Biomedical Signal Processing and Control 72 (2022) 103269

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Identification of spatiotemporal dispersion electrograms in atrial fibrillation ablation using machine learning: A comparative study

Amina Ghrissi [a], Douglas Almonfrey [b], Fabien Squara [a,c], Johan Montagnat [a], Vicente Zarzoso [a,*]

[a] Université Côte d'Azur, CNRS, I3S Laboratory, Sophia Antipolis, France
[b] Federal Institute of Espírito Santo, Vitoria, Brazil
[c] Université Côte d'Azur, CHU Pasteur, Cardiology Department, Nice, France

ARTICLE INFO

Keywords:
Atrial fibrillation
Catheter ablation
Machine learning
Multichannel electrogram
Multipole catheter
Spatiotemporal dispersion
Transfer learning

ABSTRACT

Atrial Fibrillation (AF) is the most widespread sustained arrhythmia in clinical practice. A recent personalized AF therapy consists in ablating areas displaying spatiotemporal dispersion (STD) electrograms (EGM) with the use of catheters. Interventional cardiologists use a multipolar mapping catheter called PentaRay to identify visually atrial sites with STD pattern by visual inspection. In this contribution, we propose to automatize the identification of STD EGMs using machine learning while comparing several features. The aim is to design a data representation and an adapted classification algorithm for accurate STD detection with affordable computational resources and low prediction time. Four data formats are considered: 1) EGM matrices; 2) EGM plots; 3) three-dimensional EGM plots; 4) maximal voltage absolute values. Convolutional neural networks and transfer learning based on the VGG16 architecture are benchmarked. Classification results on the test set show that extracting features automatically with VGG16 is possible and yields comparable results to classifying raw EGM recordings with values of accuracy and AUC of 90%. However, the overall precision and F1 score are low (50%), which can be explained by the high class imbalance ratio. This issue is addressed with data augmentation. Due to its low computational cost, our solution can also be deployed in real time.

1. Introduction

During the last decade, artificial intelligence models [1,2] have intensively been used as decision-aid solutions in biomedical data analysis. In the current study, we are interested in helping cardiologists to automatically identify potential ablation sites in persistent atrial fibrillation (AF) using machine learning (ML) tools. AF represents the most frequent sustained arrhythmia experienced in clinical practice and it rises in prevalence with advancing age [3]. AF is associated with a 5-fold stroke risk increase and a 2-fold increased risk of both mortality and dementia [4]. Hence, AF entails an important economic burden. Among the existing treatments of persistent AF, ablation interventions beat drug therapies in terms of long-term effectiveness. Ablation consists in burning with radiofrequency (RF) energy cardiac tissue areas estimated to be responsible for the presence and maintenance of AF. The classical ablation approach, called sequential stepwise ablation, yields poor clinical results [5]. A novel wholly patient-tailored ablation protocol, giving 95% of procedural success rate, consists in identifying ablation sites based on a signal pattern called spatiotemporal dispersion (STD) [6]. Multipolar mapping catheters are used to record electrograms (EGM) in the atria thus targeting areas of STD as potential AF drivers. The high-density PentaRay catheter (Biosense Webster, Inc., Irvine, CA, USA) is used for STD localization. It has a five-branch star design with two bipoles on each spline. According to preliminary guidelines for STD identification from visual inspection, the 10-channel EGM recording acquired by the PentaRay would display a cardiac activation delay of 70% of AF cycle length (AFCL) on a minimum of three neighboring bipoles (channels) [6]. However, this visual inspection may be biased by the difficulty for the interventional cardiologist to precisely quantify the STD pattern at each single mapped location in real time, as hundreds and even thousands of cardiac sites are mapped in a typical ablation intervention.

This work was supported in part by the French government Investments in the Future program, through IDEX UCA^{JEDI} (ANR-15-IDEX-0001) and 3IA Côte d'Azur (ANR-19-P3IA-0002) projects. V. Zarzoso holds the Chair "IAblation" from 3IA Côte d'Azur.

* Corresponding author.
E-mail addresses: amina.ghrissi@univ-cotedazur.fr (A. Ghrissi), dalmonfrey@ifes.edu.br (D. Almonfrey), fabien.squara@univ-cotedazur.fr, squara.f@chu-nice.fr (F. Squara), johan.montagnat@univ-cotedazur.fr (J. Montagnat), vicente.zarzoso@univ-cotedazur.fr (V. Zarzoso).

https://doi.org/10.1016/j.bspc.2021.103269
Received 2 March 2021; Received in revised form 7 October 2021; Accepted 15 October 2021

To overcome these limitations, the present contribution aims to design a decision-aid solution that helps interventional cardiologists detect STD patterns automatically thanks to both baseline ML and modern deep learning tools [7]. We seek the most optimal EGM classification model, formed by the combination of a data representation and an adapted classification algorithm, in order to detect STD locations with the highest performance and lowest computational cost in terms of prediction time and resources. The study dataset includes a cohort of over 35000 10-channel EGM signals acquired from 15 different persistent AF patients.

Our preliminary results in a previous study on automatic detection of STD from multichannel EGM recordings were promising [8]. The classification performance on the test dataset reached 90% of accuracy and 80% of AUC using a shallow convolutional neural network (CNN). But values of precision were low for lack of STD samples. We also addressed the issue of class imbalance and lack of training samples from STD class through adapted data augmentation (DA) methods [9,10]. In a complementary recent study, we addressed the classification of handcrafted features from EGM data [11]. As suggested in [6], we studied a time series computed from multichannel EGM signals to perform STD detection. The time series is called maximal voltage absolute values at the PentaRay bipoles (VAVp). Classification results were encouraging in simulated data but disappointing in real data.

