Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Atrial fibrillation signatures on intracardiac electrograms identified by deep learning

Miguel Rodrigo [a,b,*], Mahmood I. Alhusseini [a], Albert J. Rogers [a], Chayakrit Krittanawong [c], Sumiran Thakur [a], Ruibin Feng [a], Prasanth Ganesan [a], Sanjiv M. Narayan [a,**]

[a] Cardiovascular Division and Cardiovascular Institute, Stanford University, CA, USA
[b] CoMMLab and Electronic Engineering Department, Universitat de Valencia, VA, Spain
[c] Baylor College of Medicine, TX, USA

ARTICLE INFO

This work was presented at the 2021 Young Investigator Finals of the American College of Cardiology Scientific Sessions.

Keywords:
Atrial fibrillation
Arrhythmia
Machine learning
Implanted devices
Wearables
Electrophysiology
Biomedical signal processing
Cardiology

ABSTRACT

Background: Automatic detection of atrial fibrillation (AF) by cardiac devices is increasingly common yet sub-optimally groups AF, flutter or tachycardia (AT) together as 'high rate events'. This may delay or misdirect therapy.
Objective: We hypothesized that deep learning (DL) can accurately classify AF from AT by revealing electrogram (EGM) signatures.
Methods: We studied 86 patients in whom the diagnosis of AF or AT was established at electrophysiological study (25 female, 65 ± 11 years). Custom DL architectures were trained to identify AF using $N = 29{,}340$ unipolar and $N = 23{,}760$ bipolar EGM segments. We compared DL to traditional classifiers based on rate or regularity. We explained DL using computer models to assess the impact of controlled variations in shape, rate and timing on AF/AT classification in 246,067 EGMs reconstructed from clinical data.
Results: DL identified AF with AUC of 0.97 ± 0.04 (unipolar) and 0.92 ± 0.09 (bipolar). Rule-based classifiers misclassified ~10–12% of cases. DL classification was explained by regularity in EGM shape (13%) or timing (26%), and rate (60%; $p < 0.001$), and also by a set of unipolar EGM shapes that classified as AF independent of rate or regularity. Overall, the optimal AF 'fingerprint' comprised these specific EGM shapes, >15% timing variation, <0.48 correlation in beat-to-beat EGM shapes and $CL < 190$ ms ($p < 0.001$).
Conclusions: Deep learning of intracardiac EGMs can identify AF or AT via signatures of rate, regularity in timing or shape, and specific EGM shapes. Future work should examine if these signatures differ between different clinical subpopulations with AF.

Financial support

This work was supported in part by grants from the Spanish Ministry of Economy and Competitiveness (PID2020-114291RB-I00) and Generalitat Valenciana Grants (APOSTD/2017) and projects (GVA/2018/103; AICO/2021/318) to MRB, from the National Institutes of Health F32 HL144101 to AJR, and R01 HL149134 (Machine Learning in Atrial Fibrillation) and HL R01 HL83359 (Dynamics of Human Atrial Fibrillation) to SMN.

1. Introduction

Accurately identifying Atrial Fibrillation (AF) in tracings from wearable or cardiac implanted electronic devices (CIEDs) is increasingly central to patient care, and may guide ablation, choice of medications or anticoagulation therapy [1]. Nevertheless, automatic device detection of AF is suboptimal. Wearable devices [2,3] and CIEDs including pacemakers and defibrillators [1] typically detect AF, atrial flutter or tachycardia (AT) by rate or regularity. This often classifies organized tachycardias or even premature atrial ectopic beats as AF [4], leading to diagnostic errors or delay of definitive therapy [5,6].

We hypothesized that deep machine learning (DL) can integrate

* Corresponding author. Cardiovascular Division and Cardiovascular Institute, Stanford University, CA, USA.
** Corresponding author.
E-mail addresses: miguel.rodrigo@uv.es (M. Rodrigo), sanjiv1@stanford.edu (S.M. Narayan).

https://doi.org/10.1016/j.compbiomed.2022.105451
Received 2 September 2021; Received in revised form 23 March 2022; Accepted 23 March 2022

features from atrial electrogram (EGMs) to detect AF better than traditional approaches. DL is a provocative and rapidly developing branch of computer science which can reveal unrecognized structures in complex data [7,8], without the limitations of detailed expert rules. While DL has been applied to the ECG to identify AF [9,10], it has rarely been applied to separate AF from AT from intracardiac data. We set out to develop DL to distinguish AF from organized AT in intracardiac EGMs, uniquely validated at electrophysiological study, and compared DL analysis to expert rules.

Notably, we also set out to address the 'black box' limitation of DL, because uncertainty in how DL achieves classification [7,8] reduces confidence in its clinical use. We further hypothesized that explainability analyses could identify which clinically meaningful features such as EGM waveform shape are used by DL to classify AF. We reasoned that such 'AF signatures' may indicate clinical or physiological features that could ultimately be used to personalize therapy.

2. Materials and methods

2.1. Patient population

We studied patients in the COMPARE registry (NCT02997254) of AF patients who were enrolled prospectively at ablation for symptomatic AF, refractory to at least 1 anti-arrhythmic medication. Each patient in this registry had intracardiac EGMs recorded by multipolar 64 pole basket catheters. The registry was reviewed by a panel of 3 cardiac electrophysiologists who classified each tracing as AF or AT. For the present study, we selected consecutive patients from this registry to construct a balanced dataset of intracardiac recordings with AF (N = 43) or AT (N = 43). Each patient provided written informed consent under protocols approved by the Human Research Protection Program.

2.2. Electrogram collection and export

Electrophysiology study was performed after discontinuing antiarrhythmic medications for 5 half-lives. A 64-pole basket catheter (Abbott, Menlo Park, CA; electrode size 2 mm, inter-electrode spacing 5 mm along spline) was advanced to map the right and left atria. Catheters were maneuvered by experienced operators to optimize contact [11]. We exported 60 s of unipolar electrograms from the electrophysiological recorder (Prucka, GE Marquette, Milwaukee, WI; Bard Electrophysiology, Billerica, MA), filtered at 0.05–500 Hz. Unipolar electrograms were analyzed for durations of 4000 ms which provide ~20 cycles of AF or AT. This is a common duration for EGM sequences analyzed in the frequency domain, and longer durations may not improve rhythm identification. Original EGMs had sample Frequency (Fs) of 1 kHz (Bard) or 977 Hz (Prucka), and were resampled to compare analyses between datasets. To reduce dimensionality and, since the physiological content of AF and AT EGMs is < 200 Hz [12], we downsampled EGMs to 400 Hz with a 200 Hz anti-aliasing filter. Our results should thus be applicable to any system with Fs > 400 Hz. Ventricular artifacts were eliminated by subtracting a mean QRS complex, obtained by identifying each QRS in 3 orthogonal ECG leads by a voltage threshold and averaging them across 1 min [13]. Bipolar EGMs were constructed by subtracting unipolar signals at adjacent electrodes across each catheter spline. Examples of unipolar and bipolar EGMs are presented in Fig. 1A.

