELLA research study. The authors wish also to thank the Team at Medtronic Ibérica® S.A. for all their support throughout all the project. This work has been partly supported by research projects PI22/01042, PID2019-104356RB-C41 /AEI/ 10.13039/ 501100011033 (meHeart), PID2022-140553OA-C42 /AEI/ 10.13039/ 501100011033 (PCardioTrials), and PID2022-140786NB-C32 /AEI/ 10.13039/ 501100011033 (LATENTIA) from the Spanish Ministry of Science and Innovation. Research partially funded by Comunidad de Madrid (ELLIS Unit Madrid).

Conflict of Interest Statement

Co-authors MMM and LC are employees of Medtronic Ibérica SA. The company provided the data used in this study, however, they have no personal financial interests, relationships, or affiliations that could have influenced the presented results or their interpretation. The remaining authors declare no commercial or financial relationships that could pose a conflict of interest in the research.

References

[1] J. Komuro, D. Kusumoto, H. Hashimoto, S. Yuasa, Machine learning in cardiology: Clinical application and basic research, Journal of Cardiology 82 (2) (2023) 128–133.

[2] S. W. Chen, S. L. Wang, X. Z. Qi, et al., Review of ecg detection and classification based on deep learning: Coherent taxonomy, motivation, open challenges and recommendations, Biomedical Signal Processing and Control 74 (2022) 103493.

[3] B. Settles, Active learning, Synthesis Lectures on Artificial Intelligence and Machine Learning 6 (1) (2012) 1–114. doi:10.2200/S00429ED1V01Y201207AIM018.

[4] B. Kellenberger, D. Marcos, S. Lobry, D. Tuia, Half a percent of labels is enough: Efficient animal detection in uav imagery using deep cnns and active learning, IEEE Transactions on Geoscience and Remote Sensing 57 (12) (2019) 9524–9533.

[5] G. Lenczner, A. Chan-Hon-Tong, B. Le Saux, et al., Dial: Deep interactive and active learning for semantic segmentation in remote sensing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 3376–3389.

[6] J. E. Vargas-Muñoz, P. Zhou, A. X. Falcão, D. Tuia, Interactive coconut tree annotation using feature space projections, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 5718–5721.

[7] B. Kellenberger, D. Tuia, D. Morris, Aide: Accelerating image-based ecological surveys with interactive machine learning, Methods in Ecology and Evolution 11 (12) (2020) 1716–1727.

[8] D. T. John E. Vargas-Muñoz, A. X. Falcão, Deploying machine learning to assist digital humanitarians: making image annotation in openstreetmap more efficient, International Journal of Geographical Information Science 35 (9) (2021) 1725–1745.

[9] G. Ning, H. Liang, X. Zhang, H. Liao, Inverse-reinforcement-learning-based robotic ultrasound active compliance control in uncertain environments, IEEE Transactions on Industrial Electronics 71 (2) (2024) 1686–1696.

[10] T. Mahmood, A. Rehman, T. Saba, et al., Recent advancements and future prospects in active deep learning for medical image segmentation and classification, IEEE Access 11 (2023) 113623–113652.

[11] Z. Zhao, Z. Zeng, K. Xu, et al., Dsal: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation, IEEE Journal of Biomedical and Health Informatics 25 (10) (2021) 3744–3751.

[12] Z. Guo, R. Zhang, Q. Li, et al., Reduce false-positive rate by active learning for automatic polyp detection in colonoscopy videos, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1655–1658.

[13] R. S. Bressan, G. Camargo, P. H. Bugatti, P. T. M. Saito, Exploring active learning based on representativeness and uncertainty for biomedical data classification, IEEE Journal of Biomedical and Health Informatics 23 (6) (2019) 2238–2244.

[14] R. X. Stroobandt, S. S. Barold, A. F. Sinnaeve, Implantable Cardioverter - Defibrillators Step by Step: An Illustrated Guide, 1st Edition, Wiley-Blackwell, 2009.

[15] P. M. Parker, The 2023-2028 World Outlook for Implantable Cardioverter Defibrillators, 1st Edition, ICON Group International, 2022.

[16] J. M. Lillo-Castellano, J. L. Rojo-Álvarez, F. Chavarría-Asso, et al., Classifying cardiac arrhythmic episodes via data compression, Neurocomputing 307 (13) (2018) 1–13.

[17] W. Fan, Y. Si, W. Yang, G. Zhang, Active broad learning system for ecg arrhythmia classification, Measurement 185 (2021) 110040.

[18] G. Sayantan, P. Kien, K. Kadambari, Classification of ECG beats using deep belief network and active learning, Medical & Biological Engineering & Computing 56 (10) (2018) 1887–98.

[19] V. Vapnik, The Nature of Statistical Learning Theory, 2nd Edition, Springer Verlag, 2000.

