Deep Anomaly Detection of Bipolar Electrograms for Characterizing the Atrial Substrate

Filippo Uslenghi[1], Lorenzo Gigli[2], Massimo W Rivolta[1], Roberto Sassi[1]

[1] Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
[2] De Gasperis Cardio Center, Electrophysiology Unit, Niguarda Hospital, Milan, Italy

Abstract

Electrograms (EGM), collected during catheter ablation for atrial fibrillation (AF), allows the characterization of the atrial substrate and its remodeling. While methods for EGM analysis based on electrophysiological models have been extensively studied, the use of deep learning methods remains limited. In this study, we investigated an anomaly detection framework based on normalizing flows (NF) to distinguish pathological EGMs from healthy ones. NF provide an invertible mapping from the space of EGM signals to the samples of a multivariate normal distribution, from which convenient measures of probability can be used as anomaly scores. One-hundred-twenty-one patients were split into three cohorts: C1) 75 patients with paroxysmal AF, normal substrate, mapped in sinus rhythm; C2) 46 patients with persistent AF, abnormal substrate, or mapping during AF; and C3) 12 patients with persistent AF, abnormal substrate and mapped during AF (subset of C2). The norm of each sample, mapped by NF, was used to derive the anomaly score (a probability), on which a threshold was imposed. ROC analysis displayed AUC values of 0.79 (95% CI: 0.67–0.89) for C1 vs C2, and 0.95 (95% CI: 0.85–1.00) for C1 vs C3, indicating increasing separability among cohorts with increased phenotype severity. These results suggest NF provides an effective approach for patient-level EGM anomaly scoring in AF patients.

1. Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia worldwide, and catheter ablation (CA) is a standard treatment option [1]. During CA, electrograms (EGMs) are recorded with mapping catheters to construct an electroanatomic map of the atria, enabling clinicians to characterize atrial electrophysiology and substrate. Computational methods to support such characterization have a long standing history [2]. For example, Faes *et al.* proposed a method to quantify AF organization from single bipolar EGMs by comparing the morphology of local activation waves and deriving a regularity index based on their similarity over time [3]. More recently, Gigli *et al.* proposed a substrate characterization measure based on Shannon entropy of the voltage-amplitude distribution computed over the atrial surface, and reported an association with substrate health status, as assessed during clinical care by electrophysiologists in patients with AF [4].

While these approaches rely on hand-crafted signal descriptors, deep learning offers a data-driven alternative with potential to advance EGM characterization. Since substrate alterations modulate local activation and conduction, EGM characteristics provide a signal-level proxy for substrate characterization. In this context, anomaly detection using AI models based on deep learning models, or "deep anomaly detection", may offer a promising direction, by modeling the distribution of normal data. These methods assign anomaly scores based on deviation from the learned distribution. Applied to electrophysiological recordings, learning the distribution of physiological EGMs should enable the identification and characterization of pathological EGMs, thus helping in localizing unhealthy regions of the atrial substrate.

To the best of our knowledge, the literature on deep anomaly detection for EGM characterization is very limited, with the study by Bindini *et al.* [5] being one of the few available examples. In their work, the authors developed three anomaly detection models, namely variational autoencoders, memory-augmented autoencoders, and deep support vector data description, and evaluated their performance against clinical indicators used to characterize pathological EGMs. While preliminar, their results supported the applicability of deep anomaly detection in this setting, showing agreement across models and strong correlations between anomaly scores and clinical indicators.

In this work, we further investigated the application of anomaly detection to electrophysiology. Inspired by Rudolph *et al.* [6], who studied normalizing-flow–based anomaly detection in an industrial setting, we introduced normalizing flows for modeling the distribution of EGMs recorded from patients suffering from paroxismal AF, an

early phase of the disease, with still an healthy atrial substrate, and mapping performed in sinus rhythm (SR). The aim is to learn the characteristics of physiological EGMs and discriminate pathological EGMs. Normalizing flows are deep generative models that learn a bijective transformation between the data distribution and a chosen base distribution, typically a multivariate Gaussian distribution. The main idea is to map EGMs to points in the space of the base distribution, where its tractable formulation allows identifying outliers using percentiles and probabilities, rather than thresholding anomaly scores provided by deep neural networks.