Stratified Monte Carlo cross-validation was accomplished by randomly assigning 20% (8 or 9 patients from each group) to the validation sample and the remainder to the training sample. The corresponding datasets of 29,340 unipolar and 23,760 bipolar EGM signals (4-s length) were evaluated in the validation set. This process was repeated 10 times and the results averaged across validation sets. This approach has been used by Feeny et al. to predict response to cardiac resynchronization and by ourselves to predict sudden death from

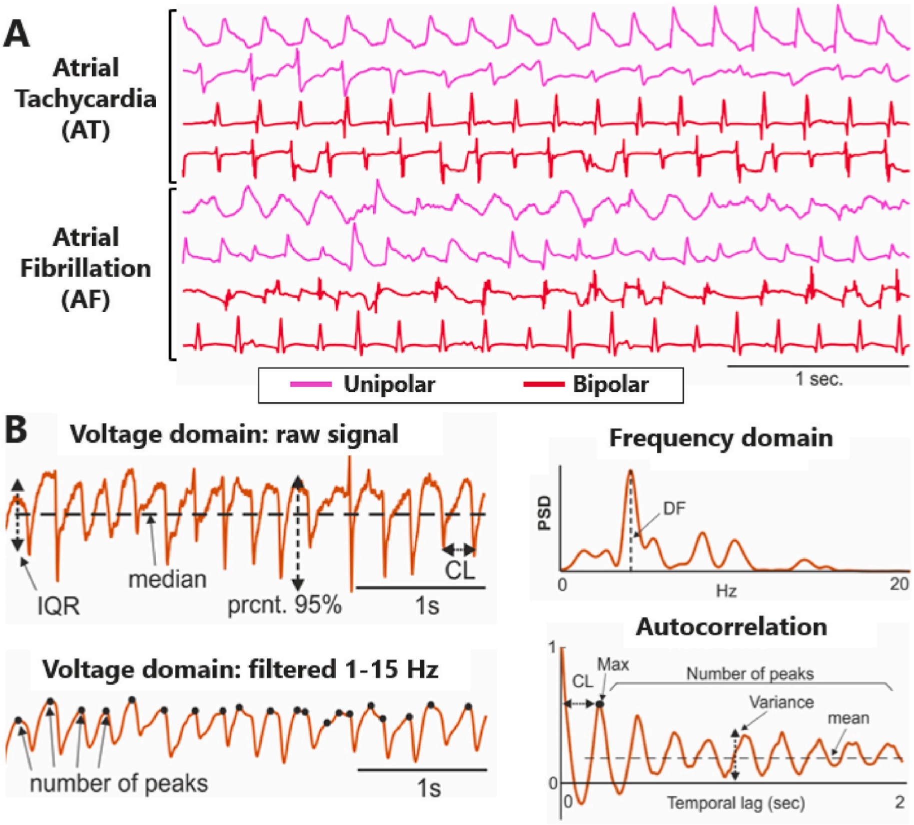

Fig. 1. Atrial Electrograms and feature extraction. A. Unipolar and bipolar EGMs from the Atrial Tachycardia (AT) and Atrial Fibrillation (AF) groups. B. Examples of traditional features extracted from atrial EGMs.

intracardiac ventricular signals [14,15].

2.3. Traditional features to identify AF

Clinicians use several EGM rules to identify AF from other arrhythmias, primarily a higher rate in AF which can be measured as the number of beats per second, the dominant frequency (DF) or its inverse (cycle length, CL). AF typically has CL < 200 ms (DF > 5 Hz), while AT has CL ≥ 200 ms. AF also exhibits a higher beat-to-beat variation in rate (Fig. 1B), which can be quantified by the standard deviation of maxima in unit time. AF exhibits beat-to-beat variations in EGM shape, unlike AT, although this is rarely quantified. Moreover, AF cases arise with rapid activation but similar EGM shapes. We quantified consistency of EGM shape using autocorrelation of successive electrograms (Fig. 1B). In total, we extracted 45 EGM features of morphology, amplitude, timing, frequency and autocorrelation to separate AF from AT (Supplementary Table 1), and compared different configurations of feature-extraction algorithms to reduce dependence of results on any one algorithm.

2.4. Statistical and classic machine learning of traditional AF features

Feature-based classification of AF versus AT was performed using well-reported techniques widely used for detecting AF from the ECG [16]. First, we used individual features, and optimized a binary threshold to predict AF for each. We then combined multiple features. Four feature-based, trainable and well-known classifiers were used (Supplementary Figure 1A and Supplementary Methods):

- Linear Regression. This was constructed by combining all parameters into one linear function ($y = g(a_0 + a_1 f_1 + a_2 f_2 + \ldots + a_N f_N)$), a_i representing the coefficient, f_i each feature and $g(\cdot)$ the canonical function (binomial). The linear model was trained using Poisson regression using function *fitglm* from Matlab® (Mathworks, Natick, MA).
- Bagged Trees (Forest). In this approach, the averaged output of N decision trees is provided, based on binary thresholding of individual parameters. An ensemble of 200 decision trees were trained using function *TreeBagger* from Matlab®.
- K-Nearest Neighbor (KNN). The predicted output is calculated through the K nearest neighbors in the domain of size N, where N is the number of features. Neighbors are points/features combinations used for training. K = 20 neighbors were considered in our analysis using function *fitcknn* from Matlab®.
- Support Vector Machine (SVM). The output is predicted in a low- or moderate-dimensional predictor data set by identifying a subset of inputs, termed support vectors, that form a decision boundary whose separation increases in training. The function *fitcsvm* from Matlab® was used.

The predictive value of each feature was calculated by its Area Under the Curve (AUC) to classify AF. Features were combined by sequential inclusion from highest AUC, in descending order until classification accuracy in the validation cohort reached a plateau. Because some features are correlated, we excluded those with correlation >0.9 against any feature already included in the model.

2.5. Deep learning

We applied two customized DL architectures, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to raw EGM signals in an end-to-end fashion. The CNN comprised two 1-dimensional convolutional layers, of 32 and 8 elements-length each, and 2 fully connected dense layers. All layers comprised 256 filters and had dropout values of 0.3, 0.2, 0.1 and 0.0, respectively (Supplemental Fig. 1B). The RNN comprised by one Long-Short Term Bidirectional Memory layer (size 256, dropout 0.3) and a fully connected layer (2 filters, dropout 0.3). These designs are similar to models we have reported [15,16], although all models were trained from scratch (transfer-learning was not used). Details of Deep Learning models and training can be found in the Supplemental Material.