The present contribution updates, completes and consolidates our preliminary results in [8,11] by putting forward a feature selection technique for a more optimal EGM classification solution. For this purpose, we design and benchmark different data representations: 1) raw EGMs stored in matrices; 2) two-dimensional (2D) images obtained by subplotting the curves of the 10 leads one under the other, as it is currently done by the mapping system in the hospital; 3) three-dimensional (3D) image tensor as a result of stacking three shifted images of EGM curves; 4) one-dimensional (1D) VAVp time series used as a compact representation of the multichannel recordings. These data representations are combined with suitable classifiers including multivariate logistic regression (MLR) [2], dimensionality reduction with principal component analysis (PCA) [12] followed by support vector machines (SVM) [13], CNN [14], and transfer learning (TL) [7] of the VGG16 model [15]. The originality of this contribution is threefold: first, to our knowledge this is the first systematic application of state-of-the-art ML techniques to EGM classification for STD identification; second, different data representations are benchmarked and TL is applied for the first time to the identification of dispersion patterns in multichannel EGMs; third, our decision-aid solution can be implemented in real time with moderate computational resources, thus potentially improving catheter ablation success rates while reducing the duration of STD-based ablation interventions for treating persistent AF.

2. AF Ablation

2.1. Catheter ablation of persistent AF

2.1.1. AF diagnosis

The heart is a vital muscle that pumps blood to irrigate the body thus providing oxygen and nutrients to body tissues. It is composed of four chambers, the two upper ones are called atria and the two lower ones are called ventricles. However, the heart mechanisms for a patient suffering from AF is different from that of a normal subject, said to be in sinus rhythm. AF is characterized by an irregular activation of the atria that start quivering or fibrillating, causing non-synchronous fluctuations in the associated electrical baseline. As a result of abnormal atrial activation, the ventricular rate becomes more rapid and disorganized [16].

2.1.2. AF therapies

The severity degrees of AF differ according to the arrhythmogenic episode duration and the response to treatment. In case of persistent AF, pharmacotherapy proves less effective than catheter ablation due to the complexity of this arrhythmia. Ablation is an invasive procedure that consists in burning the cardiac myocytes displaying irregularities with RF energy delivered with the use of catheters. The classical ablation protocol of persistent AF uses bipolar mapping catheters. It is called stepwise and consists: in 1) burning the *triggers* around the pulmonary veins thought to be responsible for initiating AF; 2) ablating areas of the atrial substrate harboring *drivers* maintaining and self-perpetuating the arrhythmia, such as complex fractionated electrograms (CFAE) [17]. However, a growing number of reports show the limitations of the stepwise approach, as several ablation interventions are typically required to terminate AF or at least to transform it to a more stable tachycardia [5].

2.2. STD-guided ablation

A recent discovery in AF ablation therapy shows that targeting only cardiac areas displaying STD EGMs can terminate AF in 95% of a cohort of 105 patients [6]. The resulting recurrence rate within 18 months of follow-up is only 15%. STD is a visually discernible AF pattern that guides interventional cardiologists in ablating persistent AF drivers. STD-based ablation uses the PentaRay mapping catheter as shown in Fig. 1(A). Before ablation with RF energy, interventional cardiologists first position sequentially the PentaRay in different sites of the atria. Ten bipolar EGMs are then simultaneously recorded per location by maintaining the catheter stable for a few seconds. Finally, atrial sites displaying an irregular cardiac activity are annotated as dispersion points and tagged for ablation. According to guidelines in [6], dispersion areas refer to clusters of electrograms, either fractionated or not, displaying interelectrode time and space dispersion at a minimum of three contiguous leads [6], as shown in Fig. 1(B). Hence, STD-based ablation is a fully patient-tailored therapy.

However, quantifying visually if the delay of the intracardiac activation through neighboring bipoles can be approximated by 70% of the AFCL is a difficult and time-consuming task. Thousands of atrial locations are mapped with the PentaRay catheter in a typical ablation intervention and all multichannel EGM recordings acquired at these locations must be evaluated visually by the practicing cardiologist in real time. Hence, the visual identification of STD areas is prone to errors and lack of reproducibility. Instead, the present work aims to design an optimized decision-aid solution that helps interventional cardiologists detect STD patterns automatically thanks to ML tools.

3. Methods

3.1. Data acquisition

The process of data acquisition consists of several steps: 1) selecting patients suffering from persistent AF and mapped with PentaRay for STD-based ablation; 2) exporting intervention data from Biosense CARTO® system, including electrocardiogram (ECG), EGM and annotated dispersion points; 3) decompressing, anonymizing and preparing data of interest; 4) structuring data for classification purposes.

During the mapping phase, the PentaRay catheter is maintained stable for 2.5 s at each location. A location refers to an anatomical point inside the heart. Multichannel EGMs are sampled at 1 kHz and displayed through the CARTO system monitor, to be analyzed by the cardiologist. Hence, EGM data exported from the CARTO system can be stored in a matrix of dimensions 10×2500.

3.1.1. Data labeling

Data labeling is performed by interventional cardiologists such that EGMs presenting spatiotemporal dispersion are annotated as “STD”, also called gradient or substrate. We automatically merge the remaining labels into the “non-STD” class. Moreover, a meticulous work is needed for standardizing the labels because the annotation nomenclature exported from CARTO (encoded labels) differ from one patient to another.

