Accurate ECG Data Generation with a Simple Generative Adversarial Network

Haiyan Wang\textsuperscript{1,2,3}, Zhaoyang Ge\textsuperscript{1,4} and Zongmin Wang \textsuperscript{*}

\textsuperscript{1} Collaborative Innovation Centre for Internet Healthcare, Zhengzhou University, 75 Daxue Road, Zhengzhou 450052, China
\textsuperscript{2} State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450003, China
\textsuperscript{3} Simulation Experiment Centre, Zhengzhou University of Aeronautics, Zhengzhou 450046, China
\textsuperscript{4} School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China
E-mail: zmwang@ha.edu.cn

\textbf{Abstract.} Since ECG data is highly sensitive medical data, the acquisition of ECG data is highly restricted. However, with the increasing demand for ECG big data research, the improvement of computer computing capabilities, and the development of deep learning, the direction of ECG intelligent analysis is facing a serious lack of standard clinical data. In order to generate more precise ECG data, this paper proposes a GAN architecture for generating ECG heartbeat data. The network structure is simple and does not require any domain knowledge. In this paper, the MIT-BIH Arrhythmia database is selected, from which all left bundle branch block heartbeats are selected to form a training dataset. The training process shows that the proposed GAN structure is effective and accurate, and the generated results show that not only the generated simulated ECG heartbeat data is diverse, but also highly similar to the real data.

\textbf{Keyword.} ECG data generation; heartbeat; left bundle branch block; deep learning model; GAN.

1. Introduction

In the information age, the security and privacy protection of medical data affects the life safety of patients, so it must be encrypted storage, encryption management, and desensitized during use. Compared with the general medical data, ECG data also has the function of personal identification \cite{1}, so it is highly sensitive data, and its acquisition and use are highly restricted. With the surge in research needs, especially the increase in the number of scenarios that require the use of ECG big data for analysis and early warning, and the continuous breakthroughs in artificial intelligence, especially deep learning technology, all of these have led to the extreme lack of ECG research data status. Therefore, there are scholars devoted to the research on the generation of simulated ECG data. The generated simulation data not only does not involve privacy protection issues, but also can generate a large number of ECG data that meet the clinical data standards and can be used for scientific research.

Generally speaking, ECG generation methods can be roughly divided into traditional methods and deep learning methods. Initially, the traditional method used was to build mathematical models using features. The most classic model is proposed by McSharry et al. \cite{2}. It is a dynamic model that
generates a single heartbeat signal based on three coupled ordinary differential equations, and then repeatedly connects each heartbeat to form an ECG signal. This model can generate very similar ECG signals of different heart rates. Li et al. [3] proposed a model for describing ECG signals with a data flow graph (DFG). The model is based on a piecewise curve modeling and generates ECG by periodically switching characteristic waves such as P wave, QRS wave, and T wave. Therefore, this model requires many parameters. Jafarnia-Dabanloo et al. [4] improved the Zeeman [5] model and obtained a comprehensive model to generate the time series for HRV (heart rate variability). This model takes into account respiratory sinus arrhythmia, Mayer waves and the important low-frequency component in the power spectrum of HRV, and generates a single cycle ECG signal by using a simple neural network. Sayadi et al. [6] modeled the temporal dynamics of ECG signals, and proposed a Gaussian-based ECG dynamic model as a generation model, which generated a synthetic ECG signal and separated characteristic waveforms. The characteristics of the above traditional methods are that you need to master ECG knowledge and accurately extract various types of ECG features as the required parameters of the model. Therefore, more parameter settings are required in the modeling process, which makes the model relatively complicated.

With the successful application of Generative Adversarial Networks (GAN) in the field of image generation, GAN is also gradually applied to the generation of ECG data [7, 8]. GAN is an end-to-end technology, and the construction of data generation models does not require domain expert knowledge. The GAN architecture proposed by Zhu et al. [7] and Delaney et al. [8] are similar. Both generators are composed of LSTM or BiLSTM, and the discriminators are composed of CNN. The training data used are all from the MIT-BIH Arrhythmia Database, and finally generate a segment of ECG data. It proves the feasibility of GAN as a tool for generating ECG data. The GAN architecture proposed above is complex, so the requirements for the hardware environment required for training are high, and the operation efficiency is low in the CPU environment.