2.6. Explainability analysis of DL to identify electrogram signatures

To study the rationale for DL classification, we created a database of atrial EGMs in which we systematically modified each clinically-intuitive EGM feature one at a time, leaving others fixed. EGM sequences were reconstructed from actual patient-EGMs in the validation cohort. This enabled us to dissect the impact of separate features of atrial EGMs on classification [17].

Fig. 2 summarizes generation of this reconstructed database. We first randomly selected 415 unipolar signals of AF and AT (N = 207 and N = 208 respectively) from all 10 validation cohorts. Activation times were assigned at the maximum absolute first derivative of the EGM. Fiducial points f_1 and f_2 were defined at 35% and 65% of the cycle length between EGM activations respectively.

To introduce EGM shape variations (Fig. 2A), the trace of a randomly identified beat was copied onto different randomly identified beats (red trace, 2.A). This allowed us to change the shape of a range of beats, from one to all. To evaluate the effect of shape irregularity, the average cross correlation between the final beats was reported.

To introduce EGM timing variations (Fig. 2B), the EGM between f_2 and the next f_1 fiducials were shifted by a random percentage of cycle length from $-p$ to $+ p$, p ranging from 0% to 35%. The EGM was reconstructed by fitting a ramp function. Timing irregularity was quantified as the deviation of beat-to-beat cycle length (CL) from overall cycle length, as a percentage of the signal CL.

Finally, EGM rate shift was generated by adding or removing atrial beats and linearly redistributing remaining activations (Fig. 2C). The EGM was then reconstructed by fitting ramp functions. Atrial beats were added to a shortest CL 100 ms, and removal was performed to CL 350 ms that represents the slowest rate during AT. Rate of the reconstructed signal was quantified as its average CL.

N = 246,067 EGM sequences were reconstructed. First, EGM sequences were generated by varying each of rate, shape and timing individually, fixing the other two (N = 8611). Second, EGM signals were generated by varying shape and timing irregularity while keeping rate fixed (N = 29,190), and varying rate and timing irregularity while keeping EGM shape fixed (N = 75,166). A final database was reconstructed in which signals had combined variations of the 3 parameters (N = 133,100). Reconstructed EGMs were used as inputs to DL to evaluate the impact of altering each parameter on AF classification in validation sets not used for DL training.

Finally, to explain DL models, we compared the relative weights (impact) of EGM variations to DL classification. We used as inputs to a logistic regression model the variations in reconstructed EGM sequences (normalized from 0 to 1). We then fitted the regression model output to previous DL predictions, and report the relative weight of controlled variations in rate, timing and shape to the DL decision in the regression model.

2.7. Classification metrics and statistics

We trained all EGM classifiers the same way, and validated their efficacy using the same metrics on the same datasets. A 10-set stratified Monte Carlo cross validation scheme with a patient-wise division was used, in which each cross validation set had 80% of patients for training (69 patients) and 20% for validation (17 patients). From multiple random divisions, 10 sets with stratified class distribution were selected: 34 AF + 35 AT or 35 AF + 34 AT patients for training and 8 AF + 9 AT or 9 AF + 8 AT patients for validation (detailed in Supplementary Table 2).

Performance was measured in the remaining 17 validation patients, and reported as mean ± standard deviation across the 10 validation sets

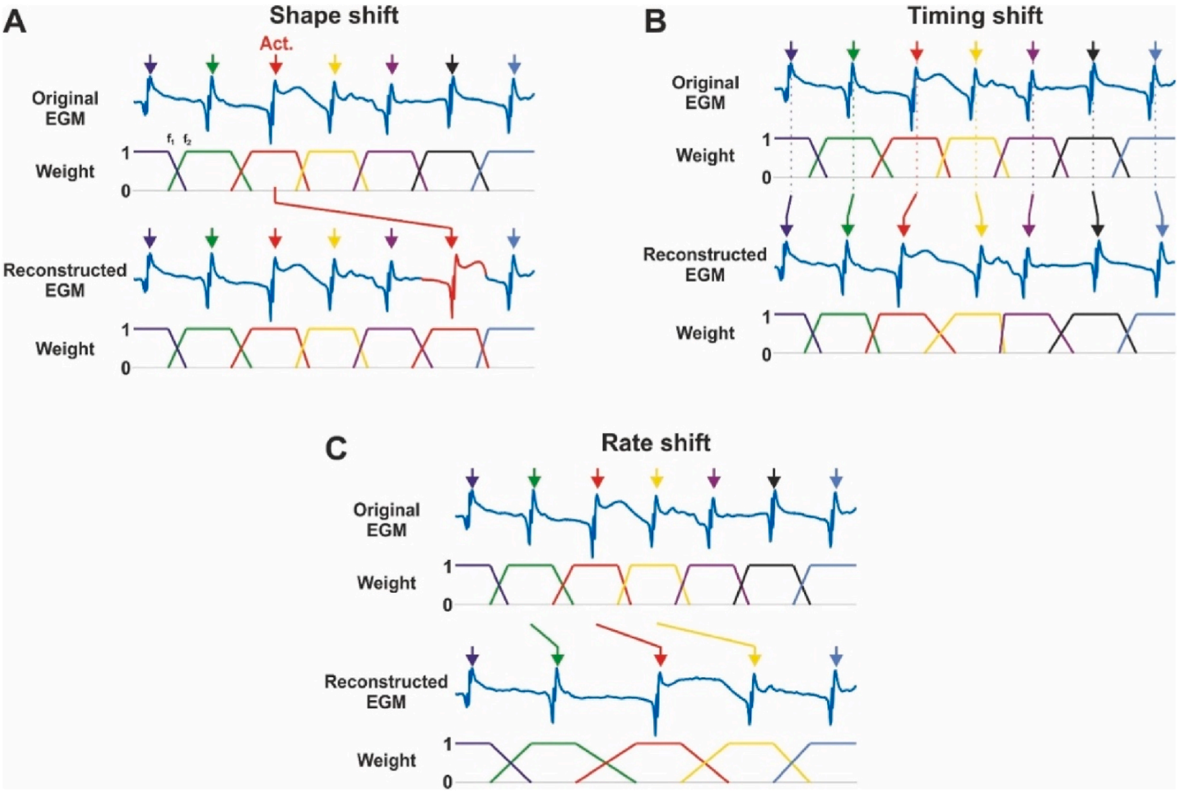

Fig. 2. AF Reconstructed EGMs. A. Shape shift. **B.** Timing shift. **C.** Rate shift.

patient-wise. The AUC was measured for each validation set by constructing the Receiver Operating Characteristics Curves. We report accuracy, sensitivity, specificity, Positive Predictive Value (PPV) and F-1 score using the optimal threshold for each classifier, selected as the threshold on the validation set that minimized *(1-sensitivity)*^{2} + *(1-specificity)*^{2}.