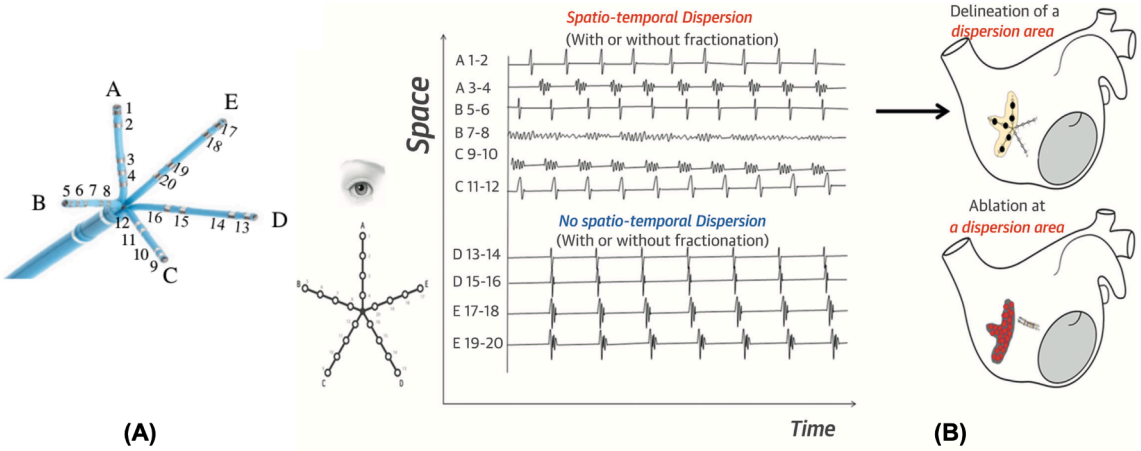

Fig. 1. (A) PentaRay multispline catheter (Biosense Webster, Inc.). (B) Dispersion areas defined and delineated via a mapping approach [6]. Channels A 1–2, A 3–4, B 5–6, B 7–8, C 9–10 and C 11–12 display STD contrarily to the remaining ones.

3.1.2. Circularity transformation

Circularity transformation consists in replicating for each sample the first two channels (matrix rows) at the end of the sample. The new sample, originally with dimensions 10×2500, becomes 12×2500. This transformation allows us to mimic the circular arrangement of the PentaRay branches. Indeed, it captures the neighborhood information between bipoles of splines A and E of the catheter, as shown in Fig. 1(A).

3.2. Data representations

Starting from the basic matrix format presented above, data can be reformatted in different ways to match different classification algorithms:

3.2.1. Matrix

The first classification scheme uses matrices of dimensions 12×2500, the result of applying the circularity transformation on raw recordings of the 10-lead EGMs. Matrices can also be treated as images with 2D CNN models like LeNet-STD, as proposed in [18] for a human activity recognition task.

3.2.2. 2D image

The image format consists in subplotting the curves of the 10 leads, one under the other, as it is currently done by the CARTO system during the mapping phase.This 2D plot of EGMs is used in the hospital and represents the most easily understandable representation by the eye of an interventional cardiologist. The present study will check if this format is also convenient for ML classifiers.

3.2.3. 3D image

This format enables the use of CNN models conceived to train on 3D RGB images like in the ImageNet dataset (natural colored images) [19]. Hence, this data format enables the use of TL [7]. The 3D tensor is the result of stacking three 2D images along the depth dimension. The first slice (image) is a subplot of the ordered leads $\{\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6,\ell_7,\ell_8,\ell_9,\ell_{10}\}$, similarly to 2D image format. The second slice is the subplot of the leads shifted in a circular way. The leads order becomes $\{\ell_2,\ell_3,\ell_4,\ell_5,\ell_6,\ell_7,\ell_8,\ell_9,\ell_{10},\ell_1\}$. The third slice represents a subplot of leads shifted twice as $\{\ell_3,\ell_4,\ell_5,\ell_6,\ell_7,\ell_8,\ell_9,\ell_{10},\ell_1,\ell_2\}$. Similarly to the circularity transformation, the idea comes from the circular structure of the PentaRay branches. The interest in shifting the leads remains in allowing the 3D filter to systematically analyze neighboring leads along the depth dimension as it already does along the spatial dimensions (matrix slice). For instance the first convolutional window captures leads $\{\ell_1,\ell_2,\ell_3\}$ across the three slices. This data format allows us to represent EGMs in a 3D tensor with a depth of three thus mimicking the shape of RGB images and allowing the use of algorithms designed for this input shape.

3.2.4. 1D signal

VAVp times series is a compact representation of multichannel recordings that has shown promising results in [11]. The VAVp distribution is claimed to depend on STD pattern in [6]. As explained in [6], VAVp time series is calculated as follows: 1) the VAV matrix contains in its rows the absolute values of each channel of the multilead EGM recording; 2) one-dimensional VAVp signal is computed as the maximal values at each time sample of the VAV matrix over the leads (rows) dimension. The histograms of VAVp distribution $(h(\mathrm{VAVp}))$ are shown in [6] to significantly depend on the presence of STD patterns. Numerical results showed that $h(\mathrm{VAVp})$ is peaky and concentrated around zero if the virtual PentaRay is positioned in non-STD areas but it gets more spread for EGMs recorded in STD areas, as shown in Fig. 2. However, a recent work [11] showed that the VAVp distribution is not a key feature in STD detection when analyzing the database used in the present study, yet the supervised classification of raw VAVp time series itself into STD vs. non-STD categories is promising. Results on the test set were good with values of accuracy, AUC, sensitivity and specificity around 90% with low variability (10^{-3}). Complementary metrics such as precision and F1 score are also computed and analyzed in the present work. Besides, a larger dataset of over 35000 EGMs are included in the present paper compared to 23000 in [11].