[20] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, Wiley–Interscience, 2001.

[21] G. Schohn, D. Cohn, Less is more: Active learning with support vectors machines, in: Proc. 17th ICML, Standford, CA, USA, 2000, pp. 839–846.

[22] B. Schölkopf, A. J. Smola, Learning with Kernels, 1st Edition, The MIT Press Cambridge, 2002.

[23] T. Luo, K. Kramer, D. B. Goldgof, et al., Active learning to recognize multiple types of plankton, J. Mach. Learn (6) (2005) 589–613.

[24] Medtronic Ibérica, Dispositivos médicos - Tecnología médica y servicios de la empresa (2016).
URL http://www.medtronic.es/

[25] Medtronic Bakken Research Center, UMBRELLA - Incidence of arrhythmias in Spanish population with a Medtronic implantable cardiac defibrillator implant, identifier: NCT01561144 (March 2012).
URL https://clinicaltrials.gov/

[26] S. Briongos-Figuero, A. García-Alberola, J. Rubio, et al., Long-term outcomes among a nationwide cohort of patients using an implantable cardioverter-defibrillator: Umbrella study final results, Journal of the American Heart Association 47 (7) (2009) 2218–2232. doi:10.1109/TGRS.2008.2010404.

[27] J. M. Lillo-Castellano, J. L. Rojo-Álvarez, F. Chavarría-Asso, et al., Big-data classification of arrhythmic episodes using compression-based kernel methods, IEEE Journal of Biomedical and Health Informatics PP (PP) (2016) x–x. doi:xx.

[28] C. Ferri, J. Hernández-Orallo, R. Modroiu, An experimental comparison of performance measures for classification, Pattern Recognition Letters, Elsevier 30 (1) (2009) 27–38. doi:10.1016/j.patrec.2008.08.010.

[29] D. Tuia, E. Pasolli, W. Emery, Using active learning to adapt remote sensing image classifiers, Remote Sensing of Environment 115 (9) (2011) 2232–2242.

