Does Synthetic Data Achieve Better Generalization on AF Classification?

Jeries Saleh[1]

[1] Technion - Institute of Technology, Haifa, IL

Abstract

High-quality ECG datasets of Atrial fibrillation (AF) patients is crucial for building AF detection and classification. Generative models have emerged as a tool to produce synthetic data to enrich existing datasets. However, reaching a low error doesn't imply that the generative model is good enough. Therefore, using Proposition 1, we test the quality of the synthetic data i.e. the generative model. At last, we show that synthetic augmentation in medical data does not necessarily improve performance in AF classification task.

1. Introduction

Atrial fibrillation (AF) is one of the most prevalent cardiac arrhythmia. Accurate detection and classification of AF using electrocardiogram (ECG) data is crucial for timely intervention. However, obtaining large, diverse, and high-quality ECG datasets is challenging due to factors like data scarcity, imbalanced class distributions, and inherent noise in ECG signals. These issues can limit the performance of machine learning models and lead to poor models generalization.

One promising approach to address this limitation is the use of generative models to synthesize realistic ECG data. Recent advances in synthetic data generation improved accuracy on image segmentation tasks [1]. Therefore, we designed this study to explore augmenting real ECG datasets with synthetic ECG samples. To visualize this, we first present a simplified depiction of data distributions Fig. 1. This visualization depicts the idea that some synthetic samples do not resemble the authentic real data. Therefore, synthetic data should be carefully generated.

Papers on diffusion models, one type of generative models, are exponentially increasing. While they excel in image generation tasks, we will see that medical data remain challenging to synthesize well, even if we reach a good score [2].

Specifically, we leveraged a recently published diffusion model to generate the synthetic samples. This model is known as Structured State Space Diffusion Model (SSSD) [3]. We trained it on a publicly published dataset from PhysioNet called the Long Term AF database [4], which consists of annotated recordings of 84 patients for 24-25 hours. In parallel, we trained an AF classifier on the same dataset. Then, we aim to use Proposition 1 to assess the quality of the synthetic data, and then through a comparative analysis, explore the ability to improve the generalization of AF classification models.

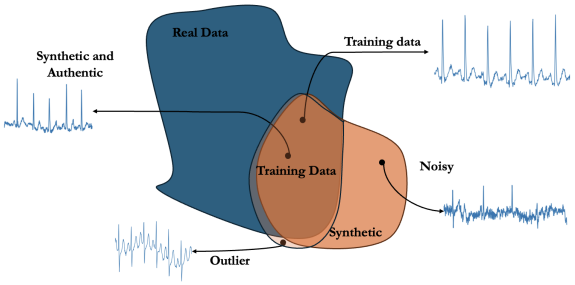

Figure 1. **Depiction of the data distributions.** The training data is mostly contained in the real data, but may contain some outliers that exhibit artifacts. The real data includes the unmeasured data that is not present in the training set. While the synthetic set might contain authentic samples, it contains unreal ones.

2. Methods

2.1. The Dataset and Preprocessing

Long Term AF Database (LTAFDB) consists of 2-Lead ECG recordings 84 patients. Most of the beats (peaks) in the recordings are annotated as normal sinus rhythm or AF. We divided each recording to segments. But some segments don't resemble an ECG form, and some have missing annotations in the beginning or in the middle. Therefore, we first divide each recording to segments starting from the first annotated beat. Then, we check whether the time difference between annotations (peaks) is consistent. As a result, we excluded records $\{30, 74, 100, 113\}$.

The dataset is highly imbalanced with positive to negative ratio of 9.38. It was split to train-test in a 70-30 split.