In this paper, in order to more accurately generate simulated ECG data, the cut heartbeat data set is used as training data, and a simple GAN structure is designed. The discriminator and generator are composed of the most basic fully connected neural network. The model structure is simple, the efficiency of data training and generation is also high. The final result shows that the generated simulation data has a high morphological similarity.

The following chapters are arranged as follows. Chapter 2 introduces the data and preprocessing process required for the experiment, Chapter 3 introduces the GAN architecture used to generate the ECG, Chapter 4 describes the experimental results in detail, and Chapter 5 is a summary and outlook.

2. Data Preparation
This paper selects the most widely used MIT-BIH Arrhythmia Database [9, 10], which contains 48 ECG records with a duration of half an hour, collected from 47 subjects, each record sampling rate is 360HZ, and the database contains common cardiovascular diseases, such as left/right bundle branch block, atrial premature, supraventricular premature, ventricular flutter and ventricular fibrillation, etc. Our goal is to generate simulated ECG heartbeat data. In order to obtain a complete heartbeat, the specific steps of heartbeat segmentation are as follows: First, we need to locate and mark the R wave peak position, then center on the R peak, take 110 sampling points to the left and 175 to the right, so a heartbeat contains 286 samples point. Taking the left bundle branch block heartbeat as an example, we selected all heartbeats marked as "L". Finally, 8072 left bundle branch block heartbeats were obtained as the data set required for the experiment. For the convenience of training, the first 8064 heartbeat data training models were finally used.

3. Modeling
Generative adversarial network (GAN) [11] is a deep learning model, composed of two networks, one is a generative network, or called a generator, and the other is a discriminant network, which can also be called a discriminator. The purpose of generator $G$ is to receive random noise $z$ and generate as real data $G(z)$ as possible to deceive the discriminator $D$, and the purpose of the discriminator $D$ is to try to
distinguish the data generated by $G$ from the real data, $G$ and $D$ constitute a dynamic “game process”. The final ideal state of the game is that $G$ can generate data that is indistinguishable from true and false, and $D$ is difficult to determine whether the data generated by $G$ is real data, that is, the probability value of $D$ to determine whether $G(z)$ is true is 0.5, that is $D(G(z)) = 0.5$, then you get a generator $G$ that can be used to generate simulation data.

The optimized objective function $V(G, D)$ of the GAN network is shown in equation (1):

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} \left[ \log D(x) \right] + E_{z \sim p_z(z)} \left[ \log (1 - D(G(z))) \right]$$  \hspace{1cm} (1)

where $z$ represents the input random noise, $G(z)$ represents the sample generated by $G$, $x$ represents the real sample, $D(x)$ represents the result of the discriminator $D$ determining $x$, and $D(G(z))$ represents the result of the discriminator $D$ determining ($G(z)$). The optimization function of discriminator $D$ and generator $G$ can be obtained from equation (1), that is, the function of optimizing $D$ is shown in equation (2):

$$\max_D V(D, G) = E_{x \sim p_{data}(x)} \left[ \log D(x) \right] + E_{z \sim p_z(z)} \left[ \log (1 - D(G(z))) \right]$$  \hspace{1cm} (2)

The function to optimize $G$ is shown in equation (3):

$$\min_G V(D, G) = E_{z \sim p_z(z)} \left[ \log (1 - D(G(z))) \right]$$  \hspace{1cm} (3)

The GAN architecture designed to generate ECG data in this paper is shown in figure 1, where both $G$ and $D$ use a simple fully connected neural network structure, and the activation function selects LeakyReLU.