3.3. Classification algorithms

The ML algorithms trained to identify STD locations are the following:

3.3.1. SVM

SVM is widely used in biomedical data analysis like ECG classification [21]. The model maps the input feature space into a higher-dimensional space so that data samples from different classes become separable. Data are divided by a hyperplane in the original n-dimensional space, with a large gap [13]. SVM aims at maximizing the distance, called margin, between data points situated near the hyperplane, called support vectors. Maximizing the margin provides some confidence that new data points can be classified with more accuracy. Then, new samples are mapped into the new space and predicted to belong to a class or another based on the side of the hyperplane on which they fall. Looking for a suitable separable representation of data is called the kernel trick.

3.3.2. PCA

PCA explores linear relationships between features and defines variables significance on the basis of their variance contribution. It projects

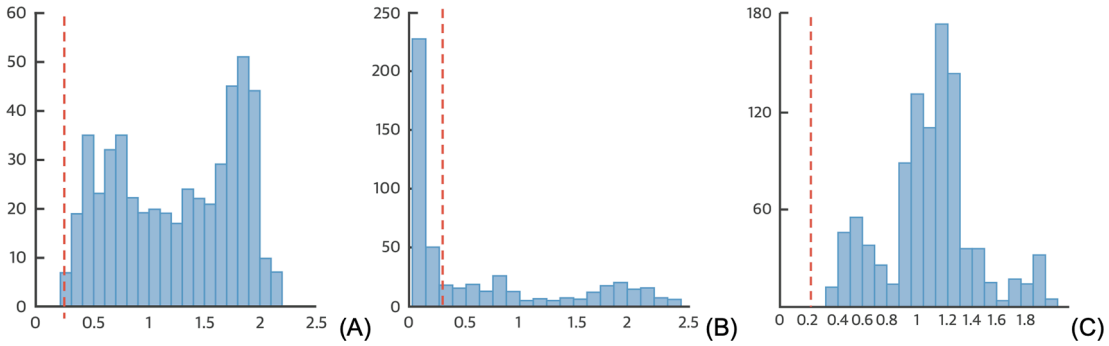

Fig. 2. VAVp distributions (histograms) obtained by positioning a virtual PentaRay catheter in human atria with homogeneous substrate (A) at the center of the driver, reminiscent of patients' dispersion areas; (B) at the periphery of the driver; (C) in the interstitial fibrosis condition [6].

data onto a new space model formed by synthesized variables called principal components (PCs). This is useful when large amounts of information may be approximated by a moderately complex model structure. Based on the estimation of the correlation structure of the variables, PCA evaluates the importance of a variable in the PC model through the size of its residual variance [12]. The choice to be examined is how many PCs (r) adequately account for the total variation in the n-dimensional data samples $\mathbf{x}$. Among the existing criteria for choosing r, we opt for the cumulative percentage of total variation, also called hard threshold (t). The optimal number of PCs is chosen as the smallest integer r for which the cumulative percentage of total variation reaches or exceeds t. PCA allows us to reduce the dimensions of the input space by projecting the input data onto a space formed by the first r PCs only.

3.3.3. MLR

In statistics, logistic regression [2] estimates the probability of a given class or event by using the logistic function. MLR is identical to univariate logistic regression, but it considers more than one covariate. Let $p(\mathbf{x})$ be the model's estimation of the probability of an event that depends on n covariates or independent variables. Then, inverting the logit formulation for modeling the probability gives:

$$\begin{cases} p : \mathbb{R}^n \rightarrow (0,1) \\ p(\mathbf{x}) = \dfrac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n}}. \end{cases}$$

The estimates of the unknown parameters $\beta_i, i = 0, 1, 2, \ldots, n$, are derived through the maximization of a likelihood function. MLR is a baseline classifier in biomedical data analysis and classification [2]. Furthermore, MLR can be implemented as a fully connected (FC) neural network.

3.3.4. CNN

Artificial neural networks (NN) are commonly used in biomedical data classification [2,22,23]. A NN consists of a series of connected layers, each layer composed of a number of artificial neurons, called nodes, that have learnable weights. A CNN is a special NN composed of convolutional and pooling layers followed by FC ones. Each node of the convolutional (conv) layer receives some inputs, performs a dot product and optionally follows it with a non-linearity. The output of each conv filter is called feature map. In pooling layers, the resolution of feature maps is reduced to increase the invariance of features to distortions. Two CNN architectures are considered in this study:

(a) **LeNet-STD** architecture is inspired by the state-of-the-art LeNet5, a baseline CNN that is both a good classifier and a computationally affordable algorithm [14]. It has 60,000 parameters compared to AlexNet for instance that has 60 million parameters [7].

(b) **VGG16** is a deep CNN designed by the Visual Geometry Group Lab, trained for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and achieving 92.7% top-5 test accuracy. The ImageNet dataset contains 14 million images that belong to 1000 different classes. VGG16 is composed of 5 conv blocks alternated with pooling layers and followed by FC ones. In total VGG16 has 16 layers with trainable weights [15].

3.3.5. TL

In terms of ML, TL is a process where a NN, say M_1, is trained on a first problem then re-used in some way in a related problem. A new model M_2 is then formed by aggregating one or several layers from the trained model M_1 to additional layers, generally FC layers with trainable weights. The main benefits of TL consist in decreasing the training time for the new model and achieving a lower generalization error [7].

3.4. Cross-validation

Cross-validation (CV) is a statistical tool commonly used in ML to quantify the generalization power of a classifier [24]. k-fold CV consists in: first, partitioning the entire datasets into k subsets called folds; second, repeating model training k times while considering, at each round, the k^{th} fold as the test dataset and the remaining samples as the training dataset; finally, the estimation of the classifier's performance is given by averaging the test results over the k rounds. A rule of thumb is to choose k equal to 5 or 10 [25]. Regarding the size of our dataset and the small number of STD samples (only 5% of non-STD samples), we opt for 5-fold CV in order to keep an acceptable amount of STD test samples. In each CV round, the test dataset is partitioned into two equal-sized subsets that will form the new validation and test sets. This guarantees that the model does not see the test samples during the training phase.

