Evidential Deep Learning Model for Atrial Fibrillation Detection from Holter Recordings

Md Moklesur Rahman[1], Massimo Walter Rivolta[1], Pierre Maison-Blanche[2],
Fabio Badilini[3,4], Roberto Sassi[1]

[1] Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
[2] Department of Cardiology, Hôpital Bichat, Paris, France
[3] Center for Biological Research, Department of Cardiology, University of California, San Francisco, USA
[4] AMPS-LLC, New York, USA

Abstract

Deep learning (DL) models have shown promising performances for detecting atrial fibrillation (AF) and atrial flutter (AFL) from electrocardiograms (ECGs) but often suffer from overconfidence and poor probability calibration. Evidential DL (EDL) addresses this by using evidence to parameterize a Dirichlet distribution for uncertainty estimation. The main objective of this study was to develop an EDL model for AF and AFL detection from 2-lead Holter ECG recordings, aiming to estimate uncertainty without incurring additional computational costs compared to traditional softmax-based DL models. In this study, we developed an evidential residual-based DL model, treating predicted probabilities as subjective opinions. The model was trained and tested on a comprehensive dataset of 661 Holter recordings. Our experiments showed that the EDL model achieved recalls of 0.953, 0.838, and 0.934 for detecting AF, AFL, and Non-AF, respectively. The corresponding AUC scores were about 0.980 for AF and Non-AF, and 0.972 for AFL. In terms of confidence estimation, the EDL model exhibited superior performance with an expected calibration error of 0.09, compared to 0.16 for the softmax-based model. These results indicate that EDL models offer enhanced calibration and effectiveness in detecting AF compared to standard softmax models.

1. Introduction

Atrial fibrillation (AF) is a prevalent cardiac arrhythmia characterized by irregular and often rapid heartbeats, which can lead to severe complications such as stroke, heart failure, and increased mortality [1]. Timely and accurate detection of AF is crucial for effective management and intervention. Holter recordings, which capture long-term electrocardiogram (ECG), are particularly useful for identifying intermittent AF episodes that may be missed during short clinical visits [2].

Recent advances in deep learning (DL) have significantly improved AF detection from Holter recordings. However, DL methods often fall short in quantifying prediction uncertainty—a critical factor in medical diagnostics due to the potential consequences of false positives and negatives. Accurate uncertainty quantification is essential for assessing the reliability of model predictions and supporting clinicians in making informed decisions. Epistemic uncertainty, which reflects the model's confidence in its predictions, is especially important in this context [3]. High uncertainty can indicate areas where the model lacks sufficient data or confidence, which is crucial for avoiding misdiagnoses and ensuring that patients receive appropriate care. Conventional approaches, such as Monte Carlo dropout [4] and ensemble methods [5], often use softmax functions to estimate class probabilities, which can lead to overconfident predictions that do not accurately reflect the true uncertainty [6].

To address these limitations, we propose an evidential DL (EDL) model for AF detection from Holter recordings. The EDL model utilizes a variational Dirichlet distribution to capture and quantify the uncertainty associated with predictions [6]. By parameterizing a Dirichlet distribution over categorical output probabilities, the EDL model provides a principled measure of epistemic uncertainty, offering varying levels of confidence in its predictions [6]. This approach not only enhances the model's robustness but also improves its ability to identify uncertain predictions, aiding clinicians in distinguishing between reliable and ambiguous cases.

The primary objectives of this study are to: (i) develop a DL model incorporating evidence-based theory for improved AF detection from Holter recordings, and (ii) eval-

ISSN: 2325-887X DOI: 10.22489/CinC.2024.384

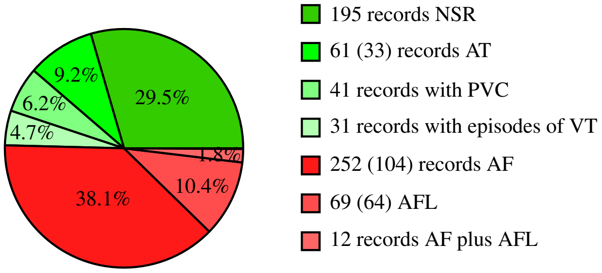

Figure 1: Distribution of patients by records in the dataset: Records with AF/AFL are shown in red, while records without AF/AFL are shown in green. The number of "chronic" records (*i.e.*, entire records under the labeled rhythm) is indicated in parentheses.

uate the advantages of the EDL model over traditional deterministic (softmax-based) DL approaches.

2. Materials and Methods

2.1. Dataset

In this study, we utilized a private dataset of 661 Holter recordings from 661 patients, collected at Groupe Hospitalier Ambroise Paré, Paris, France. Each recording, captured using a Microport Spiderview Holter recorder, lasted approximately 23 hours with a 2-lead system, a sampling rate of 200 Hz, and an amplitude resolution of $10\mu V$. The patient cohort had an average age of 60 years, with 39% of the participants being female.