2.2. Models

Structured State Space Diffusion Model (SSSD) is a generative model that was initially proposed to generate synthetic ECG samples [3]. The idea is to leverage structured state space models (SSSMs) in the model's architecture, which link between an input sequence $u(t)$ and an output sequence $y(t)$ using a hidden state $x(t)$ under the following continuous dynamics,

$$x'(t) = Ax(t) + Bu(t), y(t) = Cx(t)$$

In short, the continuous ODE system above can be discretized and written as a convolution, and using a specific initialization of the matrix A, allows for building the S4 layer, which efficiently captures long-range dependencies in time series data. We refer for the original paper for a detailed explanation [5].
Diffusion models are multiple latent models $p_\theta(x_0)$ that learn a data sample x_0 from pure Gaussian noise $x_T = N(0, I)$ in a process called the backward process which is parametrized as follows,

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

The forward process propagates a data sample x_t to noise according to a variance scheduler,

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

AF classifier is a non-causal classification model that uses Gated Recurrent Units (GRU) to classify 15s segments of ECG as normal sinus rhythm or ones with AF. The short 15s segment is challenging for any classifier to accurately classify arrhythmias as indicated by [6], and there doesn't exist, up to our knowledge, a classifier capable of successfully solving this.

The classifier was trained for 300 epochs with learning rate of 0.0001 on AdamW optimzer, with early stopping.

2.3. Training and Inference Procedure

First, each recording is segmented into 15 seconds segments ($15[s] \times 128[Hz] = 1920$ points), as this is the longest sequence length that can be trained on our GPUs. Then, the dataset was split into 80-20 train-test split.
We trained the SSSD model and the AF classifier, in parallel, on LTAFDB on generation and classification tasks respectively. Then we tested the performance of the AF classifier on the test set, and using the trained SSSD model, we generated a synthetic dataset Fig. 2.

2.4. Set Quality Assessment

Here, we suggest a proposition that can be used for quality assessment of the generated set.

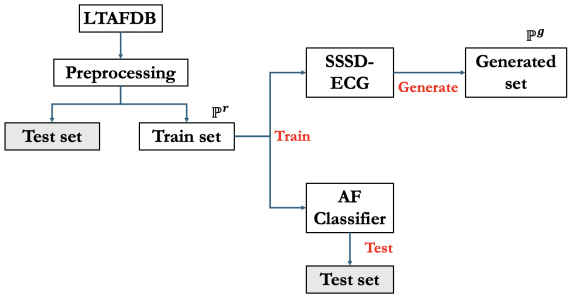

Figure 2. **Illustration of the training scheme**

Proposition 1. *Let P_r and P_g be probability distributions over input-label pairs $(x, y) \in \mathbb{R}^L \times \{0, 1\}$. Suppose the 1-Wasserstein distance between P_r and P_g bounded by ε:*

$$W_1(P_r, P_g) \leq \varepsilon.$$

Let $f : \mathbb{R}^L \to \{0, 1\}$ be a classifier, and $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ be a loss function that is Lipschitz continuous with Lipschitz constant l with respect to its inputs. Then:

$$|E_{(x,y)\sim P_r}[L(f(x), y)] - E_{(x,y)\sim P_g}[L(f(x), y)]| \leq l\varepsilon$$

Proof: The space $\mathbb{R} \times \{0, 1\}$ is a metric space with the metric being the sum of the euclidean distance on $\mathbb{R}$ and the discrete distance on $\{0, 1\}$. Then, using Kantorovich–Rubinstein duality, for all Lipschitz continuous functions $\phi : \mathcal{M} \to \mathbb{R}$ with Lipschitz constant at most l, the following holds:

$$|E_{x\sim\mu}[\phi(x)] - E_{y\sim\nu}[\phi(y)]| \leq lW_1(\mu, \nu)$$

Setting $\mathcal{M} = \mathbb{R} \times \{0, 1\}$, $\phi = L(\cdot, \cdot)$, and since $W_1(P_r, P_g) \leq \varepsilon$, we have:

$$\left|E_{(x,y)\sim P_r}[L(f(x), y)] - E_{(x,y)\sim P_g}[L(f(x), y)]\right| \leq l\,\varepsilon.$$

□

The term $E_{(x,y)\sim P_r}[L(f(x), y)]$ is estimated as the average of the expression over the test set, where $f(x)$ minimizes $L(f(x), y)$ on the training set. In our experiment setting, we set the loss function to be *BCEWithLogitsLoss*, which is Lipschitz continuous (Appendix 1). Therefore, we can use the proposition in two ways.
1. One way is by calculating the 1-Wasserstein distance between two distributions: P_r, the real distribution, and P_g, the generated set distribution. If the distance is low, then an accurate classifier would have a similar performance on the generated set, and then the set can be faithfully used in data augmentation.
2. However, the calculation of 1-Wasserstein distance can be computationally inefficient relative to inference of a trained AF classifier. Therefore, we can use the proposition by measuring the performance of the classifier on the generated set to conclude whether the generated samples are of good quality.

