Evaluating the Quality of CycleGAN Generated ECG Data for Myocardial Infarction Classification

Sara Battiston[1], Roberto Sassi[1], Massimo W Rivolta[1]

[1] Department of Computer Science, Università degli Studi di Milano, Italy

Abstract

The demand for extensive annotated datasets in ECG interpretation has led to the development of synthetic datasets using generative neural networks. Our study is aimed at assessing the quality of synthetic ECGs generated via a CycleGAN network by means of visual inspection (confidence bands and UMAP 2D plots), GAN-specific evaluation methods (GAN-train and GAN-test scoring), and statistical tests comparing ST segment amplitudes (modified Hotelling T-squared test). To this goal, we utilized a selection of 12-lead ECGs from the PTBXL dataset (available on Physionet) falling under three conditions: normal sinus rhythm, anteroseptal myocardial infarction and inferior myocardial infarction. Through the CycleGAN network we generated synthetic ECGs and compared them with the original ones. The qualitative analysis, by means of plots, showed that there was a difference in the distributions of real and synthetic data. The GAN-train/test method provided results confirming this conclusion. Lastly, the ST-segments analysis showed distributions which were dissimilar among all the conditions. In conclusion, our work demonstrated that generative networks developed in the context of image processing cannot be simply adapted to augment ECG datasets, and that proper care should be enforced to verify the quality of the generated signals, before utilising such data in applications.

1. Introduction

In the field of deep learning for electrocardiogram (ECG) interpretation, the demand for extensive annotated datasets for model training often surpasses their availability. Consequently, there is a growing trend towards the development and use of synthetic datasets, or "ECG augmentation", using statistical and deep learning methodologies [1]. Unquestionably, assessing the quality of the generated ECGs is critical for their reliable application.

In the existing literature, several techniques have been employed to check the quality of synthetic electrocardiograms, such as replacing real beats with synthetic ones in various classification tasks [2, 3], evaluating spatial distances between ECGs with appropriate measures [4], and also visual inspection [5].

Our study aimed at evaluating the quality of synthetic ECGs obtained via style transfer and implemented through the CycleGAN methodology [6]. We generated multi-lead ECG signals of normal sinus rhythm, antero-septal myocardial infarction and inferior myocardial infarction. Quality was assessed through four techniques involving: i) two methodologies of visual inspection; ii) a specific method introduced to evaluate GAN-generated data; and iii) a multivariate statistical test specifically designed to compare ST segment's amplitudes.

2. Methods

2.1. Dataset

The dataset utilized was part of the PTB-XL ECG dataset [7, 8] which contained standard 12-lead ECGs of 10 seconds, sampled at 100 Hz, from 18885 different patients. We selected a sub-sample made of 10776 ECGs, whose diagnoses were: healthy patients (NORM, 80%), antero-septal myocardial infarction (ASMI, 12%) and inferior myocardial infarction (IMI, 8%).

We applied to each of the signals a zerophase Butterworth pass-band filter of order 3, with low and high cutoff frequencies of 0.67 Hz and 15 Hz, respectively. Heartbeats were identified via the gqrs detector contained in the WFDB library [8, 9]. Then each signal was split into windows of length 0.40 s (from -100 ms to 300 ms with respect to each detected beat), such that each included only one QRS complex and T-wave. In order to have a balanced dataset, we subsampled at random with no replacement the NORM class, as it was significantly larger that the other two. The final dataset comprised 35353 0.40 s 12-lead ECG signals (34% NORM, 36% ASMI, 30% IMI).

2.2. CycleGAN

Style transfer refers to a set of techniques meant to convert an input data from one source domain to a target one.

In this study, we implemented style transfer to mutate normal ECGs to myocardial infarction signals and viceversa. One of the techniques allowing so is the well-known CycleGAN approach [6], which aims to learn a pair of inverse mapping functions F and G between the source domain X and target domain Y and vice versa, such that $G(F(X)) \approx X$ and $F(G(Y)) \approx Y$. Such mappings are determined by means of two Generative Adversarial Networks (GANs) [10] trained concurrently. In our context, such functions were meant to translate ECGs from one condition to another one and viceversa. We chose CycleGAN because it allows us to generate synthetic samples morphed directly from real ones.