3.5. Performance metrics

A cost function is minimized during the learning phase of a classifier. In order to evaluate the model's classification performance, several metrics are assessed. The accuracy (Acc) metric is computed as the result of dividing the total number of correct predictions by the total number of predictions. However, accuracy is not enough to quantify classification performance when learning from imbalanced datasets. For this, it remains important to compute the confusion matrix whose elements are the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). From these values, the following metrics are evaluated: 1) sensitivity, recall or true positive rate (TPR), measuring the proportion of actual STD samples that are correctly identified as such; 2) specificity or true negative rate (TNR), quantifying the proportion of actual non-STD samples that are correctly identified as such; 3) precision or positive predictive value (PPV), assessing how trustable the result is

when the model answers that a point belongs to the STD class; 4) negative predictive value (NPV), or probability that a sample is actually non-STD when the model has classified it as such; and 5) F1-score which stands for the harmonic mean of precision and sensitivity. The F1 score quantifies the balance between precision and recall values: high recall accompanied by high precision values reflect that the STD class is perfectly handled by the model, low recall accompanied by high precision values reflect that the model cannot detect the STD class well but is sufficiently trustable when it does; high recall accompanied by low precision values reflect that the STD class is well detected but the model output for that class also includes non-STD samples, and low recall accompanied by low precision values reflect the STD class is poorly handled by the model. Then, the area under the Receiver Operating Characteristic (ROC) curve, shortly named AUC, is computed. The ROC curve is a graphical tool widely used to evaluate the performance of a binary classifier when varying the discrimination threshold [20].

3.6. Data augmentation

One of the biggest and most frequent challenges encountered in deep learning remains the insufficient amount of data or the uneven class balance of samples within the training dataset. The high misrepresentation of STD samples in the multichannel EGM dataset leads to poor classification results in terms of sensitivity, precision and AUC. This data imbalance issue can be handled through adapted DA solutions. DA is a signal preprocessing tool that applies transformations to original samples of the minority STD class in order to synthesize new samples. It consists in forming a balanced super-dataset by replicating randomly STD samples until they reach the number of non-STD ones [9,10]. A recent study benchmarked a set of DA solutions for the classification of the current STD multichannel EGM dataset [8]. The methods explored included random oversampling (ROS), undersampling, leads shift, time shift and time reversing. These methods provide different class imbalance ratios (Tab. I in [8]).

These methods were designed to preserve the integrity of STD patterns and were approved by partner interventional cardiologists. The classification performance of both a shallow CNN and MLR demonstrated that ROS is the best technique. Indeed, ROS increased the sensitivity value by 30% compared to learning from raw data while maintaining high values of specificity and AUC around 90%. For this reason, we opt for ROS for the remaining data formats considered in the present study.

4. EGM classification models

The association between the different data formats and classifications algorithms is schematized in Fig. 3. An EGM classification model corresponds to the coupling between a data representation and a suitable classifier, as summarized next.

4.1. Matrix classification

MLR and a shallow CNN, called LeNet-STD, are used for the classification of EGM matrices. LeNet-STD is chosen as a proof of concept that 2D CNN architectures are adapted to STD detection. We study the effect of the receptive field of the network's first conv layer, denoted f_{size}. We recall that the STD pattern can be detected on a minimum of three bipoles positioned on two adjacent PentaRay branches and the average AFCL value is typically 200 ms. Hence, a natural choice is $f_{\text{size}} = 4 \times (\alpha_1 \cdot \text{AFCL})$, with $\alpha_1 \in \mathbb{Q}, \alpha_1 \geqslant 1$. The value $\alpha_1 = \frac{3}{2}$ is proposed in [8], thus capturing cardiac activations along a minimum of one and a half AFCL, and yielding $f_{\text{size}} = 4 \times 300$. The corresponding model is denoted LeNet-STD$_{(4\times300)}$. We investigate further settings such as $f_{\text{size}} = 4 \times 3$ and $f_{\text{size}} = 4 \times 200$, whose models are denoted LeNet-STD$_{(4\times3)}$ and LeNet-STD$_{(4\times200)}$, respectively. An f_{size} of 4×200 is conceived to capture exactly an entire AFCL.

4.2. 2D image classification

A classical window size for image classification is 3×3 [15]. Indeed, $f_{\text{size}} = 3 \times 3$ is the smallest receptive field to capture the patterns of left/right, up/down and center, as explained in [15]. Hence, MLR and LeNet-STD$_{(3\times3)}$ are used for EGM-plot image classification.

4.3. 3D image classification

We perform the classification of 3D images with a deep CNN called VGG16-EGM. VGG16-EGM is the result of transferring the VGG16 architecture. Indeed, we re-use the first trained conv blocks of VGG16 model. The transferred layers serve as a feature extractor. Their output is fed to three consecutive FC layers with 1024, 128 and 2 nodes, respectively. Only FC layers have learnable weights. Hence, VGG16-EGM is

