GAN-based Synthetic Heart ECG Data

Kevin Juandi [1]

[1]University of Colorado

Introduction

Machine learning models have shown remarkable capabilities, often outperforming medical experts in various tasks. However, to reach this level of performance, they typically require large, high-quality datasets. Unfortunately, obtaining such datasets can be challenging due to privacy concerns, regulatory restrictions, and the time-consuming process of expert annotation. This is where synthetic data comes into play. By simulating realistic and diverse cases, synthetic data helps fill gaps in underrepresented conditions and demographics, ultimately enhancing the robustness and generalization of models while protecting patient privacy.

Regulatory Barriers and Privacy Risks in Data Sharing

Privacy Laws Restrict Collaboration
"GDPR, HIPAA, and other regulations block cross-institutional medical data sharing, creating fragmented, siloed datasets."

Re-identification Risks
"Even anonymized data can be reverse-engineered, exposing patient identities and violating compliance."

Biased, Non-Generalizable Models
"Models trained on localized data might fail for underrepresented demographics (e.g., ethnic minorities, rare arrhythmia)."

$ High Costs of Compliance
"Legal and technical safeguards for sharing real data strain healthcare budgets and slow innovation."

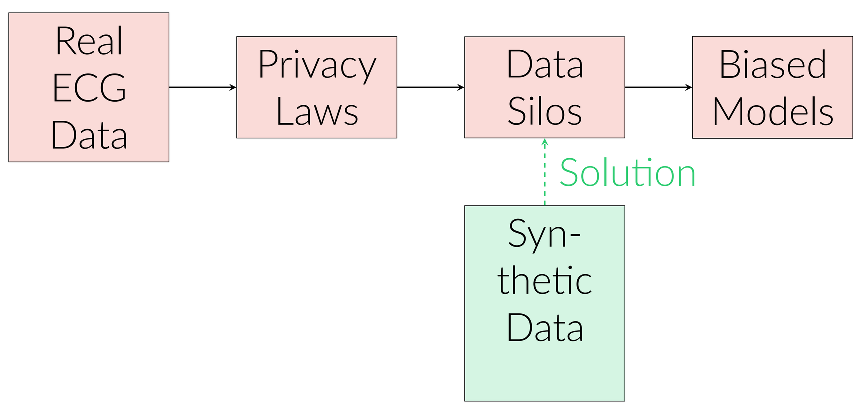


Figure 1. Current barriers vs. proposed solution.

Key Insight

The bottleneck isn't data scarcity—it's safely using existing data. Our GAN-generated ECG sidesteps privacy risks while improving detection accuracy.

MIT-BIH Arrhythmia Dataset

Original Source: MIT-BIH Arrhythmia Database (Moody & Mark, 2001)
Preprocessed Version: Kachuee et al. (2018) heartbeat segmentation
Patients: 47 subjects (selected records with clean signals)
Classes: 5 heartbeat types [a]
Samples: 109,446 heartbeats with R-peak alignment