In this study, we implemented two CycleGANs to perform style transfer back and forth from NORM to ASMI and NORM to IMI. Therefore, a total of four generators were trained.

3. Experiments

3.1. Training and Data Generation

Both CycleGANs were trained for 100 epochs with batch size of 128. The generator architectures were the same used in [6], but the input size was adapted to work with signals of shape 40×12 (12 leads of 0.4 s). The source-target pairs of datasets used were NORM-ASMI and NORM-IMI, and were randomly split into train and validation sets (70% and 30%, respectively). Once both generators were available, the validation sets were also used to generate the synthetic datasets for our experiments. We generated ASMI and IMI ECGs from NORM signals, and NORM ECGs from both ASMI and IMI signals, obtaining a total of 10606 synthetic signals (3564 ASMIgen, 3564 IMIgen, 7042 NORMgen).

3.2. Visual Inspection

To gain a qualitative insight into the quality of the ECGs generated via CycleGAN's generators, we visually inspected the data. To do so, we randomly sampled 2500 ECGs from both synthetic and real datasets, and plotted together their corresponding 95% confidence band for each ECG lead. In this way, we were able to assess the resemblance between the variability of the generated and real data for each lead. A second visual inspection was performed by a plot of a 2D representation of the real and generated signals, obtained via UMAP [11] dimensionality reduction technique (number of neighbours $= 15$ and distance $= \mathrm{euclidean}$). Specifically, we fitted the UMAP reducer on the real ECG signals, and then plotted a 2D density heatmap from the 2D real data. Upon this heatmap we further plotted the scatter points of the 2D synthetic ECGs.

3.3. GAN-train and GAN-test Scores

In order to quantify the goodness of the CycleGAN generated data we exploited the GAN-train and GAN-test technique [12]. Briefly, this technique provides an evaluation method which works by computing two scores called GAN-train and GAN-test. The scores are obtained by first training a classifier on synthetic data (GAN-train classifier) and computing its accuracy with a test set of real samples (GAN-train score), then training a second classifier (GAN-test classifier) on real data and recording its accuracy score on a synthetic test set (GAN-test score). Both the GAN-train and GAN-test scores are finally compared to a baseline which is the accuracy of the first and second classifier computed on a synthetic data test set (baseline for the GAN-train score) and the real data test set (baseline for the GAN-test score), respectively. The GAN-train score is supposed to provide a measure of how diverse the generated data are, while the GAN-test score indicates how close the generated data are to the original data manifold. To this end, we used as ECG classifier a CNN adapted from [13], where we modified the input layer to match our input ECGs and removed two out of the three ResNet blocks. Both classifiers were trained to distinguish the three conditions (NORM, ASMI, IMI). For the training of the GAN-train classifier, we split the synthetic dataset (10606 signals) into train-test with a $85\% - 15\%$ split, and trained the network for 4 epochs since no improvements were observed afterwards. For the GAN-test classifier, we similarly split the real dataset (35353 signals) into train-test with a $85\% - 15\%$ split, and trained the network for 25 epochs. The accuracies on the test sets were used as GAN-train and GAN-test scores.

3.4. Evaluation Metrics on ST Segments

Since myocardial infarction affects the ST-segment, here we set forth to assess the quality of the generated ECGs by evaluating their ST-segment's amplitude (elevation). In particular, from each ECG, we selected by visual inspection the ST-segment as a sub-window of length 80 ms. Then, for each lead, we computed the average value in millivolt as representative of ST segment's amplitude (elevation). We finally compared the population of the averages of such amplitude between the real and generated data of corresponding conditions by means of a modified Hotelling's T-square test for different covariances. To this end, we computed the simultaneous 95% intervals, as follows:

$$\mu_R^\ell - \mu_G^\ell \pm \sqrt{\chi^2_{12,\alpha}\left(\frac{s^2_{R,\ell}}{n_R} + \frac{s^2_{G,\ell}}{n_G}\right)} \qquad (1)$$

where μ_R^ℓ and μ_R^ℓ are the mean of the ST-segments for real and generated data respectively, $s^2_{R,\ell}$ and $s^2_{G,\ell}$ are the sample variances of the ST-segments for real and generated data respectively, ℓ is the ECG lead, $1-\alpha = 95\%$ is the significance level, n_R and n_G are the numbers of the real and generated samples respectively. When zero (no difference in elevation) was contained in the confidence intervals, on average, real and synthetic data had statistically comparable ST segment's amplitudes.