2. Methods

2.1. Dataset

The dataset was collected as part of routine clinical care in the electrophysiology unit of Niguarda Hospital (Milan, Italy) over 4 years. All participants provided informed consent prior to data collection. The study protocol was approved by the Ethics Committee of Niguarda Hospital in compliance with the Declaration of Helsinki. The cohort comprised 121 patients with paroxysmal or persistent AF undergoing catheter ablation. For each patient, a left atrial electroanatomic map was acquired prior to ablation using the CARTO® 3 electroanatomic mapping system and the PENTARAY™ or OCTARAY™ mapping catheter (Biosense Webster, Diamond Bar, CA, USA). Map density and the number of recorded signals varied across cases depending on electrophysiologist preference. The rhythm during mapping was SR, paced atrial rhythm, or AF. At the end of the mapping, the electrophysiologist categorized the atrial substrate as normal or abnormal.

The collected signals were sampled at $1,000$ Hz, and the bipolar EGM activations were extracted using the OpenEP library [7]. Each EGM was centered in a 128-sample window around the local activation time provided by OpenEP, bandpass filtered with a third-order Butterworth filter (30–300 Hz), and then downsampled to 700 Hz to reduce the segment length, yielding 90-sample segments. EGMs with peak-to-peak voltage amplitude above the 97.5$^{\text{th}}$ percentile (~ 7 mV) were considered noisy and discarded. Each segment was normalized by subtracting its mean and dividing by its standard deviation. The dataset comprised $138,816$ bipolar activations, each ~ 0.13 s long.

Three patient cohorts were defined. Cohort 1 (C1) included 75 patients with paroxysmal AF, classified as having a normal substrate, and mapped in SR. Cohort 2 (C2) comprised the remaining 46 patients who did not meet all C1 criteria, i.e., those with persistent AF and/or abnormal substrate and/or mapping performed during AF. Cohort 3 (C3), a subset of C2, included 12 patients with persistent AF, abnormal substrate, and mapping performed during AF. This split reflected the fact that C1 represented a relatively early phenotype, characterized by normal substrate and mapping in SR, and thus provided signals representative of less advanced disease. In contrast, C2 included patients who deviate from one or more of these conditions. C3 corresponded to the most compromised group, with abnormal substrate and mapped during AF, and was therefore expected to exhibit the most pronounced AF-related signal characteristics. Patients in C1 were split into training (60%), validation (15%), and test (25%) sets, whereas patients in C2 and C3 were reserved for testing.

2.2. Normalizing Flows

Normalizing flows are deep generative models that represent complex probability distributions [8]. Let p_X denote the target distribution and p_Z a chosen base distribution (e.g., Gaussian), with the same dimensionality. A normalizing flow is an invertible, parameterized transformation $\boldsymbol{f}_\theta : Z \to X$ that maps samples $\mathbf{z} \sim p_Z(\mathbf{z})$ to samples $\mathbf{x} = \boldsymbol{f}_\theta(\mathbf{z})$ distributed according to p_X. The model is trained by maximizing the log-likelihood of observed data under the change-of-variable formula:

$$p_X(\mathbf{x};\theta) = p_Z\big(\boldsymbol{f}_\theta^{-1}(\mathbf{x})\big)\left|\det \frac{\partial \boldsymbol{f}_\theta^{-1}(\mathbf{x})}{\partial \mathbf{x}}\right|, \qquad (1)$$

where, as common in this context, $\det \frac{\partial \boldsymbol{f}_\theta^{-1}(\mathbf{x})}{\partial \mathbf{x}}$ is the determinant of the Jacobian of the transformation, which accounts for local volume changes induced by the mapping.