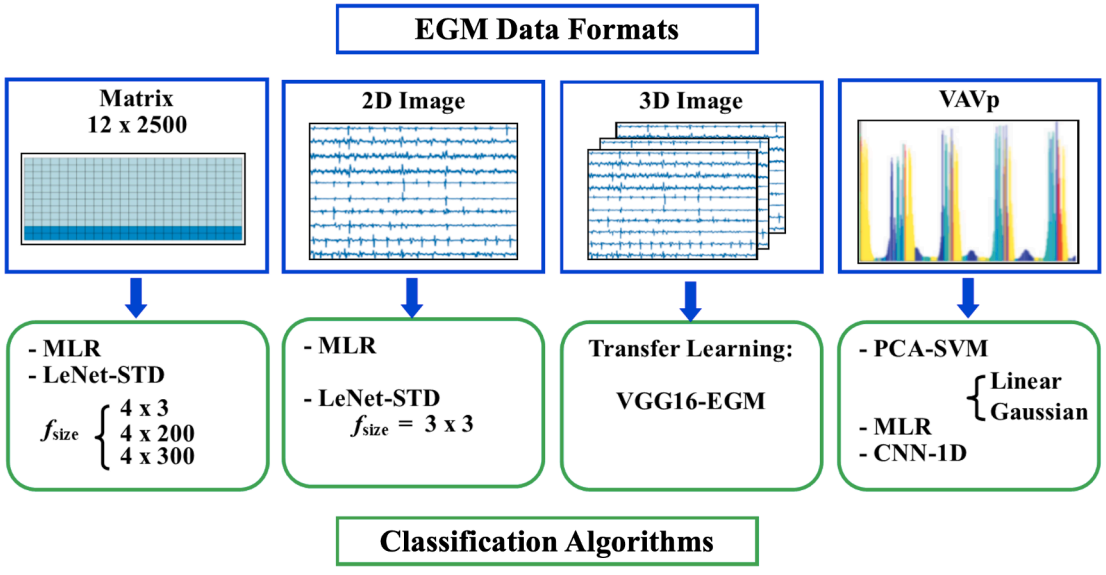

Fig. 3. Classification algorithms adapted to the different EGM data formats considered in this study. An EGM classification model is composed of the combination of an EGM data format and a suitable classifier.

adapted to 3D image classification and can be trained on the EGM dataset.

4.4. VAVp classification

In addition to MLR, SVM is another baseline linear classifier in ML, as recalled in Sec. 3.3.1. However, classifying VAVp time series was not possible with SVM because this input has 2500 covariates and SVM is not adapted to such a large amount of variables. We propose to reduce the dimensions of VAVp signals with PCA, as explained in Section 3.3.2. We use the hard threshold technique at 95% of explained variance to decide the number r of PCs to keep. Then, we feed the reduced version of VAVp to SVM. Linear and Gaussian kernels are benchmarked. The models are denoted respectively PCA-SVM$_{\text{Lin}}$ and PCA-SVM$_{\text{Gauss}}$. We also perform classification with a shallow 1D CNN (CNN-1D).

Implementation: ML experiments are conducted within the Docker environment, running on Quadro P6000 GPU and Ubuntu 19.10 operating system using Python 3.6.9 programming language, Tensorflow 2.1.0 deep learning library for GPU and Keras API.

5. Results

5.1. Study dataset

CARTO data of 15 patients are exported in the present study from the database of the Cardiology Department of Nice University Hospital Center (CHU Pasteur), whose Ethics Committee approved the proposed research. Cartographies of both right and left atria are merged. The study dataset includes a cohort of 35563 10-channel EGM signals of length 2.5 s acquired from both right (RA) and left atria (LA). The recordings include 1804 STD and 33759 non-STD samples. The baseline information of the study patients is given in Table 1. The population is aged 64 years on average and is composed of 80% male and 20% female patients. The average initial AF cycle in left atrial appendage (LAA) is 156 ms. On average, the ablation intervention lasts 2 h 36 min and requires 155 RF shots for a total RF delivery of 53 min 57 s. The class imbalance ratio (CIR) of our dataset is given by:

$$\text{CIR} = \frac{\#\text{STD}}{\#\text{non-STD}} \approx 5\%$$

where # refers to the number of samples. Based on our finding in the comparative study [8] and due to the extremely low value of CIR, we opt for ROS to handle the lack of a sufficient amount of STD samples, as justified in Section 3.6.

5.2. Classification results

The identification of STD locations is performed with several ML techniques including TL. For this purpose, we benchmark the different EGM classification models detailed in Section 4. Their performance on test sets are given in Table 2, where values are calculated as the average over 5-fold CV, as explained in Section 3.4. All standard deviations are inferior to 10^{-2}. Hence, all trained models have a low generalization error. The results of the different EGM classification models can be summarized as follows.

Table 1
Baseline information about patient's population.

Feature	Value (mean ± std)
Gender	12 Male + 3 Female
Age (year)	64 ± 12
Initial AF cycle in LAA (ms)	156 ± 12
Procedure duration	2 h 36 min ± 42 min
RF duration	53 min 57 s ± 12 min 15 s
Number of RF shots	155 ± 61

5.2.1. Matrix classification

Both LeNet-STD with its three settings of f_{size} and MLR yield good and comparable performance. However, LeNet-STD slightly outperforms MLR, as the latter obtains values of AUC and F1 equal to 92% and 53% respectively, whereas the average values of AUC and F1 are 94% and 58% for LeNet-STD. The choice of f_{size} in LeNet-STD has a small impact on the classification performance but has an important effect on the computational cost. Indeed, F1 scores are 58%, 59% and 60%, respectively, with LeNet-STD$_{(4\times3)}$, LeNet-STD$_{(4\times200)}$ and LeNet-STD$_{(4\times300)}$ respectively. However, the training times of LeNet-STD$_{(4\times300)}$ and LeNet-STD$_{(4\times200)}$ are very high compared to LeNet-STD$_{(4\times3)}$ as shown in Table 3.

5.2.2. 2D image classification

Second, we check if classifying 2D images would give good performance because this is the most intuitive input format. In practice, interventional cardiologist analyze visually EGM subplots to detect STD pattern, so this format seems particularly suitable in this context. However, results in Table 2 show no significant enhancement in performance with 2D images compared to EGM matrices. Here again, we notice that LeNet-STD yields a better performance, with an F1 score of nearly 60%, comparing favorably to the 52% provided by MLR.

