Quantifying Uncertainty of a Deep Learning Model for Atrial Fibrillation Detection from ECG Signals

Md Moklesur Rahman[1], Massimo Walter Rivolta[1], Fabio Badilini[2,3], Roberto Sassi[1]

[1] Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
[2] University of California, San Francisco, USA
[3] AMPS-LLC, New York, USA

Abstract

Recently, deep learning (DL) demonstrated capable to identify atrial fibrillation (AF) from electrocardiograms (ECGs) with significant performance. Nevertheless, these models may present an exaggerated self-confidence in their predictions and showing poor calibration in their output probabilities. In addition, such models cannot quantify the uncertainty of the predictions: a fundamental property in the clinical practice. In this study, we compared two DL models with the same architecture, but the second one had the first and last layers trained using a Bayesian approach, i.e.*, variational inference (VI), allowing the estimate of uncertainty of the predictions. We then compared the models in terms of predictive uncertainty H and Expected Calibration Error (ECE). Our experiments showed that the both models performed very well on the MIT-BIH Atrial Fibrillation dataset (sensitivity and specificity $> 96\%$). The first model proved being i) more confident than the second one (H: 0.006 vs 0.090); and ii) more poorly calibrated (ECE: 0.360 vs 0.028). Despite the computational demand required for using the Bayesian approach in DL, our study demonstrated the importance of quantifying uncertainty of DL-based predictions for AF detection from ECG signals.*

1. Introduction

Atrial fibrillation (AF) is a common cardiac arrhythmia characterized by irregular electrical activity in the atria, leading to an increased risk of stroke and other cardiovascular complications [1]. Early and accurate detection of AF is crucial for timely intervention and improved patient outcomes. With the advancement of artificial intelligence (AI), the potential for automated AF detection using electrocardiogram (ECG) signals has gained significant attention. Recently, deep learning (DL) techniques for AF detection have shown promising performance [2]. Nevertheless, these approaches are deterministic architectures, and hence they lack strategies to measure classification uncertainty.

The presence of uncertainty in DL models can often be attributed to factors associated with data input. This type of uncertainty is commonly referred to as "aleatoric uncertainty" and stems from issues such as noise and imprecise measurements. In addition, there exists another category of uncertainty, known as "epistemic uncertainty", which arises from a lack of knowledge, *e.g.*, DL model architecture, model hyperparameters, limited data [3]. Consequently, performing uncertainty analysis becomes vital to ensure robust classifier, particularly in clinical applications, where low error tolerance is essential. Methods that employ Bayesian modeling have demonstrated favorable results in handling uncertainty analysis for DL models. One such example is the Bayesian Neural Network (BNN), which introduces stochastic components over the network parameters, simulating various possible models with their probability distributions [4]. Despite being computationally more demanding than deterministic models, these architectures provide valuable uncertainty predictions, which may make outputs more reliable and trusted.

The objective of the study was twofold: i) to develop a novel DL architecture designed as a BNN for AF detection from ECG signals; and ii) to determine whether the Bayesian approach would be beneficial, hence justifying the increased computational demand, with respect to a deterministic model.

2. Methods

2.1. Data

In this work, the MIT-BIH Atrial Fibrillation Database (AFDB), which is freely available on Physionet, was utilized [5, 6]. This database contains 2-lead ECG signals from 23 patients sampled at a frequency of 250 Hz. The rhythm annotations within AFDB are classified into four types: AF, AFL (atrial flutter), J (atrioventricular junctional rhythm), and N (sinus rhythm). In this study, the annotations of N were reclassified as "non-AF", while AF