The dataset included a variety of cardiac conditions: about 50% of the recordings (n=333) featured at least one episode of AF or atrial flutter (AFL), with episodes ranging from brief occurrences to continuous (chronic AF or AFL). The remaining recordings consisted of cases with normal sinus rhythm (NSR, n=195), premature ventricular contractions (PVCs, n=41), atrial tachycardia (AT, n=61), or ventricular tachycardia (VT, n=31). The distribution of these cardiac conditions is illustrated in Figure 1.

For our analysis, we divided the dataset into training (250 records), validation (18 records), and test (393 records) sets. During preprocessing, we applied a third-order zero-phase Butterworth bandpass filter with cutoff frequencies of 0.5 Hz and 40 Hz to correct for baseline wander and minimize power line interference. Each recording was then segmented into 10-second segments without overlap. The number of 10-second segments for the training, validation, and test sets is summarized in Table 1. We considered normal sinus rhythm (NSR) and atrial tachycardia (AT) together as Non-AF due to their distinct heart-rate variability, risk profiles, and treatment strategies compared to AF and AFL.

2.2. Evidential Deep Learning

EDL combines DL with uncertainty quantification using evidence theory, specifically Dempster-Shafer Theory [6, 7]. Unlike softmax-based DL models, which produce point predictions, EDL models aim to explicitly model epistemic uncertainty in predictions in a principled way. In EDL, a set of belief masses $b_k \geq 0$ is assigned to each class $k \in [1, K]$, representing the potential class labels for a given input. These belief masses, along with an overall uncertainty mass $u \geq 0$ (corresponding to an "I don't know" category), must satisfy the constraint: $u + \sum_{k=1}^{K} b_k = 1$. The belief mass b_k for class k and the uncertainty mass u are defined as follows:

$$b_k = \frac{e_k}{S}, \quad u = \frac{K}{S}, \quad S = K + \sum_{k=1}^{K} e_k, \tag{1}$$

where e_k represents the evidence supporting the assignment of the input to class k. This formulation shows that as the evidence e_k for a class increases, the associated uncertainty u decreases. Conversely, in the absence of any evidence, $u = 1$ reflects complete uncertainty. This framework relates to the Dirichlet distribution through the concentration parameter $\alpha_k = e_k + 1$ for class k. The Dirichlet distribution, serving as the conjugate prior to the categorical distribution, enables sampling of probability assignments across all possible classes, denoted as $p \sim \text{Dir}(p \mid \alpha)$ and $\hat{y} \sim \text{Cat}(y \mid p)$. The expected probability for class k is computed as the mean of its corresponding Dirichlet distribution: $\hat{p}_k = \alpha_k / S$.

2.3. Evidential Loss Function

Let $f(x \mid \Theta)$ represent the evidence vector $\mathbf{e} \in \mathbb{R}^K$ predicted by the DL model for an input x, with Θ denoting the model parameters. The parameters of the corresponding Dirichlet distribution are defined as $\alpha = f(x \mid \Theta) + 1$.

Given an observation x_i, let y_i be the one-hot encoded vector representing the true class, where $y_{ij} = 1$ for the true class j and $y_{ik} = 0$ for all $k \neq j$. The Dirichlet distribution $\text{Dir}(\mathbf{p}_i \mid \alpha_i)$ is used to model the uncertainty over the class probabilities $\mathbf{p}_i$. The Dirichlet distribution acts as a prior for the multinomial likelihood of the class labels.

To train the model, we minimize the following evidential loss function for each instance i, which penalizes in-

Table 1: No. of 10-s ECG segments (in thousand units).

Training			Validation			Testing		
Non-AF	AF	AFL	Non-AF	AF	AFL	Non-AF	AF	AFL
1009	848	136	49	66	34	2324	498	124