Computing the Jacobian determinant (or equivalently its log-determinant) for an arbitrary high-dimensional transformation is computationally prohibitive. To make likelihood evaluation tractable, Dinh *et al.* introduced affine coupling layers [9], a class of invertible transformations whose Jacobian is triangular, yielding a simplified determinant formulation. Briefly, let $\mathbf{u} \in \mathbb{R}^N$ be the input to an affine coupling layer and choose $1 \le n < N$. Splitting $\mathbf{u}$ into $\mathbf{u}_{1:n}$ and $\mathbf{u}_{n+1:N}$, indicating the vectors built considering the elements of $\mathbf{u}$ from 1 to n and from $n+1$ to N, and defining $s, t : \mathbb{R}^n \to \mathbb{R}^{N-n}$, the output $\mathbf{v} \in \mathbb{R}^N$ of the affine coupling layer is given by

$$\mathbf{v} = \begin{bmatrix} \mathbf{u}_{1:n} \\ \mathbf{u}_{n+1:N} \odot \exp\big(s(\mathbf{u}_{1:n})\big) + t(\mathbf{u}_{1:n}) \end{bmatrix}, \qquad (2)$$

where $\odot$ represents the Hadamard product. Using this definition, the log-determinant reduces to $\sum_{i=1}^{N-n} s(\mathbf{u}_{1:n})_i$, since the Jacobian is triangular with diagonal entries $\exp(s)$. This construction leaves $\mathbf{u}_{1:n}$ unchanged while applying an affine transformation to $\mathbf{u}_{n+1:N}$ conditioned on $\mathbf{u}_{1:n}$. Stacking multiple coupling layers yields an expressive nonlinear transformation $\boldsymbol{f}_\theta$. Moreover, permuting the input dimensions between successive layers ensures that all components are transformed across the flow.

2.3. Technical implementation

In this work, we built a normalizing flow composed of 16 affine coupling layers. The layers alternated between (i) coupling with an alternating (even–odd) mask to select the identity and affine-transformed subvectors, followed by a random shuffle permutation, and (ii) a channel-wise split in which the input vector is divided in half and the two halves are swapped. The scaling s and translation t functions were each parameterized by a multilayer perceptron (MLP) with four fully connected layers. To match the half-dimensional input induced by the coupling mask, the network used 45 units in the first layer, 128 units in the second and third layers, and 45 units in the output layer. ReLU activations followed the second and third layers, while a $\tanh$ activation was applied at the output to bound the scale and to shift terms to $(1/e, e)$ and $(-1, 1)$, respectively, for improved numerical stability. The base distribution of the normalizing flow was set to a 90-dimensional standard multivariate normal, with mean vector $\mu = \mathbf{0}$ and covariance matrix $\Sigma = I_{90}$, both fixed (*i.e.*, not learned).

Hyperparameters (batch size, number of affine coupling layers, and the number of units in the MLPs parameterizing s and t) were selected via a tuning phase on the training/validation sets. During hyperparameter selection, models were trained for 100 epochs with the Adam optimizer (learning rate $5{\times}10^{-4}$, weight decay $1{\times}10^{-5}$) to minimize the negative log-likelihood computed via (1). The model achieving the highest validation log-likelihood was trained with 16 affine coupling layers for 87 epochs with a batch size of 32 and it was retained as the final model.

3. Experiments

Signals from patients in the Cohort 1 test set and from the remaining cohorts were processed with the trained model and mapped to the latent space via $\mathbf{z} = \boldsymbol{f}_\theta^{-1}(\mathbf{x})$. For each patient p, we computed the mean squared Euclidean norm of the latent vectors across EGMs:

$$r_p^2 = \frac{1}{|S_p|} \sum_{\mathbf{x} \in S_p} \left\| \boldsymbol{f}_\theta^{-1}(\mathbf{x}) \right\|_2^2 , \tag{3}$$

where S_p denotes the set of EGM segments from patient p. We considered two patient-level anomaly scores. For the first score, we computed the negative log-density under the 90-dimensional standard normal distribution and defined the anomaly score as