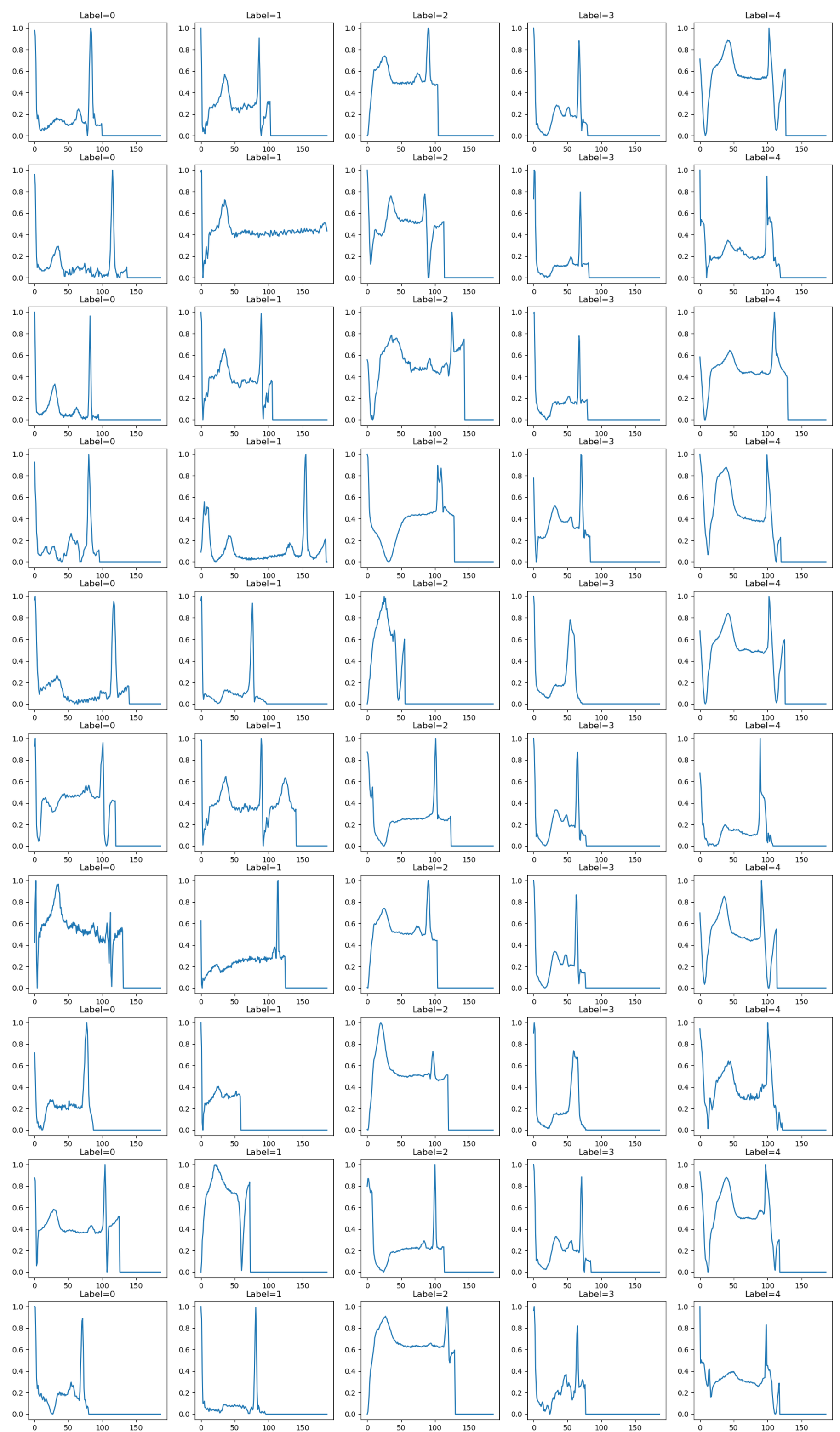


Figure 2. Preprocessed heartbeat sample from Kachuee et al. (2018)

[a]Normal (N), Supraventricular (S), Ventricular (V), Fusion (F), Unknown (Q)