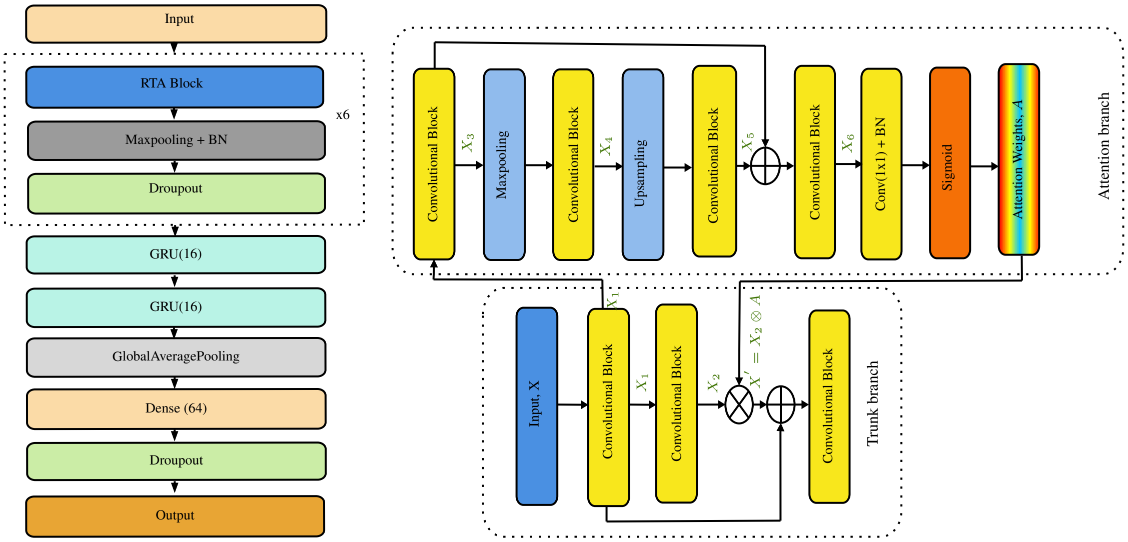

Figure 2: Diagram of the softmax-based DL architecture (a) and RTA block (b).

correct predictions while accounting for uncertainty:

$$L_{\text{ev},i}(\Theta) = \int \|\mathbf{y}_i - \mathbf{p}_i\|^2 \frac{1}{\mathrm{B}(\alpha_\mathrm{i})} \prod_{j=1}^{K} p_{ij}^{(\alpha_{ij}-1)} \, d\mathbf{p}_i$$
$$= \sum_{j=1}^{K} (y_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{S_i + 1}, \tag{2}$$

where $\hat{p}_{ij} = \alpha_{ij}/S_i$ with $S_i = \sum_{j=1}^{K} \alpha_{ij}$ is the total evidence for observation i, and $\mathrm{B}(\alpha_\mathrm{i})$ represents the multivariate Beta function. The term $\|\mathbf{y}_i - \mathbf{p}_i\|^2$ corresponds to the squared error between the ground-truth label and the predicted class probabilities, and the second term regularizes the variance of the predicted probabilities. Additionally, to prevent overconfidence and encourage appropriate uncertainty, a regularization term is introduced by penalizing the divergence between the predicted Dirichlet distribution $\mathrm{Dir}(\mathbf{p}_i \mid \alpha_i)$ and the uniform Dirichlet prior $\mathrm{Dir}(\mathbf{p}_i \mid \mathbf{1})$. The total loss function is thus formulated as:

$$L(\Theta) = \sum_{i=1}^{N} L_{\text{ev},i} + \lambda_t \sum_{i=1}^{N} \mathrm{KL}\left(\mathrm{Dir}(\mathbf{p_i} \mid \tilde{\alpha}_\mathbf{i}) \| \mathrm{Dir}(\mathbf{p_i} \mid \mathbf{1})\right), \tag{3}$$

where λ_t is an annealing coefficient that controls the impact of the regularization term during training. It is defined as $\lambda_t = \min\left(1, \frac{t \cdot \text{Batch size}}{10 \cdot N}\right)$ where t is the current training step, and N represents the total number of training samples. The $\mathrm{KL}(\cdot\|\cdot)$ represents the Kullback-Leibler divergence between two Dirichlet distributions. The adjusted Dirichlet parameters $\tilde{\alpha}_i$ are computed to update the prior belief about the class probabilities by incorporating the true label, effectively removing non-informative evidence and focusing on the true class. They are computed as: $\tilde{\alpha}_i = \mathbf{y_i} + (\mathbf{1} - \mathbf{y_i}) \odot \alpha_\mathbf{i}$. Equation 3 enables the model to simultaneously minimize the prediction error and maintain calibrated uncertainty, particularly in cases where the available evidence is limited.

2.4. Model Development

The architecture of the softmax-based DL model is the one presented in [8] and is illustrated in Figure 2. It integrates a residual temporal attention (RTA) block and a gated recurrent unit (GRU) layer. The GRU layer, added after the RTA block, helps capture temporal patterns in ECG segments, improving AF detection accuracy. The RTA block, repeated six times with kernels starting at 32 and doubling to 128, enhances the model's ability to extract detailed signal representations for better AF detection. The RTA block has two key components explained further below.

Attention Branch: Input X (a 10-second, 2-lead ECG segment) first goes through a convolutional block, producing output X_1. This output feeds into the attention branch, which applies another convolutional block to get X_3. To capture global information, down-sampling (max-pooling) and up-sampling (nearest-neighbor interpolation) are used. A convolutional block then generates a global feature X_6. This global feature is added to X_3 and fused through a residual structure to enhance the feature map. The resulting map undergoes a convolutional block and generates a feature map, X_6. After that, it is fed into a 1×1 convolution layer with sigmoid activation to yield temporal attention weights, A.