5.2.3. 3D image classification

Applying TL of VGG16 model to 3D images is performed as a proof of concept. VGG16-EGM yields good classification performance with values of Acc, AUC and F1 of 93%, 93% and 55%, respectively. This shows that extracting features from 3D plots with the use of conv blocks, trained on a different application (natural images), does work. However, besides the heavy computational cost of VGG16-EGM, as highlighted in Table 3, synthesizing 3D images reveals heavy too. Several extensions were benchmarked for storing and reading images in an optimal way.

5.2.4. VAVp classification

The overall classification performance of VAVp time series is also good but less effective than the remaining data representations. For instance, the values of F1 score do not exceed 50%. Table 2 shows that the performance of CNN-1D is slightly better than that of MLR. MLR beats significantly PCA-SVM, mainly in terms of F1 score. Moreover, the Gaussian kernel outperforms the linear one. The VAVp format does not yield the best value in any of the performance metrics considered in this evaluation and results in the lowest F1 scores, all below 46%.

6. Discussion

Automatic classification of STD EGMs is possible using suitable ML techniques. Based on the results reported in this work, we can claim that classifying raw EGMs stored in matrices with LeNet-STD$_{(4\times3)}$ yields a good balance between performance and computational cost. If we observe the values of precision and F1 across all experiments (Table 2), we notice poor values (30% with VAVp and inferior to 60% with the remaining features) even though values of TPR are high around 80% and 90% (except with PCA-SVM). However, values of PPV and F1 are high for the training set that is balanced with ROS. This low precision on the test sets can be explained by the fact that both test and validation sets remain highly imbalanced. Hence, even though the trained models can identify better STD samples when the training set is augmented, as highlighted in [8], information about STD class is still less rich than that of non-STD due the redundancy introduced by ROS. A low precision also indicates that the number of FP is important compared to that of TP. This can be acceptable in biomedical data analysis. On the one hand, TPR is high which means that true STDs are well detected. On the other hand, it is always preferable to detect non-STDs as STDs than conversely. This allows the model to tag these FP predictions during the mapping phase

Table 2
Average classification performance on test set through 5-fold CV.

Format	Classifier	Test classification performance						
		Acc	TPR	TNR	AUC	PPV	NPV	F1
Matrix	MLR	0.924	0.902	0.847	0.929	0.389	0.991	0.533
	LeNet-STD$_{(4\times3)}$	0.936	0.911	0.862	0.94	0.436	0.992	0.579
	LeNet-STD$_{(4\times200)}$	0.937	**0.929**	0.881	0.940	0.443	**0.993**	0.589
	LeNet-STD$_{(4\times300)}$	**0.940**	0.928	0.879	**0.943**	**0.453**	0.99	**0.598**
2D image	MLR	0.873	0.913	0.852	0.874	0.398	0.991	0.517
	LeNet-STD$_{(3\times3)}$	0.939	0.911	0.880	**0.943**	0.452	**0.993**	0.597
3D image	VGG16-EGM	0.926	0.880	**0.928**	0.925	0.406	0.992	0.553
VAVp	PCA-SVM$_{\text{Lin}}$	0.791	0.670	0.535	0.805	0.128	0.970	0.207
	PCA-SVM$_{\text{Gaus}}$	0.837	0.732	0.616	0.849	0.179	0.976	0.277
	MLR	0.888	0.891	0.807	0.892	0.286	0.988	0.423
	CNN-1D	0.902	0.890	0.817	0.907	0.321	0.989	0.460

Table 3
Computational cost of training on balanced dataset and predicting the label of a test data sample. Time is given as hours(hh):minutes(mm):seconds(ss). Symbol $\text{nb}_{\text{tr}-\text{par}}$ stands for the number trainable parameters, which might be inferior to the total number of parameters in case of TL when used as a feature extractor. For VAVp, the number of variables is $n = 2500$ and $r \approx 300$. Symbol m represents the number of training samples (80% of the total number of samples in 5-fold CV).

Format	Classifier	Training cost		Prediction time (s)
		time	$\text{nb}_{\text{tr}-\text{par}}$	
Matrix	MLR	00:01:49	400,672	0.139
	LeNet-STD$_{(4\times3)}$	00:11:04	1,289,458	0.643
	LeNet-STD$_{(4\times200)}$	00:24:57	1,214,322	0.625
	LeNet-STD$_{(4\times300)}$	00:37:51	1,175,922	0.616
2D image	MLR	01:08:57	400,672	0.144
	LeNet-STD$_{(3\times3)}$	01:10:15	1,289,362	0.618
3D image	VGG16-EGM	14:35:28	100,795,778	0.944
VAVp	PCA-SVM$_{\text{Lin}}$	00:10:14	$O(m^2n)$	0.003
	PCA-SVM$_{\text{Gaus}}$	00:55:32	$O(m^3r)$	0.002
	MLR	00:12:39	400,672	0.141
	CNN-1D	00:11:25	80,066	0.585

and give the cardiologist the opportunity to analyze them later, before ablation. Also, Table 3 shows that our solution can be deployed in real time since the prediction time (inference) of a new data sample can be performed in less than a second (maximal prediction time equal to 0.944 s for VGG16-EGM). If more efficient computational resources are deployed, prediction time can be further decreased. The overall performance achieved with our proposed solutions is good with values of Acc and AUC around 90%, but better results are expected in a medical decision-aid tool. An accuracy of 95% should be attained for the solution to be considered as sufficiently reliable for clinical use. Also, a higher number of patients would provide a richer EGM database with increased variability across patients.