Statistical differences between continuous variables were assessed using paired or unpaired student T-test, and differences between categorical variables using the Chi-square test (χ^2). These statistical tests were used to compare patient demographics, outcome metrics of the classifiers or arrhythmia predictions in different signal subsets, but not for feature selection. We used a two-tailed alpha of 0.05 to indicate significance.

3. Results

Table 1 shows patient demographics. Patients presenting in AF had similar characteristics to those presenting in AT, except for a less frequent history of non-paroxysmal AF in patients presenting with AT ($p < 0.02$).

3.1. *Identifying AF by traditional features*

Table 2 shows the top 15 features of unipolar EGMs that distinguished AF from AT based on a single feature threshold. Overall, single features provided modest accuracy for AF. Rate features of cycle length or DF provided AUC of 0.75 and 0.67, respectively. Autocorrelation provided AUC of 0.83. EGM amplitude was the least accurate with AUC of 0.58. ROC curves are presented in Supplementary Fig. 2.

Optimal features for bipolar electrograms were similar but had lower predictive value. For bipolar signals, CL provided AUC for separating AF from AT of 0.76, DF provided AUC of 0.76, and EGM amplitude had AUC 0.65. (Supplementary Table 3).

Table 1
Patient demographics.

	All Patients	Patients with AF	Patients with AT	p-value
Number of Patients	86	43	43	–
Age (years)	60.7 ± 11.2	61.3 ± 11.6	59.8 ± 10.7	0.59
Female	16 (19%)	8 (19%)	8 (19%)	1
Weight (kg)	94.2 ± 19.3	93.5 ± 20.0	95.4 ± 18.6	0.69
CHA2DS2-VASc Score	2.0 ± 1.6	2.2 ± 1.7	1.7 ± 1.3	0.13
Time since diagnosis (months)	53.7 ± 60.1	51.3 ± 61.9	57.5 ± 58.2	0.68
LA volume (ml)	67 ± 22	69 ± 22	63 ± 23	0.43
Number with prior AF ablation	28 (33%)	15 (35%)	13 (30%)	0.65
Number with non-paroxysmal AF	26 (30%)	18 (42%)	8 (19%)	0.02

Table 2
Top 15 Unipolar Electrogram features that identified AF.

Feature	AUC	Sens.	Spec.
Autocorrelation: Peak max amplitude	0.83	0.75	0.78
Autocorrelation: Median	0.76	0.73	0.70
Cycle Length (Bott filt.)	0.75	0.70	0.73
Autocorrelation: Standard Deviation	0.71	0.65	0.67
Dominant Frequency 0–10 Hz (filt. Bott.)	0.67	0.63	0.64
Number of local maxima, 90–150 Hz filtered	0.66	0.57	0.69
Autocorrelation: Peak mean amplitude	0.65	0.58	0.61
Cycle Length (1–15 Hz filt.)	0.60	0.57	0.57
Dominant Frequency 2–8 Hz (filt. 1–15 Hz)	0.59	0.53	0.64
Absolute amplitude: 75% percentile	0.58	0.64	0.52
Absolute amplitude: 95% percentile	0.58	0.49	0.67
Peak number (120–150 Hz filt.)	0.58	0.61	0.53
Absolute amplitude: median	0.56	0.62	0.51
Absolute 1st derivative amplitude: 25% percentile	0.56	0.57	0.54
Absolute 1st derivative amplitude: 10% percentile	0.56	0.58	0.52

3.2. Combining traditional features to identify AF

We tested classifiers that combined the traditional features in Table 2. Supplemental Figure 3A presents the results of AUC to classify unipolar EGM by 4 classifiers (Linear Regression, Bagged Tree, K-Nearest Neighbor and Support Vector Machine), as a function of varying numbers of presented features. AUC reached a plateau of ~0.95 after 5 features, and dropped when >18 features were included.

Supplemental Figure 4B shows similar results for classifying bipolar electrograms, with plateau AUC 0.93 also achieved after ~5 features which fell after >18 features. Figure S4.C-D shows the classification accuracy for AF of the 4 feature-based classifiers, as the average of cross validation sets. Unipolar EGM classifiers were similar to bipolar EGM classifiers.

Linear Regression and SVM were the best multi-feature classifiers, respectively. Performance of the best feature-based classifiers for unipolar and bipolar EGM classification are included in Table 3. In general, classifiers combining different features showed better performance than individual features alone.

Because feature selection was conducted on the whole dataset this could over-estimate performance. We therefore compared feature selection computed only on the training data of each cross-validation set (Supplemental Fig. 4). Performance was similar.

3.3. Deep learning to identify AF

Supplemental Fig. 5 shows CNN and RNN classification metrics for raw unipolar and bipolar EGMs without the feature extraction used above. For unipolar EGMs, RNN and CNN had similar performance (AUC 0.97 ± 0.04 vs 0.95 ± 0.05; $p > 0.05$) which approximated the best feature-based classifier. Fig. 3B shows similar results for CNN and RNN on bipolar EGMs.

Supplemental Figure 5C-D shows ROC curves for detecting AF by pooling the 10-cross validation cohorts for all classifiers. For both unipolar and bipolar signals, DL thus was able to classify AF from AT without traditional rules yet with similar accuracy. Results of all feature-based and DL classifiers are summarized in Table 3.

We assessed the impact of controlled variations in EGM features on DL performance. Fig. 3A shows the impact of variations in EGM shape, calibrated by controlled changes in correlation coefficient (CC) across 4270 individual EGMs. Fig. 3B shows the impact of variations in activation timing, calibrated as the standard deviation of CL in 2250 EGMs. Fig. 3C shows the impact of varying EGM shape. EGMs with $CC > 0.9$ were classified as AT. With falling correlation coefficients of EGM shape (CC), classification as AF increased to 62.2% ($CC = 0.4$) and to 83.4% ($CC = 0.1$).

Fig. 3D shows the impact of CL irregularity. EGM signals with timing variability $<10\%$ were classified as AT 94% of the time, while EGMs with variability $>20\%$ CL were classified as AF $>85\%$ of the time.

Overall, AF was optimally identified by shape $CC < 0.48$ and timing variability $>15\%$ of CL ($p < 0.001$, χ^2). We probed feature-based classifiers using this approach (Supplemental Fig. 5). We found a similar trend for SVM, for which AF was optimally identified by shape $CC < 0.37$ and timing variability $>18\%$ of CL.