3. Results and Discussions

3.1. SSSD-ECG Performance

The ECG generative model SSSD-ECG is trained using MSE loss, and by training on the LTAFDB, it reaches an MSE loss of **0.002**. Although this is a good score but that does not necessarily lead to a good generative model [2]. See (Appendix 2) for examples of generated samples.

3.2. AF Classifier Performance

On LTAFDB: The performance of the classifier, as shown in Table 1, indicates a significant imbalance in the precision and recall trade-off. The low Fβ-score and Gβ-score suggest that the classifier has difficulty handling the positive class effectively. This is primarily due to the high imbalance in the dataset, where the positive class is underrepresented compared to the negative class. The model likely prioritizes minimizing false negatives over false positives, which may not be optimal for applications where false positives need to be controlled.

Metric	Score (LTAFDB)	Score (Synthetic)
Fβ-Score	0.38	**0.905**
Gβ-Mean Score	0.109	**0.656**
Macro AUC	**0.883**	0.681

Table 1. Performance Metrics for the AF Classifier on LTAFDB and synthetic set

On the synthetic set: Using the diffusion model, we generated 6300 AF classified samples, and 3300 healthy classified samples.

The performance of the classifier, as shown in Table 1, doesn't follow the same trend which we saw on the test set on LTAFDB. Using Proposition 1, we conclude that the generated samples are not of high quality.

3.3. The Best Synthetic Candidates

The AUC result on the synthetic set is significantly lower than that on the real dataset. Therefore, we explored the generated samples as seen in Appendix 2. We notice that the performance on 'healthy' labelled generated samples differs from 'AF' labelled generated samples, which explains the resulting metrics in Table 1.

Label	Accuracy
Healthy	0.2924
AF	0.895

Table 2. Classifier accuracy on the two different labels

Based on these results, and Proposition 1, the best synthetic candidates are the ones which were classified correctly by the classifier. However, if the generated data is small compared to the original dataset, we don't expect it to improve very much. Also, if the bottleneck is the model and not the data then data augmentation shouldn't help.

3.4. Does Synthetic Sets Help?

Based on the previous section, we simulate a condition where we have scarce data from the LTAFDB dataset which we augment with generated samples. The results are shown in Table 3

Metric	Score (w/out DA)	Score (with DA)
Fβ-Score	0.168	**0.21**
Gβ-Mean Score	0.0389	**0.04**
Macro AUC	0.772	0.63

Table 3. Classifier accuracy on the two different labels

Therefore, the answer to the title's question is that we cannot show that augmenting with synthetic samples help.

4. Conclusion

Generated synthetic datasets have been suggested to increase the variety of samples, as it was found that training on synthetic and testing on real does somehow work [3]. However, here we explore this in detail *for the first time*, and we show that synthetic datasets in medical data should be used carefully. In addition, we found a way to test the quality of generated samples given a classifier, and here we show that a good score doesn't lead to good quality samples, as was discussed in [2].

References

[1] Alimisis P. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions. Computer Vision and Pattern Recognition Jul. 2024;ArXiv:2407.04103.

[2] Li S, Chen S, Li Q. A good score does not lead to a good generative model, 2024.

[3] Alcaraz JML, Strodthoff N. Diffusion-based conditional ecg generation with structured state space models. Computers in Biology and Medicine 2023;163:107115. ISSN 0010-4825.

[4] Petrutiu S, Sahakian AV, Swiryn S. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. EP Europace 05 2007;9(7):466–470. ISSN 1099-5129.

[5] Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. In The International Conference on Learning Representations (ICLR). 2022; .