4. Results

For each of ASMIgen, IMIgen, NORMgen, we built and visually inspected the 95% confidence band plots, as explained in section 3.2. We reported an example for the ASMI class in fig. 1 and the corresponding heatmap-scatter plots in fig. 2. Synthetic signals generally differed by visual inspection from the corresponding genuine ones. In particular, the tails of the distributions were overlapping but not their bulk.

We then computed the GAN-train and GAN-test scores and collected the baseline accuracy for each. In particular for the GAN-train classifier we recorded a baseline accuracy of 98%, while the GAN-train score was 62%. Similarly, for the GAN-test classifier baseline accuracy was 89%, and the GAN-test score 81%.

We finally performed the ST-segments analysis described in section 3.4. None of the differences between the ST-segment amplitudes of real and generated data were statistically close to zero. A radar plot for the ASMI-ASMIgen ST-segments amplitudes is reported in figure 3.

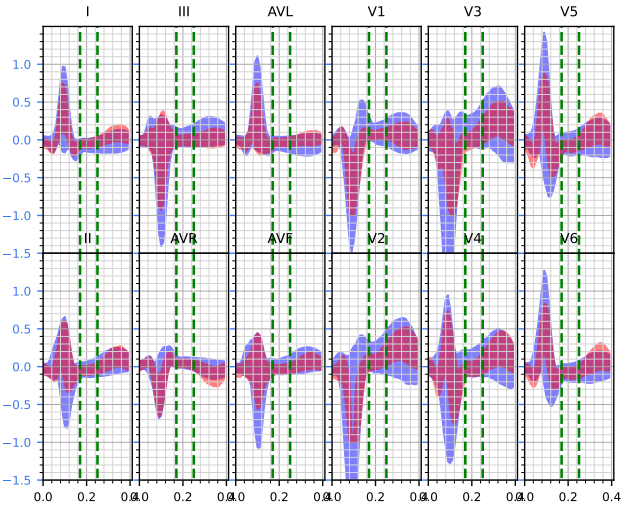

Figure 1: Confidence band plot at 95% for ASMI (blue) and ASMIgen (red) generated from NORM ECGs. Dashed green lines indicate the ST-segment.

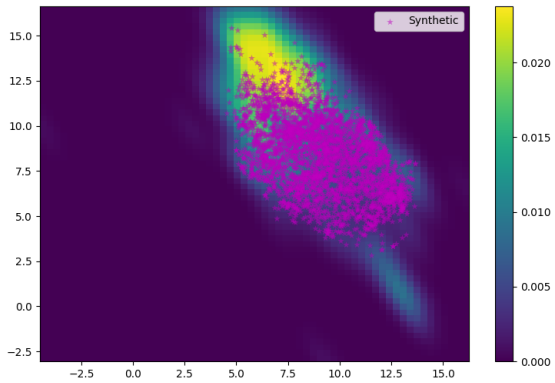

Figure 2: Heatmap of the real ASMI data obtained via UMAP and scatter plot of ASMIgen from NORM ECGs.

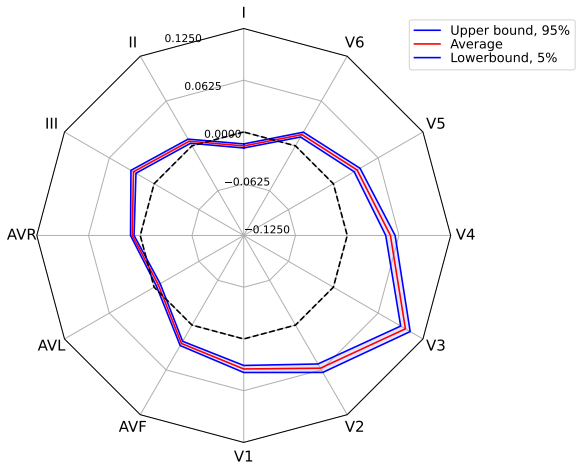

Figure 3: Average difference of ST-segment's amplitudes, and the corresponding 95% confidence interval, between ASMI and ASMIgen from NORM ECGs.