cGAN Model Architecture

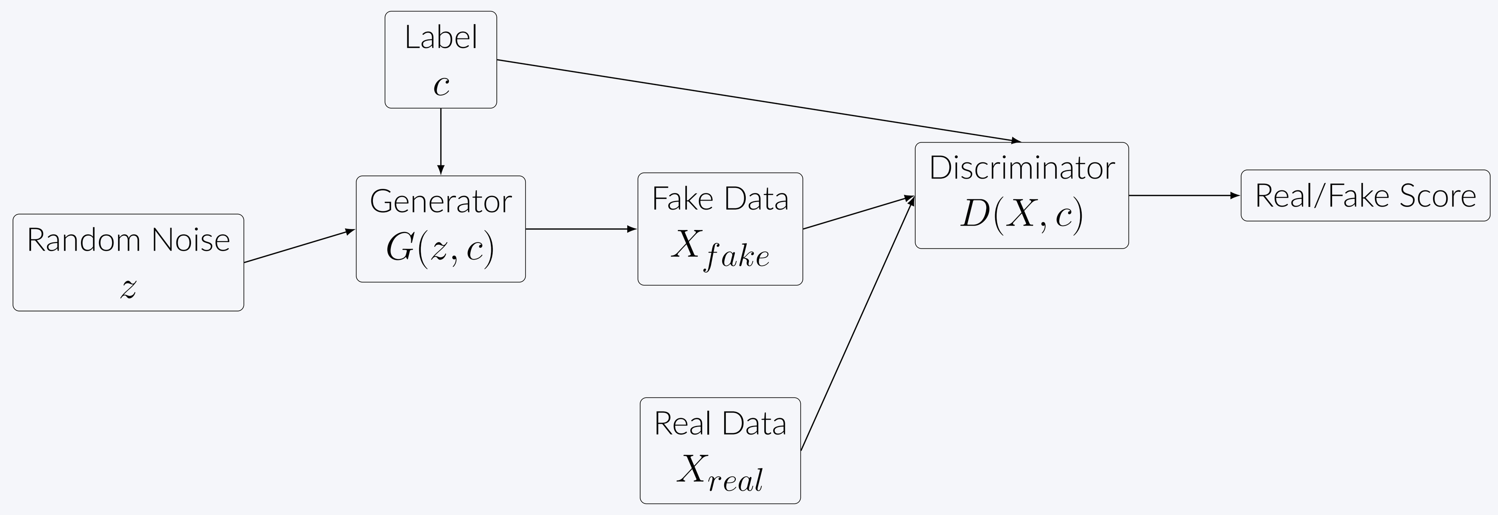


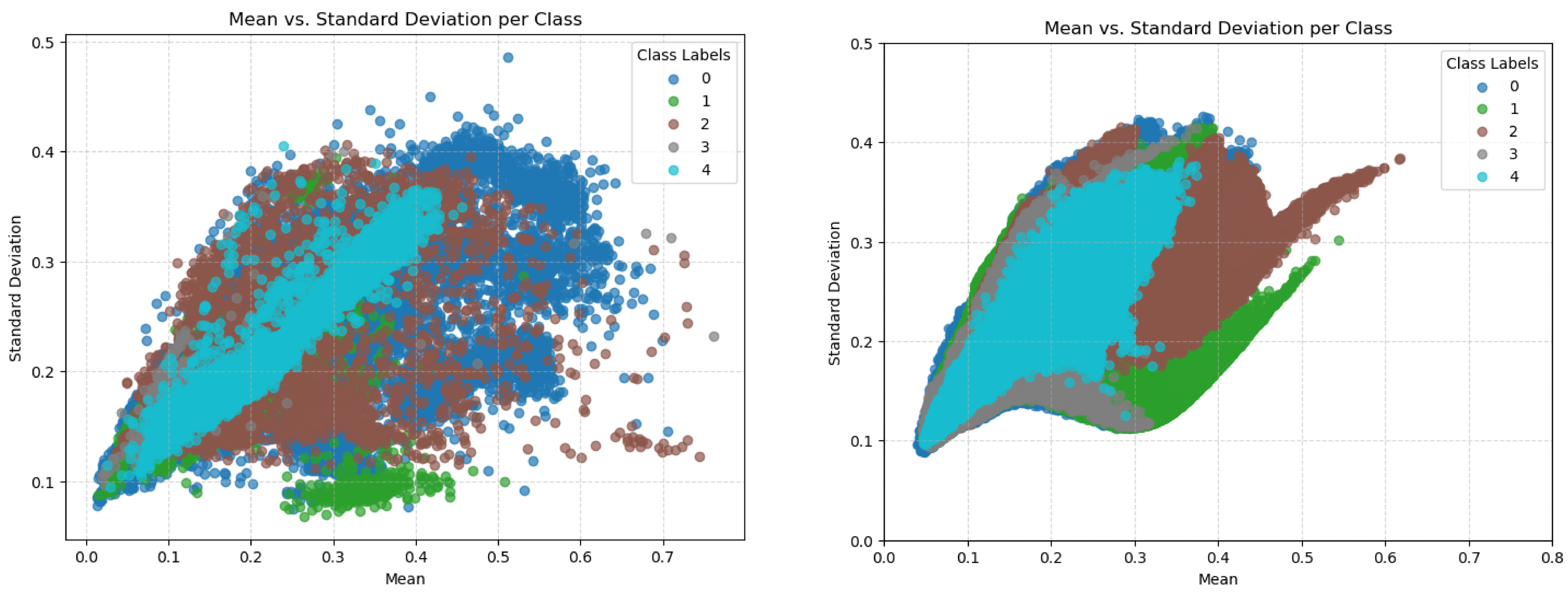
Figure 3. Conditional GAN (cGAN) architecture

This architecture represents a **Conditional Generative Adversarial Network (cGAN)**, where both the **Generator** and **Discriminator** are conditioned on an additional input c (e.g., a class label or structured data).

Generator $G(z,c)$: Takes random noise z and condition c to generate synthetic data X_{fake}.
Real & Fake Data: The Generator's output (X_{fake}) is compared against real data (X_{real}).
Discriminator $D(X,c)$: Evaluates whether input data (real or fake) is authentic while considering the condition c.
Training Objective:
The **Generator** tries to **fool the Discriminator** into classifying fake data as real.
The **Discriminator** learns to **distinguish real from fake** while ensuring the generated data aligns with c.