Trunk Branch: Input X is processed through a convolutional block to produce X_1, which then goes to the attention branch. X_1 represents the attention weight fea-

Table 2: Performance of softmax-based DL and EDL model for AF detection.

Model	Recall			AUC			ECE
	Non-AF	AF	AFL	Non-AF	AF	AFL	
Softmax-DL	0.951	0.931	0.812	0.981	0.973	0.942	0.16
EDL	0.953	0.934	0.838	0.982	0.977	0.972	0.09

ture map, which is further processed in the trunk branch. X_2 is obtained and multiplied element-wise with the attention map A. This process adjusts feature importance and reduces noise. The refined feature map $X_2 \odot A$ (pointwise product) is combined with X_1 through a residual structure and passed through another convolutional block for the next network layer.

In this study, we compare the performance of an EDL model with the same architecture as a softmax-based DL model, using cross-entropy loss instead of evidential loss. Both models were trained using the following hyperparameters: (i) the Adam optimizer with a learning rate of 0.001, (ii) a batch size of 128, (iii) a total of 50 epochs, and (iv) an early stopping with the patience of 6 was used to end training when no improvement was observed.

3. Results and Discussions

We evaluated the performance of the DL model using recall, area under the receiver operating characteristic curve (AUC) score, and expected calibration error (ECE). Recall and AUC were computed using one-vs-all approach. These metrics assess the model's accuracy and uncertainty in detecting AF.

Table 2 presents a comparative analysis of the performance of both the softmax-based DL model and the EDL model across key metrics. Both models demonstrate strong performance in the Non-AF and AF classes, achieving identical recall scores of approximately 0.95 for Non-AF and 0.93 for AF, indicating their ability to accurately detect Non-AF and AF cases. They also exhibit excellent discriminative power for Non-AF and AF, with nearly identical AUC scores of 0.98 for Non-AF and 0.97 for AF.

A slight difference emerges in the AFL class, where the EDL model outperforms the softmax-DL model, achieving a recall of 0.838 compared to 0.812. Additionally, the EDL model shows superior AUC performance for AFL, with a score of 0.972 compared to 0.942 for the softmax-DL model, highlighting the EDL model's enhanced ability to detect and differentiate AFL cases.

Furthermore, the EDL model demonstrates better calibration, as indicated by its lower ECE of 0.09, compared to the softmax-DL model's ECE of 0.16. This suggests that the EDL model's predicted probabilities are more accurately aligned with the true outcomes.

4. Conclusion

This study presents an EDL model for detecting AF from Holter ECG recordings. Compared to traditional softmax-based models, the EDL model demonstrates superior recall, enhanced discriminative power, and improved calibration. By providing well-calibrated confidence estimates, the EDL model contributes to more reliable clinical decision-making, reducing both false positives and negatives. These advancements position the EDL model as a promising tool for precise AF detection and diagnosis. Future research will focus on validating these findings in larger patient populations and exploring additional clinical applications, particularly considering out-of-distribution datasets.

Acknowledgments

Md Moklesur Rahman acknowledges support from a PhD fellowship funded by Cardiocalm srl, Italy.

References

[1] Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB, Davis EM, Donahue KE, Doubeni CA, Epling JW, Kubik M, et al. Screening for atrial fibrillation: US preventive services task force recommendation statement. JAMA 2022;327(4):360–367.

[2] Rahman MM, Rivolta MW, Badilini F, Sassi R. A systematic survey of data augmentation of ECG signals for ai applications. Sensors 2023;23(11):5237.

[3] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? NeurIPS 2017;30.

[4] Zhang W, Di X, Wei G, Geng S, Fu Z, Hong S. Cardiac arrhythmia classification with rejection of ecg recordings based on uncertainty estimation from deep neural networks. Neural Comput Appl 2024;36(8):4047–4058.

[5] Rahman MM, Rivolta MW, Badilini F, Sassi R. Quantifying uncertainty of a deep learning model for atrial fibrillation detection from ECG signals. In CinC, volume 50. 2023; 1–4.

[6] Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems 2018;31.

[7] Shafer G. Dempster-shafer theory. Encyclopedia of artificial intelligence 1992;1:330–331.

[8] Rahman M, Rivolta MW, Maison-Blanche P, Badilini F, Sassi R. Residual-attention deep learning model for atrial fibrillation detection from holter recordings. Journal of Electrocardiology 2024;84:12. ISSN 0022-0736.

Address for correspondence:

Md Moklesur Rahman
Dipartimento di Informatica, Università degli Studi di Milano
Via Celoria 18, 20133, Milan, Italy
md.rahman@unimi.it