5. Discussion and Conclusion

The accuracy scores obtained through GAN-test and GAN-train classifiers gave us an insight on how close the generated data were to the real ones. In fact, the GAN-test classifier (trained on real data) displayed on synthetic data a slightly lower score than baseline (81% vs 89%). This suggests that the synthetic data are close enough in distribution to the real ECG. On the other hand, when comparing the GAN-train classifier (trained on synthetic data) score with its baseline value, the difference is much larger (62% vs 98%) hinting that the synthetic data did not display as much variability as that of the real ECG (the distribution of the synthetic data is narrower and mostly included in the wider distribution of the real ECG). These

considerations are confirmed by the heatmap-scatter plots that we computed. In particular, in the case of the ASMI class (fig. 2), we can notice that the generated data distribution generally overlaps the real data distribution, but the synthetic samples are mostly concentrated in a smaller region. So, although the distributions of real and synthetic data are overlapping in some regions, this partial overlap is not large enough to provide variability for plausible ECG samples. Similar considerations were drawn from both IMI-IMIgen and NORM-NORMgen heatmap-scatter plots (data not displayed).

In fig. 1, we provided an example of the 95% confidence bands for the real (blue band) and generated (red band) signals for the ASMI class. The bands of the generated signals are narrower, supporting the previous conclusion that the distribution of the synthetic data generally displays less variability. While genASMI generated signals might look plausible enough, they have in general smaller amplitudes in all the leads shown. IMI-IMIgen and NORM-NORMgen led to plots very similar to this one (not reported for brevity).

Finally, we analyzed the similarity of the ST-segments between synthetic and real ECG. Referring to the radar plot relative to the ASMI-ASMIgen comparison (fig. 3), the line corresponding to the value 0 (the black dashed line in the plot) was outside the 95% confidence intervals for most leads. Therefore, the distribution of the amplitudes (elevations) of the ST-segments generated by the model were, actually, statistically different from those of the real signals (in accordance with the previous findings). Similar conclusions apply for both the IMI-IMIgen, NORM-NORMgen comparisons.

In conclusion, in this work, we tested four techniques to evaluate the goodness of synthetic electrocardiograms generated through CycleGANs to simulate healthy and myocardial infarction conditions. The experiments we made showed how the generated data were not actually valuable enough to be employed for data augmentation purposes. Both global ECG distribution metrics and analysis on the ST-segment amplitudes support this claim. A main limitation of this study was the use of a generative neural network architecture that was originally defined for images. Future works will focus on the optimization of the generative networks.

Acknowledgments

This work was partially supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

[1] 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.

[2] Adib E, Fernandez AS, Afghah F, Prevost JJ. Synthetic ECG signal generation using probabilistic diffusion models. IEEE Access 2023.

[3] Wulan N, Wang W, Sun P, Wang K, Xia Y, Zhang H. Generating electrocardiogram signals by deep learning. Neurocomputing 2020, 404:122–136.

[4] Adib E, Afghah F, Prevost JJ. Synthetic ECG signal generation using generative neural networks, 2021. ArXiv preprint arXiv:2112.03268.

[5] Piacentino E, Guarner A, Angulo C. Generating synthetic ECGs using GANs for anonymizing healthcare data. Electronics 2021, 10(4):389.

[6] Zhu J, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In 2017 IEEE International Conference on Computer Vision (ICCV). 2017 2242–2251.

[7] Wagner P, Strodthoff N, Bousseljot R, Kreiseler D, Lunze F, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). *PhysioNet*, 2022. `https://doi.org/10.13026/kfzx-aw45`.

[8] 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.

[9] Xie C, McCullum L, Johnson A, Pollard T, Gow B, Moody B. Waveform Database Software Package (WFDB) for python, 2022. PhysioNet.

[10] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Networks. Communications of the ACM 2020, 63(11):139–144.

[11] McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018. ArXiv preprint arXiv:1802.03426.

[12] Shmelkov K, Schmid C, Alahari K. How good is my GAN? In Proceedings of the European Conference on Computer Vision (ECCV). 2018 213–229.

[13] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Meira Jr W, Others. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications 2020, 11(1):1–9.

Address for correspondence:

Sara Battiston
Department of Computer Science
University of Milan
Via Celoria 18, 20133, Milan, Italy
sara.battiston@unimi.it