3.4. Probing DL for composite signatures of AF

We assessed the impact of concurrent shape and timing changes in $N = 29{,}190$ reconstructed EGM sequences. Fig. 4A illustrates simultaneous controlled variation in both shape and timing. Fig. 4B shows their impact on DL classification, and Fig. 4C shows their impact on the best feature-based classifiers. In each, classification is color-coded by the percentage of EGMs classified as 100% AT (blue) to 100% AF (red).

For DL, Fig. 4B indicates that modifying EGM shape from CC 0.1 to 0.9 and timing from 0 to 35% did not reclassify AF to AT or vice versa. DL classified $>50\%$ sequences as AF except a small population of EGM sequences with $<20\%$ timing variability and shape $CC > 0.5$ (Fig. 4B blue). Even in these sequences, 40% were classified as AF. Thus, DL classification either did not depend on linear combinations of EGM shape and timing, or used additional features. Conversely, AF diagnosis by SVM (Fig. 4C) ranged from 0% AF for regular EGMs (0.9 CC, 0% CL variation) to 100% AF in irregular EGMs ($CC < 0.4$, timing $>15\%$ CL) and thus could be explained by variations in timing and shape alone.

We now examined $N = 2091$ reconstructed EGMs with varying rate yet constant shape and regularity (Fig. 5A). For DL, Fig. 5B shows that EGM rate moderately explained AF identification. In sequences with $CL < 180$ ms, 62–100% were classified as AF. In sequences with $CL > 180$ ms, $>70\%$ were classified as AT. $CL < 190$ ms optimally separated AF from AT (AUC 0.83). Conversely, the optimal feature-based classifiers (SVM) did less well using a rate cut-point alone (Fig. 5C).

Finally, we measured the relative importance of each controlled variation (rate, shape, timing) to DL in the dataset of $N = 133{,}100$ reconstructed EGMs, using a logistic regression model. This model used as inputs the variations in timing, rate and shape of each reconstructed EGM, and as output the DL prediction. EGM shape consistency contributed 13.0% to DL classifications (95% CI [12.3%–13.7%], $p < 0.001$), timing regularity contributed 26.9% (95% CI [25.8%–28.0%], $p < 0.001$) and rate contributed 60.1% (95% CI [59.7%–60.5%], $p < 0.001$). The logistic regression model explained only ~70% of DL classification ($AUC = 0.72$). We concluded that DL may code relationships between rate, timing or shape in a non-linear fashion or, alternatively, that DL could be making classification from additional parameters.

3.5. Probing DL to reveal specific EGM features for AF

We studied if specific EGM shape morphologies may influence DL classification. We examined $N = 75{,}166$ EGM sequences with 100% consistency in shape that were classified as AF independent of rate or

Table 3
Summary of the performance of the different classifiers. (FB: Feature-Based; DL: Deep-Learning).

		Classifier	Acc.	Sens.	Spec.	AUC	PPV	F-1
Unipolar EGMs	FB	Linear Regression (Linear)	0.88	0.95	0.93	0.95	0.95	0.95
		Bagged Trees (Forest)	0.88	0.93	0.93	0.94	0.93	0.93
		K-Nearest Neighbor (KNN)	0.88	0.93	0.93	0.94	0.93	0.93
		Support Vector Machine (SVM)	0.87	0.96	0.91	0.94	0.96	0.96
	DL	Convolutional (CNN)	0.88	0.91	0.95	0.95	0.91	0.91
		Recurrent (RNN)	0.89	0.96	0.93	0.97	0.96	0.96
Bipolar EGMs	FB	Linear Regression (Linear)	0.87	0.93	0.91	0.93	0.93	0.93
		Bagged Trees (Forest)	0.85	0.94	0.87	0.93	0.94	0.94
		K-Nearest Neighbor (KNN)	0.84	0.92	0.92	0.93	0.92	0.92
		Support Vector Machine (SVM)	0.83	0.91	0.93	0.93	0.91	0.91
	DL	Convolutional (CNN)	0.87	0.95	0.88	0.93	0.95	0.95
		Recurrent (RNN)	0.81	0.92	0.88	0.92	0.92	0.92

Explainability of DL to Identify AF Signatures.

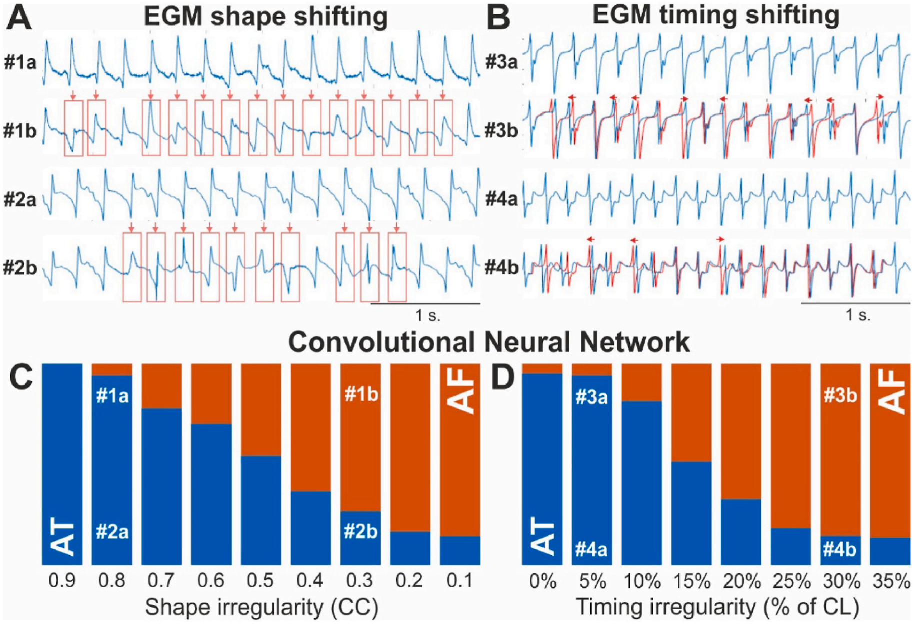

Fig. 3. Classification of Reconstructed EGM based on Shape and Timing Irregularity. Reconstructed EGM generation using shape shifting (A) and time shifting (B). Red boxes and EGM signals mark the variations on the reconstructed EGM respect to the departing EGMs. Classification of reconstructed EGMs by CNNs based on Shape Irregularity (C) and Timing Irregularity (D). Red: classified as AF; Blue: classified as AT.