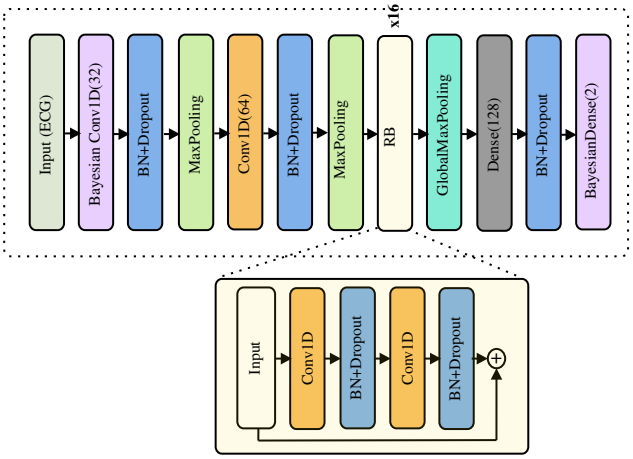

Figure 1. Diagram of the BDL model. BN and RB stand for batch normalization and residual block, respectively.

and AFL were merged as "AF".

2.2. Preprocessing

A third-order zero-phase Butterworth bandpass filter was employed on both leads with cutoff frequencies of 0.5 Hz and 40 Hz to suppress baseline wandering and power-line interference. After filtering, each recording was segmented with a 10 s window without any overlapping. To facilitate robust model training and evaluation, a patient-wise split was employed, dividing the dataset with a 80:20 ratio for training and validation. The final number of training and validation windows was 65655 and 18358, respectively.

2.3. Uncertainty quantification

Uncertainty quantification in BDL models can be achieved using Bayesian approaches. One of these approaches is the BNN model. Let $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ be a training set of N samples with inputs $\mathcal{X} = \{x_1, \cdots, x_N\}$ and targets $\mathcal{Y} = \{y_1, \cdots, y_N\}$, where $x_n \in \mathbb{R}^d$ is the ECG (d-dimensional vector) and $y_n \in \mathbb{R}^C$, where C represents the number of classes which, in this study, was 2 (AF vs non-AF; hereafter encoded with the numbers of 1 and 2, respectively). The predictive distribution for a new sample x^* can be calculated as:

$$p(y^*|x^*, \mathcal{D}) = \int p(y^*|x^*, w)p(w|\mathcal{D})dw, \qquad (1)$$

where y^* can take only the values 1 or 2.

The Monte Carlo method can be employed to approximate the integral in (1). To achieve this, it is necessary to perform T random sampling of the weights $w^{(t)}$ from the posterior distribution $p(w|\mathcal{D})$, where t indicates the sample index. As a result, instead of a single output, T outputs are obtained from the model $\left\{y^{(t)}; 1 \leq t \leq T\right\}$. In this study, we tested several values of T, specifically, 1, 10 and 50. The set $\left\{y^{(t)}\right\}$ can be interpreted as a sample of the predictive distribution. Having at disposal these T probabilities, the final prediction of the BNN model on the input x^* can be computed as the sample mean

$$p(y^*|x^*, \mathcal{D}) \approx \frac{1}{T}\sum_{t=1}^{T} p\left(y^*|x^*, w^{(t)}\right) = p_{y^*}, \qquad (2)$$

while the uncertainty can be estimated by computing, for example, the standard deviation.

This approach allows us to estimate the BNN prediction and uncertainty. However, finding and sampling the posterior distribution for complex models, such as neural networks, is computationally expensive because of their high dimensionality. To address this issue, variational inference (VI) [7], a popular approach for BNNs, was used in our study.

2.4. Variational inference

In VI, instead of directly sampling from the exact posterior distribution $p(w|\mathcal{D})$, a variational distribution $q_\theta(w)$ is utilized, characterized by a parameter vector θ. The values of θ are then learned by minimizing the Kullback–Leibler (KL) divergence between $q_\theta(w)$ and $p(w|\mathcal{D})$, aiming to make them as close as possible. This minimization process is equivalent to maximize the evidence lower bound (ELBO) function $\mathcal{L}(\theta)$, which serves as the objective function to train the model:

$$\mathcal{L}(\theta) = \int q_\theta(w) \ln p(\mathcal{Y}|\mathcal{X}, w)dw - D_{KL}\left(q_\theta(w)\|p(w)\right). \qquad (3)$$

The objective of maximizing the first term in (3) is to encourage a good fit to the data of the distribution $q_\theta(w)$, while minimizing the second term aims to make $q_\theta(w)$ close to the prior distribution $p(w)$. Although determining $q_\theta(w)$ can be intricate in general, a common and straightforward approach is to assume an independent Gaussian distribution $w \sim \mathcal{N}(\mu, \sigma^2)$ for each weight. Furthermore, the prior distribution is commonly taken as a standard normal distribution $w \sim \mathcal{N}(0, 1)$, which we followed suit.