$$\ell_p = \frac{1}{2} \left(90 \log(2\pi) + r_p^2 \right) . \tag{4}$$

The rationale is that a flow trained on normal substrates is expected to assign high density to normal samples and lower density to pathological ones, which deviate from the learned distribution and are underrepresented in the training set. For the second anomaly measure, we mapped the latent norm r_p to the corresponding χ CDF value. Since $\|Z\|_2 \overset{d}{=} \chi_{90}$ for $Z \sim \mathcal{N}(\mathbf{0}, I_{90})$, we computed

$$u_p = F_{\chi_{90}}(r_p) , \tag{5}$$

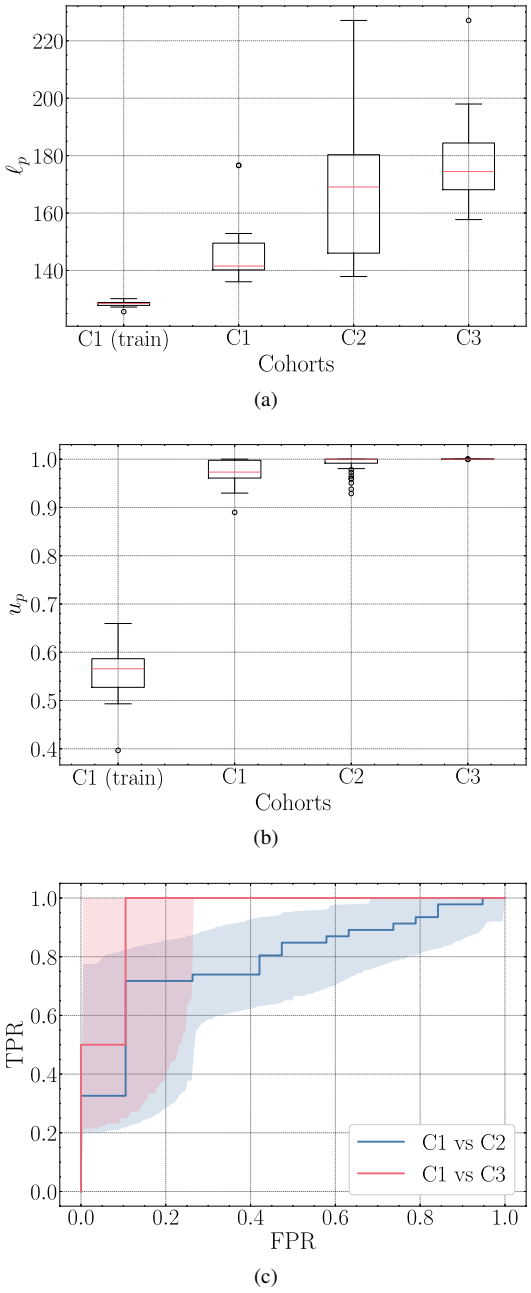

Figure 1: (a) Boxplot of the first anomaly score ℓ_p across the patient cohorts. (b) Boxplot of the second anomaly score u_p across the patients cohorts. (c) ROC curves for C1 vs C2 and C1 vs C3 based on the second anomaly score. Shaded bands indicate 95% confidence intervals.

where $F_{\chi_{90}}$ is the CDF of the χ distribution with 90 degrees of freedom. Thus, $u_p = P(\|Z\|_2 \leq r_p)$ quantified the probability that a random sample Z lies within a hypersphere of radius r_p.

Two Receiver Operating Characteristic (ROC) analyses were performed to differentiate C1 from C2 and C1 from C3, respectively. The area under the curve (AUC) was used as a performance metric. It should be noted that ℓ_p and u_p have the same AUC since they have a monotonic relationship. However, we believe that u_p, which defined a score in terms of probability, was a more interpretable quantity.