[6] Keidar N, Elul Y, Schuster A, Yaniv Y. Visualizing and quantifying irregular heart rate irregularities to identify atrial fibrillation events. Frontiers in Physiology 2021;12. ISSN 1664-042X.

5. Appendices

5.1. Appendix I:

BCEWithLogits is a Lipschitz Continuous Function: Using *BCEWithLogitsLoss*. For $z \in \mathbb{R}$ and $y \in \{0, 1\}$, the loss function can be expressed as:

$$L(z, y) = \log(1 + e^{-z}) + (1 - y)z.$$

Given the metric on $\mathbb{R} \times \{0, 1\}$:

$$d((z_1, y_1), (z_2, y_2)) = d_1(z_1, z_2) + d_2(y_1, y_2),$$

where d_1 is the Euclidean distance on $\mathbb{R}$, and d_2 is the discrete distance on $\{0, 1\}$ with:

$$d_2(y_1, y_2) = \begin{cases} 0, & \text{if } y_1 = y_2 \\ 1, & \text{if } y_1 \neq y_2. \end{cases}$$

Case 1: $y_1 = y_2$. Observe,

$$\left| \frac{\partial L}{\partial z} \right| = |\sigma(z) - y| \leq 1,$$

Thus, for $y_1 = y_2 = y$, we have:

$$|L(z_1, y) - L(z_2, y)| \leq |z_1 - z_2|.$$

This satisfies the Lipschitz condition with $K = 1$.
Case 2: $y_1 \neq y_2$, the distance becomes,

$$d((z_1, y_1), (z_2, y_2)) = |z_1 - z_2| + 1$$

In this case, we analyze:

$$|L(z_1, 1) - L(z_2, 0)| = \left|\log(1 + e^{-z_1}) - \log(1 + e^{z_2})\right|$$

A suitable choice of $K = \log 2$ works, since for $y = 0$, $z_2 < 0$ and for $y = 1$, $z_1 > 0$. Thus, for $y_1 \neq y_2$, we have:

$$|L(z_1, 1) - L(z_2, 0)| \leq \log(2)[|z_1 - z_2| + 1],$$

which satisfies the Lipschitz condition with $K = \log 2$.

5.2. Appendix II:

Randomly Picked Generated ECGs from SSSD-ECG on LTAFDB.

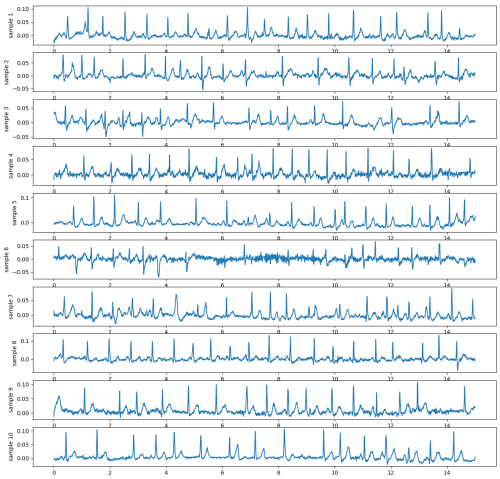

Figure 3. **Example of healthy generated samples**

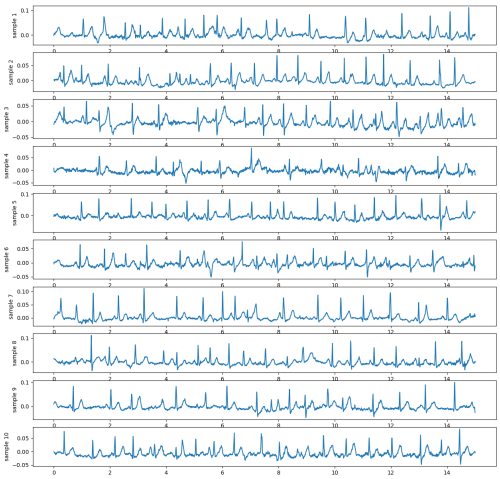

Figure 4. **Example of AF generated samples**