2.5. Model development

The architecture of the non-Bayesian DL (NBDL) model consisted of 36 layers. To make the optimization of such a complex network manageable, we incorporated shortcut connections, similar to the residual network architecture. The network was composed of 16 residual blocks, each containing two convolutional layers. The number of

residual blocks was selected by maximizing the accuracy on the validation set. These convolutional layers had a filter size of 3 and 32×2^k filters, where k was a hyperparameter that starts at 0 and was incremented by 1 every fourth residual block. Additionally, every alternate residual block reduced the input size by a factor of 2 through subsampling. To improve convergence and training stability, we applied ReLU activation function and batch normalization after each convolutional layer. Furthermore, we introduced dropout with a probability of 0.3, to prevent overfitting. Subsequently, a dense layer comprising 128 neurons was employed, followed by ReLU activation, batch normalization, and a dropout layer. Ultimately, a softmax layer was utilized to generate a probability distribution across the two output classes, enabling the detection of AF in the ECG signals.

The BDL model was then created by modifying the first and last layer of NBDL. These two layers were treated using the Bayesian approach and trained by VI. The BDL model architecture was shown in Figure 1.

Both models were trained using: i) the Adam optimizer with a learning rate of 0.001; ii) a batch size of 128; iii) a number of epochs of 100; and iv) an early stopping procedure with patience of 10 as end criterion during training. The Bayesian layers were instead implemented using the TensorFlow Probability package.

2.6. Evaluation

The performance of the models was evaluated using metrics reflecting different aspects.

First, sensitivity (Se; detection of AF) and specificity (Sp) were used to assess the model ability to correctly identify positive and negative instances.

Second, the confidence of the predictions provided by the models was quantified using the well-known predictive uncertainty, which is defined as the Shannon Entropy over the output probabilities, as follows:

$$\mathcal{H}^* = -\sum_{y^*=1}^{C} p_{y^*} \log_2 p_{y^*}, \tag{4}$$

where p_{y^*} is either the final prediction for BDL, as obtained in (2), or the deterministic prediction for NBDL. We quantify the overall entropy $\mathcal{H}$ for each tested model by averaging such entropies across ECGs of the dataset.

Third, we verified whether predictions provided by a model were well-calibrated. A model is well-calibrated when the probability associated with the predicted class label reflects its ground truth correctness likelihood [8]. Here, the calibration was quantified using the expected calibration error (ECE). ECE is typically computed by assigning the samples of a dataset to different bins according to their output probability, and then for each bin, the accuracy is quantified. A model is well-calibrated when class probabilities match accuracy. Formally, the ECE is defined as:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{\#\mathcal{B}_b}{N} \left|\mathrm{acc}(b) - \mathrm{conf}(b)\right| \tag{5}$$

where B refers to the number of bins which is set to 10, N is the total number of samples, $\#\mathcal{B}_b$ represents the number of samples in the bin b, $\mathrm{acc}(b)$ is the accuracy achieved by the model within a given bin, and $\mathrm{conf}(b)$ is the average predicted probability of that bin.

Table 1. Metrics to evaluate the model performance. The validation set is used for the quantification.

Model	Se (%)	Sp (%)	$\mathcal{H}$ (bits)	ECE
NBDL	97.6	96.9	0.006	0.360
BDL ($T = 1$)	98.2	99.0	0.043	0.005
BDL ($T = 10$)	100	100	0.078	0.082
BDL ($T = 50$)	100	100	0.090	0.028

3. Results

Table 1 reports the results for the selected metrics.

All models achieved high recognition rates on the test set, with NBDL showing the lowest Se and Sp (97.6 and 96.9, respectively).