4. Results

Confidence intervals (CIs) at 95% were estimated via bootstrap resampling (10,000 trials) using the percentile method. Figures 1a and 1b show the distributions of the two anomaly scores across the three cohorts. Figure 1c shows the ROC curves for discriminating C1 vs C2 (AUC: 0.79, 95% CI: 0.67–0.89) and C1 vs C3 (AUC: 0.95, 95% CI: 0.85–1.00) based on the u_p score. The u_p threshold that best balanced the true positive rate and true negative rate was 0.9986 for C1 vs C2 and 0.9999 for C1 vs C3.

Notably, Figure 1b shows relatively high u_p values even for C1 signals, especially when compared with the clearly lower u_p values observed in the C1 training set. This pattern may indicate overfitting to the training data, suggesting that the adopted strategy to limit overfitting (early stopping based on validation performance) may not have been sufficient and that additional regularization or model-selection strategies should be explored. Nevertheless, this effect did not compromise the overall methodology or the main separability results.

5. Discussion and Conclusion

In this work, we developed an anomaly detection framework based on normalizing flows trained on EGMs from patients with normal atrial substrate, in the context of AF. When evaluated on EGMs from patients with normal and abnormal substrate, the model achieved good separability and high AUCs, particularly for C1 vs C3.

A different approach was followed in [5], where information about atrial substrate was not available, which precluded defining a normal-substrate training set and evaluating performance on a labeled test set. Nevertheless, the authors assumed that physiological EGMs were more prevalent than pathological EGMs and, therefore, that anomaly detection models would primarily learn the support of normal signals, separating them from anomalous ones. This intuition was supported by the agreement observed between the anomaly scores produced by their models and the clinical indicators considered in their analysis.

In conclusion, our results suggested that normalizing flows are a viable tool for EGM anomaly scoring in AF. Training on signals, acquired in sinus rhythm from patients with normal substrate and an early disease profile, enabled discrimination of patients with increasingly compromised phenotypes, particularly those with abnormal substrate and mapping during AF. Future works will focus on analyzing and validating samples generated by the model, including both physiological and pathological EGMs.

References

[1] Arbelo E, Brugada J, Hindricks G, et al. The Atrial Fibrillation Ablation Pilot Study: an European Survey on Methodology and results of catheter ablation for atrial fibrillation conducted by the European Heart Rhythm Association. European Heart Journal June 2014;35(22):1466–1478. ISSN 0195-668X.

[2] Ravelli F, Masè M. Computational mapping in atrial fibrillation: how the integration of signal-derived maps may guide the localization of critical sources. EP Europace May 2014; 16(5):714–723. ISSN 1099-5129.

[3] Faes L, Nollo G, Antolini R, et al. A method for quantifying atrial fibrillation organization based on wave-morphology similarity. IEEE Transactions on Biomedical Engineering December 2002;49(12):1504–1513. ISSN 1558-2531.

[4] Gigli L, Preda A, Coluzzi D, et al. Left atrial spatial entropy: a novel tool for electrophysiological substrate characterization in atrial fibrillation. Frontiers in Physiology November 2024;15. ISSN 1664-042X.

[5] Bindini L, Pagani S, Bernardini A, et al. All-in-one electrical atrial substrate indicators with deep anomaly detection. Biomedical Signal Processing and Control December 2024; 98:106737. ISSN 1746-8094.

[6] Rudolph M, Wandt B, Rosenhahn B. Same Same But DifferNet: Semi-Supervised Defect Detection with Normalizing Flows. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). ISSN 2642-9381, January 2021; 1906–1915.

[7] Williams SE, Roney CH, Connolly A, et al. OpenEP: A Cross-Platform Electroanatomic Mapping Data Format and Analysis Platform for Electrophysiology Research. Frontiers in Physiology February 2021;12. ISSN 1664-042X.

[8] Dinh L, Krueger D, Bengio Y. NICE: Non-linear Independent Components Estimation, April 2015. ArXiv:1410.8516 [cs].

[9] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using Real NVP, February 2017. ArXiv:1605.08803 [cs].

Address for correspondence:

Filippo Uslenghi
Dipartimento di Informatica, Università degli Studi di Milano
Via Celoria 18, 20133, Milan, Italy
filippo.uslenghi@unimi.it