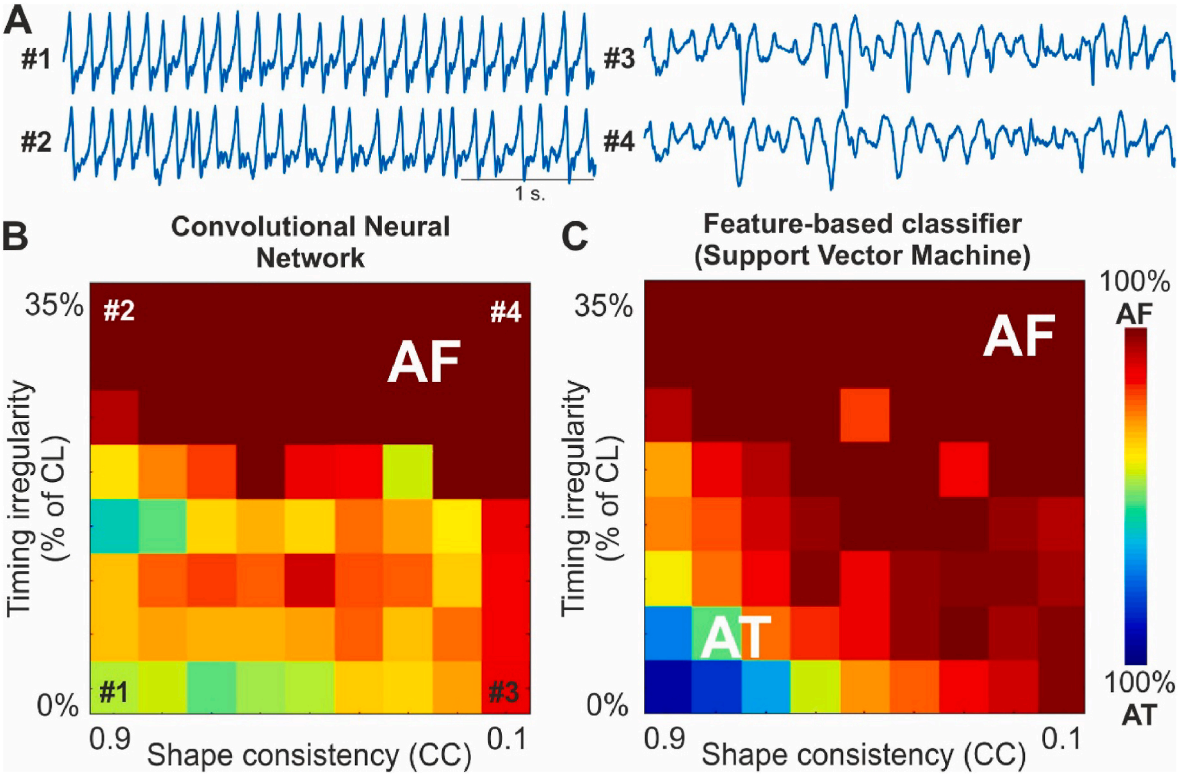

Fig. 4. Classification of Reconstructed EGM based on Shape and Timing Consistency. A. Reconstructed EGM signals generated using both shape and time shifting. Classification of reconstructed EGMs by CNNs (B) and SVM (C) based on both Shape and Timing consistency, color-coded according to the percentage of signals classified as AF.

timing (Fig. 6A).

Of the N = 415 individual unipolar EGM shapes, 101 (24%) were classified by CNN as AF in >80% of experiments independent of variations in EGM shape or timing (Fig. 6B). For RNNs, 30/415 EGM shapes were classified as AF in >80% of experiments, independent of other parameters (Fig. 6C). From both experiments, 15 EGM traces were classified as AF in >80% of experiments by both RNN and CNNs, higher than expected (p = 0.03, χ^2). Fig. 6D shows these 15 unipolar EGM morphologies, which were complex with multiple deflections and fractionation. Similar results were found when examining classification of

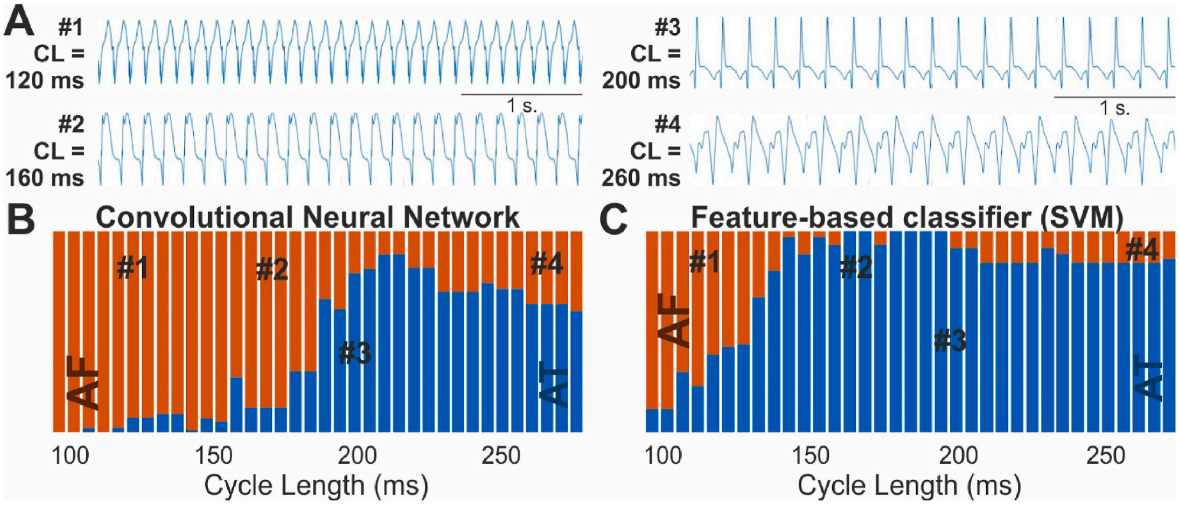

Fig. 5. Classification of Reconstructed EGM based on Cycle Length. A. Reconstructed EGM signals generated with timing and shape consistency and varying their Cycle Length. Classification of reconstructed EGMs by CNNs (B) and SVM (C) based on Cycle Length. Red: classified as AF; Blue: classified as AT.

AF in >75% to >95% of experiments.

4. Discussion

We show that deep learning can accurately identify AF from organized ATs even that overlap in rate or regularity. Probing DL to explain their classification revealed specific cutpoints of timing variability, variability in electrogram shape and fast rate used by DL to classify AF as opposed to AT. Notably, we also identified a set of unique EGM shapes that classified as AF by multiple DL architectures regardless of variations in EGM shape, timing, or high rate. Thus, deep learning approximated the performance of expert rules, yet uncovered non-linear variations in rate and regularity or additional features not revealed by logistical regression analysis. Our use of computer modeling to explain DL by controlled variations in clinically intuitive parameters could be useful in different physiological and clinical applications. Clinically, studies should explore if EGM fingerprints, and variations in rate or regularity differ between sub-types of AF patients such as paroxysmal versus persistent AF, patients with or without fibrotic atrial remodeling, or with differing response to therapy.