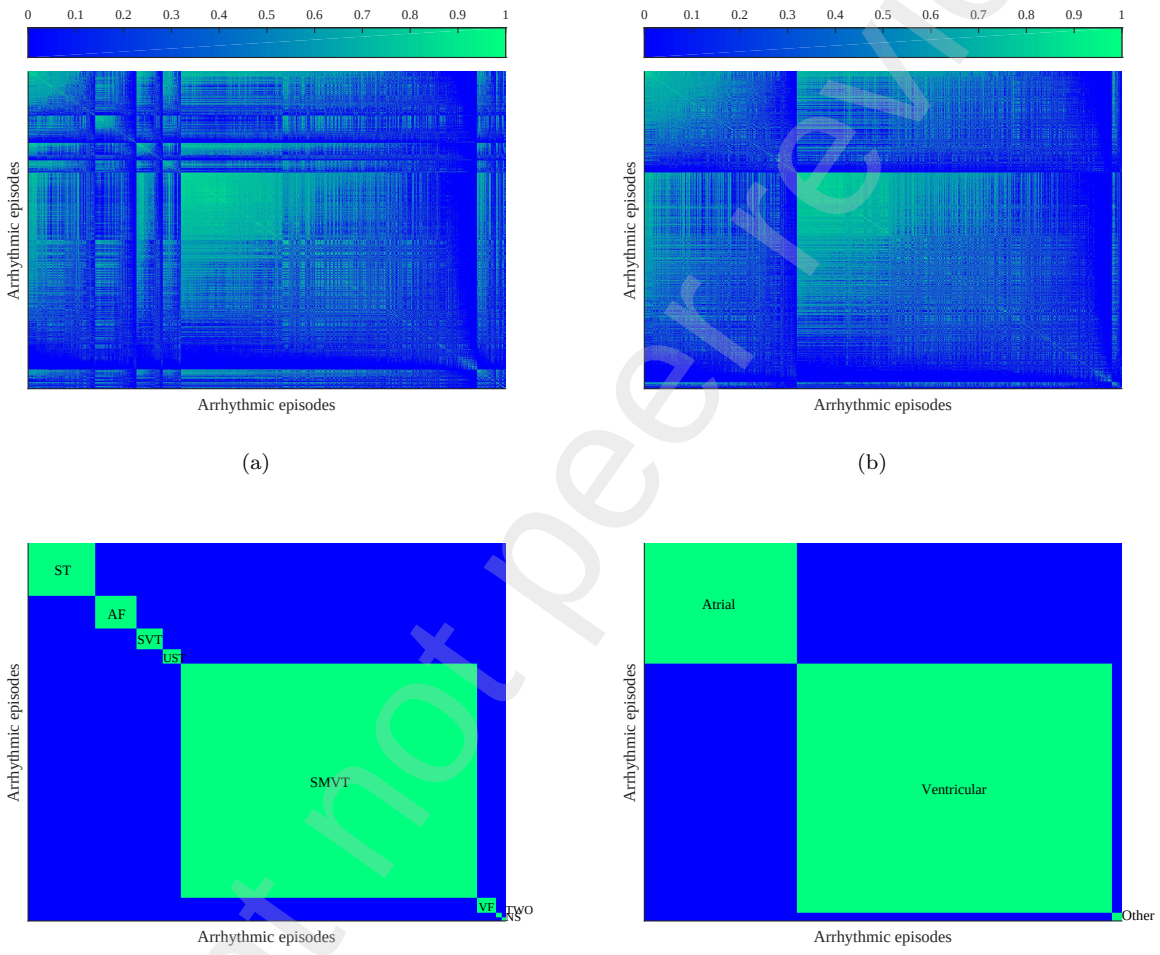


Figure 1. Actual ((a) and (b)) and ideal ((c) and (d)) kernel matrices for both 8-class ((a) and (c)) and 3-class ((b) and (d)) scenarios computed using the 9908 arrhythmic episodes of the SCOOP database. For a better interpretation, episodes are sorted by class and degree of similarity.