Regarding predictive entropy, NBDL achieved the lowest $\mathcal{H}$ value (0.006), whereas the BDL models, for all T values considered, had a higher entropy, with the second lowest entropy for BDL with $T = 1$ (0.043). The highest entropy was instead obtained by BDL with $T = 50$ (0.090). Moreover, entropy showed an increasing trend with respect to T.

In terms of calibration, NBDL displayed the highest ECE value, while BDL obtained better calibration for the

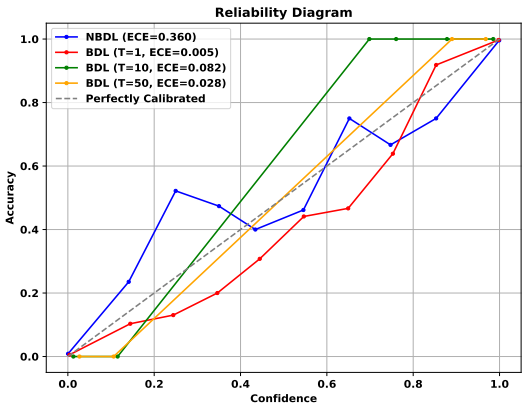

Figure 2. Calibration plot for BDL and NBDL models.

three considered T values. Figure 2 shows the calibration plot depicting confidence vs accuracy, along with the ECE values obtained.

3.1. Discussion

Both models achieved a very high recognition rate. However, this could be due to the fact that the number of residual blocks had been selected on the validation set. Considering the low number of patients of the dataset at disposal, we preferred to keep the evaluation of the performance on unseen data to future works, but we ensured that models had similar performance before quantifying predictive uncertainty and calibration.

As expected, NBDL showed very high confidence in its predictions, as indicated by the low average predictive uncertainty $\mathcal{H}$. Differently, BDL had consistently achieved higher entropy values, which points to a lower confidence. Moreover, the gradual increase in $\mathcal{H}$ throughout the iterative process (increasing T) in the BDL model reflected its ongoing adaptation and exploration of the parameter space, highlighting a greater degree of uncertainty in its predictions.

The proposed architecture had only two layers trained with VI. This choice was a compromise between the possibility of estimating the uncertainty and the computational demand, while keeping the number of parameters of BDL comparable to NBDL.

The results presented in this study suggest that incorporating uncertainty quantification techniques, specifically with the proposed BDL model, can significantly enhance the performance for AF detection in terms of overconfidence and calibration. Moreover, the results have significant implications for the development of robust DL models for AF detection and their integration into clinical practice. Indeed, by providing clinicians with uncertainty measures, BDL enables them to make more informed decisions, possibly improving patient risk stratification, and enhancing overall patient care. Ultimately, this research contributes to the advancement of AI-driven cardiovascular disease management for AF detection and highlights the importance of uncertainty quantification in DL models.

4. Conclusion

In this paper, we proposed a Bayesian DL model for AF detection that quantifies uncertainty by taking into account both epistemic and aleatoric uncertainties. The significance of this work is two-fold: i) it promotes the trust in AI-based predictions of AF detection thanks to uncertainty estimates; and ii) despite the computational demand required, it shows it is still convenient to include a Bayesian approach into the pipeline. Future research should investigate the potential of preprocessing techniques, such as augmentation and resampling, to enhance the performance of uncertainty quantification in Bayesian DL models.

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] Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, et al. A survey of uncertainty in deep neural networks. Artificial Intelligence Review 2023;1–77.

[4] Hernández-Lobato JM, Adams R. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In International Conference on Machine Learning. 2015; 1861–1869.

[5] Moody GB, Mark RG. A new method for detecting atrial fibrillation using R-R intervals. Computing in Cardiology 1983;10:227–230.

[6] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, Physiotoolkit, and Physionet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[7] Hoffman MD, Blei DM, Wang C, Paisley J. Stochastic variational inference. Journal of Machine Learning Research 2013;.

[8] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In International Conference on Machine Learning. 2017; 1321–1330.

Address for correspondence:

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