**Figure 1.** GAN architecture for generating ECG data.

4. Results

We analyze the experimental results from two aspects. One is process analysis, which is to analyze the validity of the model, and the other is result analysis, which is to analyze the quality of the results generated by the model. First, you can analyze whether the model is good or bad according to the Loss graph generated during the training process. For the generator, if the loss drops quickly, it is likely that the discriminator is too weak, causing the generator to "fool" the discriminator easily; and for the discriminator, if the loss drops quickly, it means that the discriminator is strong, which means that the image generated by the generator is not realistic enough, so that the discriminator can easily distinguish the authenticity. So whether it is a discriminator or a generator, the value of loss cannot represent the quality of the generator. For a good GAN model, its loss is always fluctuating. As shown in figure 2 is the loss graph generated during the GAN training process. It is easy to see that the loss...
curves of the generator and the discriminator are in a fluctuating state, and both have converged to 1. Prove that the training process is successful, indicating that the model is effective and correct.

![Figure 2. GAN errors plot in training.](image)

Although the convergence of the error graph can prove that the model training process is successful, it does not mean that the model can certainly generate high-quality ECG data. Therefore, the quality of the generated data still depends on the ECG generated last. As shown in figure 3, it is a comparison chart of the generated ECG heartbeat data and real ECG heartbeat data. The dotted line is the original heartbeat waveform, and the solid line is the heartbeat waveform generated in this paper. Taking three different forms of left bundle branch block heartbeats as an example, each heartbeat lists two representative generated heartbeats. It can be seen from the comparison that the generated data is highly consistent in morphology with the original data, and there are subtle differences between the generated data. It shows that the generated data is not only of high quality but also meets the requirements of the diversity of generated data.

![Figure 3. Comparison of generated and real ECG heartbeat data.](image)

5. **Summary and Outlook**

In order to solve the problem of insufficient ECG standard data, from the perspective of generating more accurate data and improving model training efficiency, a simple GAN architecture is proposed for generating ECG heartbeat data. Taking the left bundle branch block heartbeat as an example, the
loss graph produced by the training process showed that the proposed model is effective and accurate. Moreover, the generated data is not only highly similar in morphology to the original data, but also meets the diverse requirements of the generated data.

The current problem is that there are no authoritative indicators to measure the quality of the generated ECG data, so the next step is to research and propose more comprehensive indicators that can be used to verify the quality of the generated data.

References
[1] Biel L, Pettersson O, Philipson L and Wide P 2001 ECG analysis: A new approach in human identification IEEE Trans. Instru. Meas. 50 808-812.
[2] Mcsharry P E, Clifford G D, Tarassenko L and Smith L A 2003 A dynamical model for generating synthetic electrocardiogram signals IEEE Trans. Biomed. Eng. 50 pp 289-294.
[3] Li Z and Ma M 2005 ECG Modeling with DFG International Conference of the IEEE Engineering in Medicine and Biology Society pp 2691-94.
[4] Jafarnia-Dabanloo N, Mclernon D C, Zhang H, Ayatollahi A and Johari-Majd V 2007 A modified Zeeman model for producing HRV signals and its application to ECG signal generation J. Theor. Biol. 244 180-189.
[5] Zeeman E C 1973 Differential equations for the heartbeat and nerve impulse Dynamical Systems 4 683-741.
[6] Sayadi O, Shamsollahi M B and Clifford G D 2010 Synthetic ECG generation and Bayesian filtering using a Gaussian wave-based dynamical model Physiol. Meas. 31 1309-29.
[7] Zhu F, Ye F, Fu Y, Liu Q and Shen B 2019 Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network Scientific Reports 9 1-11.
[8] Delaney A M, Brophy E and Ward T E 2019 Synthesis of realistic ECG using generative adversarial networks arXiv: Signal Processing.
[9] Goldberger A L, Amaral L A, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K and Stanley H E 2000 PhysioBank, PhysioToolkit, and PhysioNet components of a new research resource for complex physiologic signals Circulation 101 215-220.
[10] Moody G B and Mark R G 2001 The impact of the MIT-BIH arrhythmia database IEEE Eng. Med. Biol. Mag. 20 45-50.
[11] Goodfellow I J, Pougetabadjie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial networks Advances in Neural Information Processing Systems 3 2672-80.