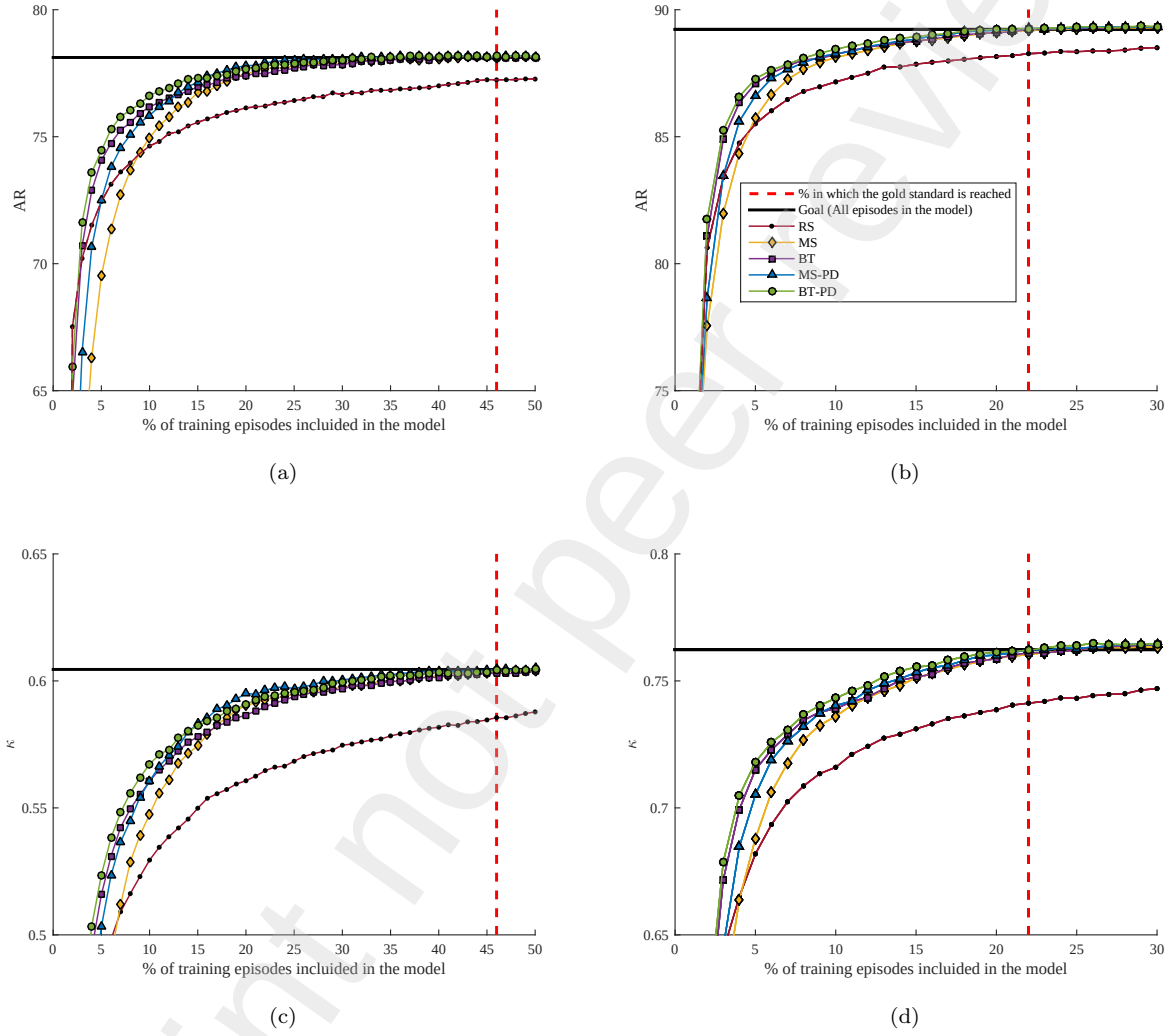
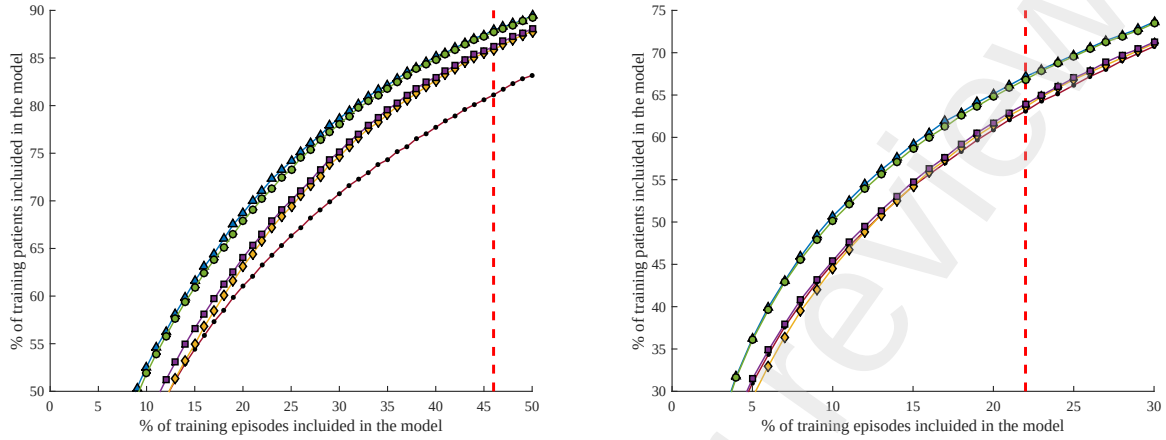
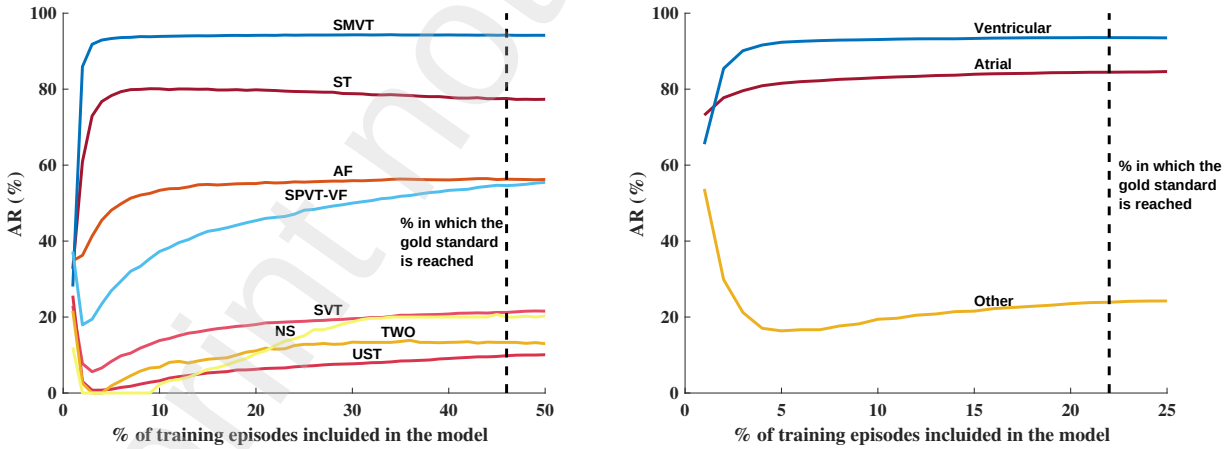


Figure 2. Evolution of the test AR ((a) and (b)) and κ ((c) and (d)) median values considering 1,000 runs when the proposed methodology is AL-streamlined in both 8-class ((a) and (c)) and 3-class ((b) and (d)) schemes. Horizontal black lines denote the gold standard. Dashed vertical lines indicate the percentage of the gold standard reached by the four AL heuristics different from RS.



(a) (b)

Figure 3. Evolution of the percentage of patients included in the model for each heuristic in both 8-class (a) and 3-class (b) schemes. Each dashed vertical line indicates the proportion of the gold standard achieved by the four AL heuristics different from RS. The same legend is used in Fig. 2.



(a) (b)

Figure 4. Evolution of the test AR median values considering 1,000 runs when the proposed methodology is AL-streamlined in both 8 (a) and 3-class (b) schemes. BT-PD heuristic was used for the AL streamlining. Dashed vertical black lines denote the gold standard performance value represented in Fig. 2.