4.1. Classification performance

The use of machine or deep learning to classify intracardiac EGMs is relatively new, although DL has been extensively use to separate AF and other atrial arrhythmias from Sinus Rhythm (SR) on the ECG. A recent review by Fatma et al. [18], reported that accuracy for detecting AF vs sinus rhythm using deep learning methods ranged from 90% to 99.7%. However, the accuracy of DL on the ECG for the current problem of separating AF from AT falls to 89.7% [19].

Surprisingly, features based on single expert rules such as cycle length or DF had low predictive accuracy for AF. Combining multiple features improved classification and plateaued for <20 features (Fig. S3). This suggests that the N = 45 features included in this study (Supplementary Table 1) spanned key features that separate AF from other arrhythmias. Nevertheless, this approach may be vulnerable to patient differences or variations in the classification problem and so may not be scaleable.

DL classification of raw EGMs were at least as effective as feature-based classifiers, but may be more scaleable as it did not require problem-specific features to be identified. Moreover, in our study, DL identified novel features that extend beyond those described by experts in the literature.

Classification of bipolar EGMs showed lower performance than unipolar EGMs. Given that the same methods were used for both types of signals, this difference likely results from the maximal predictive value of individual features (0.83 AUC for unipolar vs. 0.76 AUC for bipolar) but could potentially also reflect differences in data size (29,340 unipolar vs. 23,760 bipolar EGMs). Whether different features should be used for bipolar EGMs, not included in this manuscript, will be further explored.

4.2. Explaining DL to identify potential AF signatures

We varied one feature at a time in reconstructed signals to quantify the contribution of each to AF identification. In this way, we defined a novel composite signature for AF comprising >15% timing variation, <0.48 correlation between successive EGMs, CL < 190 ms and also novel EGM shapes classified as AF regardless of rate and timing. Other features may also exist, such as fractionation or Shannon Entropy which are composites of shape, rate and timing and require further study.

4.3. Probing and explaining DL

DL shows excellent classification performance, yet its medical use has been limited by a lack of explainability for its decisions [7,8]. To address this limitation, our group has studied approaches to explain how DL can predict ventricular arrhythmias from cellular mechanisms [15] or how DL can interpret complex activation maps of AF [16].

The present study provides insights to explain how DL classifies fibrillatory rhythms from intracardiac EGMs, compared to traditional expert rules. The DL approach may provide a platform to identify AF from intracardiac devices or wearable devices by examining ECG-based features using transfer learning or de novo models.

An advantage of DL is that it does not rely upon linear dependencies between input features and the classification (AF or AT), unlike logistic regression analysis, which may be more appropriate for complex physiological problems. Classic feature-based models, such as logistic regression, are limited to the linear dependence of the features onto the classification which may be insufficient for an accurate electrophysiological description.

4.4. Clinical implications

An immediate implication of this work is to better identify AF from cardiac implanted electronic devices such as pacemakers and ICDs. This approach also could be applied to other EGM signals such as ambulatory ECGs from wearable devices [9,10]. Notably, our results summarized in

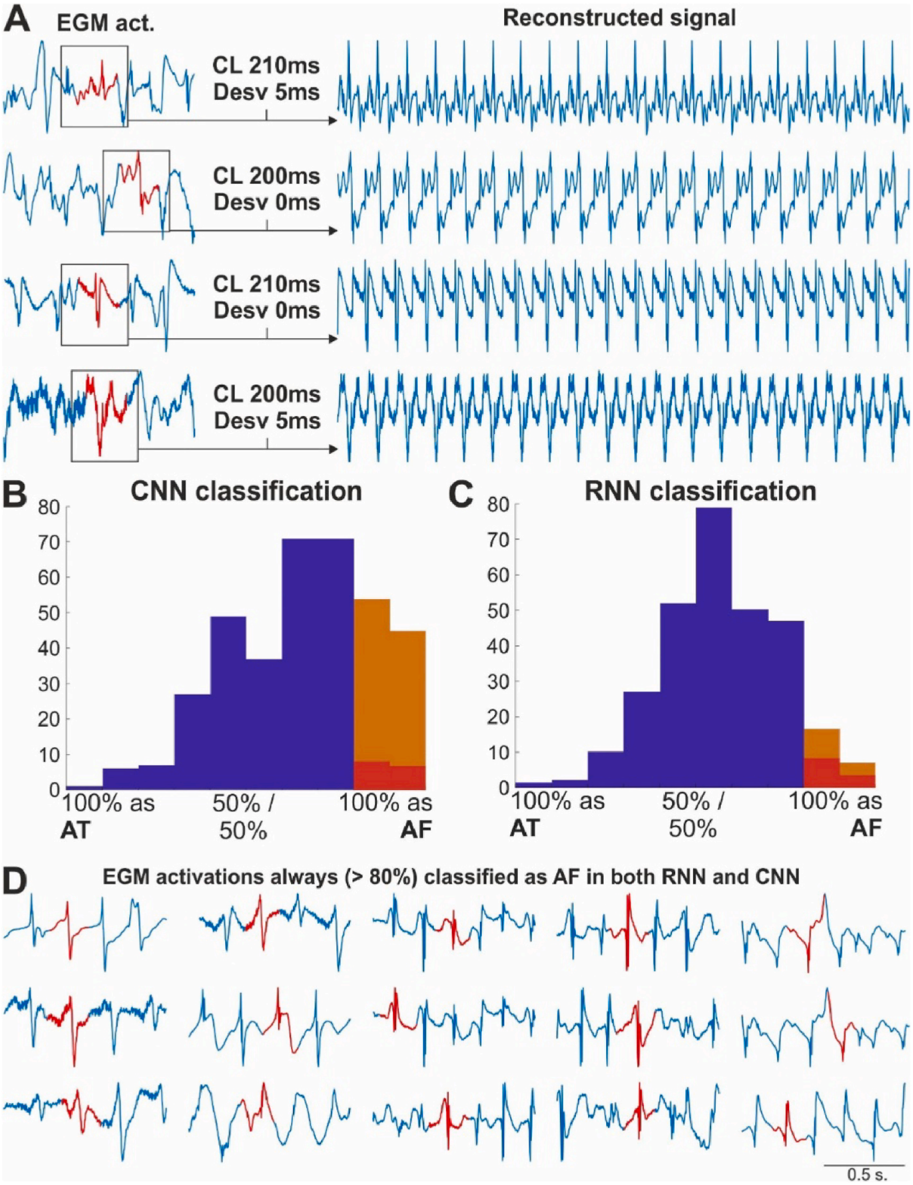

Fig. 6. Classification of AF based on specific EGM shape, independent of beat-to-beat variations in timing or shape. A. Reconstructed EGM signals with unique activation shapes. (B) Histogram of CNN and (C) RNN classification of these EGM shapes as AF or AT. Orange represents EGM shapes classified >80% as AF, blue indicates EGM shapes classified <80%, and red indicates EGM shapes coinciding in orange regions. D. EGM shapes classified >80% as AF independently of their time, shape or CL irregularity by both CNN and RNN.