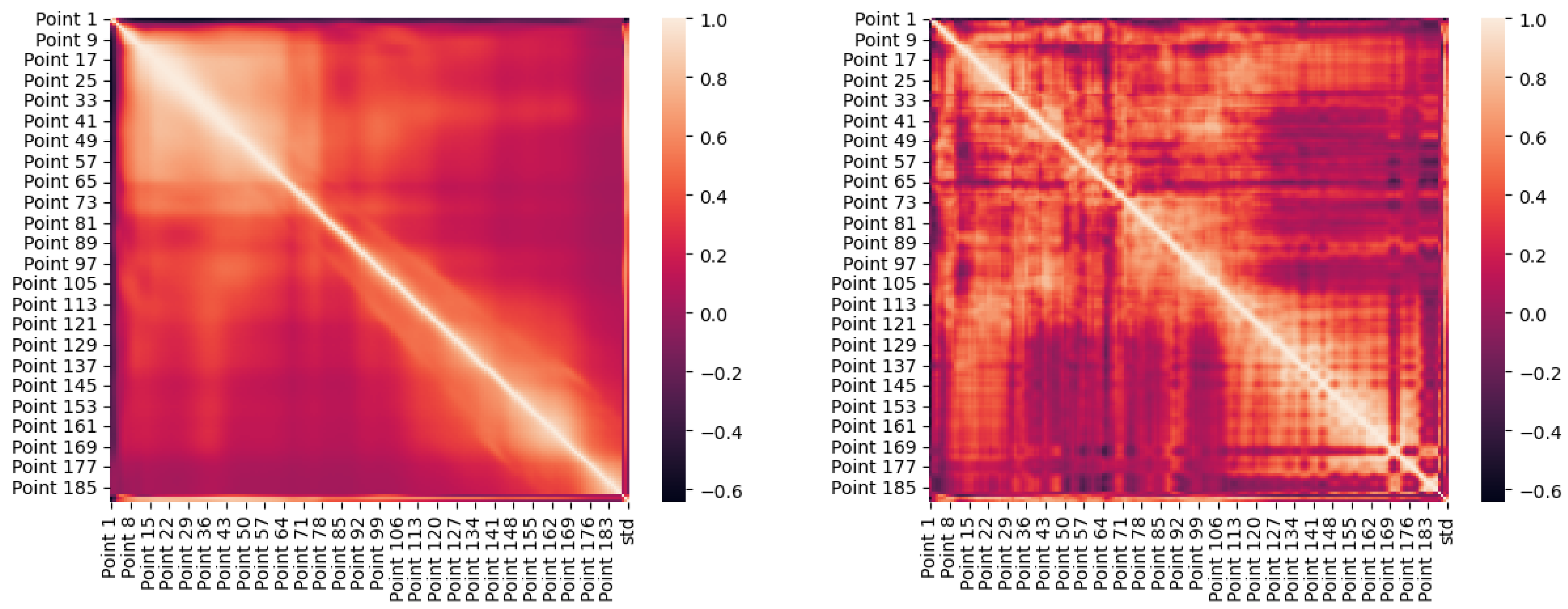
This conditioning mechanism enhances **control over generated outputs**, making cGANs useful for **image synthesis, text generation, and structured data generation**.

Synthetic Data Quality



(a) Real ECG Beats (b) Synthetic ECG Beats

Figure 4. Mean-Std Graph: Real vs. Fake



(a) Real ECG Beats (b) Synthetic ECG Beats

Figure 5. Correlation Heatmap: Real vs. Fake

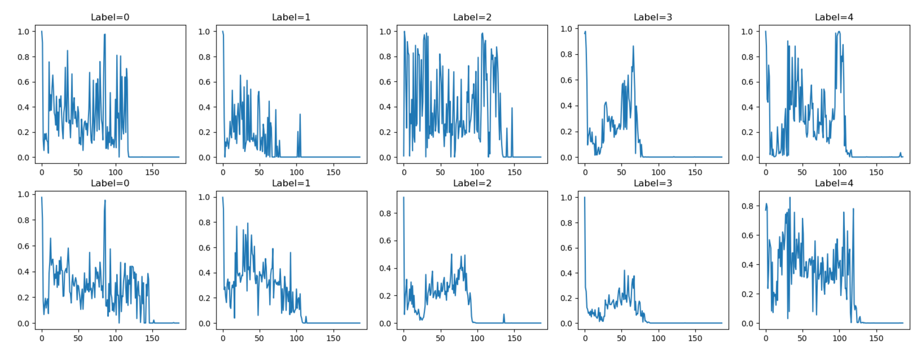


Figure 6. Sample of Generated ECG Data

While the results show promise, further refinements are needed. The Savitzky-Golay filter could aid in denoising, but preprocessing challenges limit its effectiveness. Preprocessing the training data to sinus rhythm might be beneficial.

References

[1] Edmond Adib, Fatemeh Afghah, and John J. Prevost. Synthetic ECG signal generation using generative neural networks. *arXiv preprint*, 2021.

[2] Vladimir Bok and Jakub Langr. *GANs in Action: Deep Learning with Generative Adversarial Networks*. Manning Publications, 2019.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 2672–2680, 2014. Seminal GAN paper.

[4] Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. ECG heartbeat classification: A deep transferable representation. In *IEEE International Conference on Healthcare Informatics (ICHI)*, pages 443–444, 2018.

[5] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint*, 2014. Seminal paper on conditional GANs.

[6] George B Moody and Roger G Mark. The MIT-BIH arrhythmia database. *Circulation*, 101(23):e215–e220, 2001.

kevin.juandi@colorado.edu