7. Conclusions and perspectives

Automatic detection of atrial areas presenting STD pattern is a valuable decision-aid tool that can help interventional cardiologists identify potential ablation sites for treating persistent AF in a faster and more reliable manner than the current visual inspection. To identify STD, several features are extracted from EGM recordings either automatically with TL and end-to-end NN training or in a hand-crafted fashion with VAVp time series. These different data representations are classified with the use of adapted ML algorithms giving rise to a variety of EGM classification models. Moreover, we study the effect of some hyperparameters and settings like the choice of receptive fields in CNN and kernels in SVM. Based on the analysis of both the classification results on the test set and the computational cost of the different classification models, the best performance is achieved with LeNet-STD and $f_{\text{size}} = 4 \times 3$ for classifying raw EGM matrices. The average performance over 5-fold CV reaches 94% of accuracy and AUC added to an F1-score of 60%. The low precision and F1-score can be explained by the insufficient amount of STD samples even though the class imbalance issue is alleviated with DA during the training phase. On the other hand, VGG16-EGM demonstrates that extracting features with conv blocks of the VGG16 model, trained on natural images, works as well as shallower architectures like LeNet-STD, but is computationally very expensive. By providing interventional cardiologists with a real-time objective measure of STD, the proposed solution offers the potential to improve the efficiency and effectiveness of this fully patient-tailored catheter ablation approach for treating persistent AF. Future work should investigate alternative ML models like long short-term memory (LSTM) neural networks, as they are well adapted to time series classification [26].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Yao, R. Wang, X. Fan, J. Liu, Y. Li, Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network, Information Fusion 53 (2020) 174–182.
[2] S. Dreiseitl, O.M. Lucila, Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics 35 (5–6) (2002) 352–359.
[3] V. Zarzoso, D.G. Latcu, A.R. Hidalgo-Muñoz, M. Meo, O. Meste, I. Popescu, N. Saoudi, Non-invasive prediction of catheter ablation outcome in persistent atrial fibrillation by fibrillatory wave amplitude computation in multiple electrocardiogram leads, Archives of Cardiovascular Diseases 109(12) (2016) 679–688.
[4] C.T. January, L.S. Wann, H. Calkins, L.Y. Chen, J.E. Cigarroa, J.C. Cleveland, et al., 2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society, Journal of the American College of Cardiology 74 (1) (2019) 104–132.

[5] A. Verma, C.Y. Jiang, T.R. Betts, et al., Approaches to catheter ablation for persistent atrial fibrillation, New England Journal of Medicine 372 (19) (2015) 1812–1822.
[6] J. Seitz, C. Bars, G. Théodore, et al., AF ablation guided by spatiotemporal electrogram dispersion without pulmonary vein isolation: a wholly patient-tailored approach, Journal of the American College of Cardiology 69 (3) (2017) 303–321.
[7] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press (2016) www.deeplearningbook.org.
[8] A. Ghrissi, D. Almonfrey, R. Almeida, F. Squara, V. Zarzoso, J. Montagnat, Data augmentation for automatic identification of spatiotemporal dispersion electrograms in atrial fibrillation ablation using machine learning, in: Proc. 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Montreal, Canada, July 2020.
[9] M. Agnieszka, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: Proc. IEEE International Interdisciplinary PhD Workshop, May 2018.
[10] C. Shorten, L.K. Taghi, A survey on image data augmentation for deep learning, Journal of Big Data 6 (1) (2019) 60.
[11] A. Ghrissi, F. Squara, V. Zarzoso, J. Montagnat, Identification of spatiotemporal dispersion electrograms in persistent atrial fibrillation ablation using maximal voltage absolute values, in: Proc. 28th European Signal Processing Conference, Amsterdam, The Netherlands, Jan. 2021.
[12] I.T. Jolliffe, C. Jorge, Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2065) (2016), 2015–0202.
[13] B.E Boser, M.G. Isabelle, N.V. Vladimir, A training algorithm for optimal margin classifiers, in: Proc. 5th Annual Workshop on Computational Learning Theory, USA, 1992, pp. 144–152.
[14] Y. LeCun, L. Bottou, Y. Bengio, et al., Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. International Conference on Learning Representations, San Diego, CA, USA, May 2015.
[16] J. Malmivuo, R. Plonsey, Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields, Oxford University Press, NY, USA, 1995.
[17] K. Nademanee, J. McKenzie, E. Kosarand, et al., A new approach for catheter ablation of atrial fibrillation: mapping of the electrophysiologic substrate, Journal of the American College of Cardiology 43 (11) (2004) 2044–2053.
[18] J. Yang, M.N. Nguyen, P.P. San, X. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, in: Proc. 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, July 2015.
[19] J. Deng, W. Dong, R. Socher et al., ImageNet: A large-scale hierarchical image database, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, June 2009, pp. 248–255.
[20] K. Hajian-Tilaki, Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, Caspian Journal of Internal Medicine 4 (2) (2013) 627.
[21] F. Melgani, B. Yakoub, Classification of electrocardiogram signals with support vector machines and particle swarm optimization, IEEE Transactions on Information Technology in Biomedicine 12 (5) (2008) 667–677.
[22] B. Pyakillya, N. Kazachenko, N. Mikhailovsky, Deep learning for ECG classification, Journal of Physics: Conference Series 913 (1) (2017). IOP Publishing.
[23] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, et al., Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nature Medicine 25 (1) (2019) 65.
[24] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proc. 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, Aug. 1995.
[25] M. Kuhn, J. Kjell, Applied Predictive Modeling, Springer 26 (2013) 70.
[26] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, et al., Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nature Medicine 25 (2019) 65–69.