Figs. S3 and S5 show that extracting a small number of specific EGM features to train a simple classifier provides diagnostic performance similar to more complex DL models. Indeed, we have previously shown that quantitative variations in the ECG f-wave can separate AF from atypical and them from typical AT [13,18]. More broadly, DL-based EGM signatures could be potentially applied to CIEDs, modified for use to the ECG, or applied during catheter ablation.

The proposed methodology can be used to identify features for DL analysis of intracardiac EGM analysis based on explainability analysis. Our study identified AF fingerprints such as EGM features of multiple deflections and complex morphology which may reflect underlying disease or substrate conditions (presence of fibrotic tissue or tissue anisotropy). These may also be rate-dependent. These novel findings broaden the application of DL for rhythm classification, and reveal specific EGM features consistent with conditions such as fibrosis. This approach could thus form the basis for future hypotheses testing such as separating patient subtypes based on structural or electrical remodeling [20].

4.5. Limitations

We designed our study using basket catheter signals instead of using data from implantable devices, because this provided the opportunity to sample multiple regions of both atria simultaneously, enabling spatial comparisons and providing unequivocal diagnosis of AF or AT at invasive EP study. These tools must be extended to ICD or pacemaker recordings. The need for a large database may limit the application of DL to smaller datasets, and the use of transfer learning or other approaches could be used to apply the current analyses more broadly. We cannot exclude that these DL models are catheter specific, and future work should examine other catheters including higher-resolution smaller electrode designs. This work is ongoing in our laboratory. It is not clear whether AF signatures are region-specific, and future work could examine EGM near the pulmonary veins, left atrial appendage and other regions to further enhance patient classification. Finally, while patients were free of anti-arrhythmic medications at the time of their electrophysiological study, it is not clear whether their specific co-morbidities or medications may contribute to these results.

This manuscript focuses on the relative performance of different architectures to classify atrial EGMs, and how 'explainability' analysis of classifier decisions can guide interpretation. Our goal was to allow a foundation to construct future classification models for AF with deeper knowledge of architectures and how these reflect biological and clinical features. This manuscript was therefore focused on model comparison without exhaustive fine-tuning of each model, using a cross-validation scheme allowing comparison across patients. Ultimately, fully generalizable models should be tested with several independent datasets for clinical practice.

5. Conclusions

Deep learning was developed to identify AF from AT, and revealed classification features including novel EGM shapes, >15% timing variation, <0.48 correlation between EGMs and CL < 190 ms. This integrated computer modeling and machine learning approach could be applied to reveal sub-types of AF patients ('computational signatures') with differing underlying substrate, mechanisms or response to therapy.

Declaration of competing interest

MR: Equity Interests from Corify Care, S.L. MIH: Intellectual Property Rights from Stanford University. SMN: Compensation for Services from Up to Date, Abbott Laboratories, TDK Inc. Intellectual Property Rights from University of California Regents and Stanford University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105451.

References

[1] G. Hindricks, et al., Performance of a new leadless im-plantable cardiac monitor in detecting and quantifying atrial fibrillation: results of the XPECT trial, Circ Arrhythm Electrophysiol 3 (2) (2010) 141–147.

[2] M. V Perez, et al., Large-scale Assessment of a smartwatch to identify atrial fibrillation, N. Engl. J. Med. 381 (20) (2019) 1909–1917.

[3] Y. Guo, et al., Mobile photoplethysmographic technology to detect atrial fibrillation, J. Am. Coll. Cardiol. 74 (19) (2019) 2365–2375.

[4] E.S. Kaufman, et al., Positive predictive value of device-detected atrial high-rate episodes at different rates and durations: an analysis from ASSERT, Heart Rhythm 9 (8) (2012) 1241–1246.

[5] E. Bertaglia, et al., Atrial high-rate episodes: prevalence, stroke risk, implications for management, and clinical gaps in evidence, Europace 21 (10) (2019) 1459–1467.

[6] T.T. Tomson, et al., Management of device-detected atrial high-rate episodes, Card Electrophysiol Clin 7 (3) (2015) 515–525.

[7] C. Krittanawong, et al., Deep learning for cardiovascular medicine: a practical primer, Eur. Heart J. 40 (25) (2019) 2058–2073.

[8] E.J. Topol, et al., High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56.

[9] J.M. Bumgarner, et al., Smartwatch algorithm for automated detection of atrial fibrillation, J. Am. Coll. Cardiol. 71 (21) (2018) 2381–2388.

[10] G.H. Tison, et al., Passive detection of atrial fibrillation using a commercially available smartwatch, JAMA Cardiol 3 (5) (2018) 409–416.

[11] S. Honarbakhsh, et al., Panoramic atrial mapping with basket catheters: a quantitative analysis to optimize practice, patient selection, and catheter choice, J. Cardiovasc. Electrophysiol. 28 (12) (2017) 1423–1432.

[12] M. Rodrigo M, et al., Non-invasive spatial mapping of frequencies in atrial fibrillation: correlation with contact, Mapping'' Front Physiol 11 (2021) 611266.

[13] M.I. Alhusseini, et al., Machine learning to classify intra-cardiac electrical patterns during atrial fibrillation: machine learning of atrial fibrillation, Circ Arrhythm Electrophysiol 13 (8) (2020) e008160.

[14] A.K. Feeny, et al., Artificial intelligence and machine learning in arrhythmias and cardiac Electrophysiology, Circ Arrhythm Electrophysiol 13 (8) (2020) e007952.

[15] A.J. Rogers, et al., Machine learned cellular phenotypes predict outcome in ischemic cardiomyopathy, Circ. Res. 128 (2) (2020) 172–184.

[16] S. Liaqat, et al., Detection of atrial fibrillation using a machine learning approach, Information 11 (12) (2020) 549.

[17] Ribeiro, et al., Why should I trust you? Explaining the Predictions of Any Classifier'' arXiv 1602 (2016), 04938.

[18] F. Murat, et al., Review of deep learning-based atrial fibrillation detection studies, Int. J. Environ. Res. Publ. Health 28 (21) (2021) 11302.

[19] M.D. Ivanovic, et al., Deep learning approach for highly specific atrial fibrillation and flutter detection based on RR intervals, Annu Int Conf IEEE Eng Med Biol Soc (2019) 1780–1783.

[20] B. Deb, et al., Identifying atrial fibrillation mechanisms for personalized medicine, J. Clin. Med. 10 (23) (2021) 5679.