ECG-DPM: Electrocardiogram Generation via a Spectrogram-based Diffusion Probabilistic Model

Lujundong Li[1,2‡], Tong Xia[3‡], Haojie Zhang[1,2], *Student Member, IEEE*,
Dongchen He[4], Kun Qian[1,2*], *Senior Member, IEEE*, Bin Hu[1,2*], *Fellow, IEEE*,
Yoshiharu Yamamoto[5], *Member, IEEE*, Björn W. Schuller[6,7], *Fellow, IEEE*, and Cecilia Mascolo[3]
1. Key Laboratory of Brain Health Intelligent Evaluation and Intervention, Ministry of Education,
Beijing Institute of Technology, Beijing, China
2. School of Medical Technology, Beijing Institute of Technology, Beijing, China
3. Department of Computer Science and Technology, University of Cambridge, UK
4. School of Life Sciences, Peking University, Beijing, China
5. Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo, Japan
6. GLAM – the Group on Language, Audio, & Music, Imperial College London, UK
7. CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany
{*qian, bh*}*@bit.edu.cn*

***Abstract*—An electrocardiogram (ECG) records the electrical signals from the heart to assess various cardiovascular conditions. Deep learning methods have been proposed to model ECGs, but the insufficient availability of ECG data and annotations often hinders their performance. To address this challenge, this paper explores the latest data synthesis technique, i.e., diffusion probabilistic models (DPMs), to enable the generation of an unlimited number of ECGs representing various cardiovascular conditions. In contrast to previous approaches that treat ECGs as time series data or convert them into power spectrograms, we introduce a novel multi-channel spectrogram-based diffusion framework. In our method, the diffusion model enhances generation diversity, while the multi-channel spectrogram preserves both magnitude and phase information, ensuring high fidelity in the reconstructed ECGs. Extensive experiments conducted on real-world ECG data demonstrate the superiority of our approach. Notably, our ECG-DPM outperforms the best baseline by a margin ranging from 12.5% to 62.5% when generating ECGs for 30 seconds.**

***Index Terms*—Electrocardiogram, Data Generation, Mel Spectrogram, Diffusion Model**

I. Introduction

Cardiovascular diseases (CVDs) are among the leading causes of death globally [1], [2]. To prevent severe CVDs from developing, clinicians usually measure electrocardiograms (ECGs) to diagnose conditions and deliver timely interventions. Machine learning models, particularly deep learning models, are now being widely studied to automatically classify ECGs and improve clinical efficiency [3]–[6]. Unfortunately, the performance of these methods highly depends on the quality and quantity of annotated ECG data available for model training. Yet, it is usually difficult to gather sufficient data for diagnostic model developing, due to health data privacy constraints and the high cost of annotation by cardiologists [7]. To this end, generating realistic ECGs becomes a crucial task.

Recognising the significance of ECG generation, extensive studies have been conducted, facilitated by remarkable advancements in deep generative models. Among those, the applications of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have emerged as outstanding examples [8]–[10]. Despite their impressive performance in generating ECGs, these methods require large data to fit the additional model parameters and often result in unstable generations when compared to traditional statistical models [11].

Built on deep generative models, existing works generate ECGs in two ways: i) by directly generating ECG signals, treating ECG as one-dimensional time series data [8], [10], and ii) by generating power spectrograms through Short-Time Fourier Transform [12] applied to the signals, followed by the subsequent recovery of ECG signals from the spectrograms [9], [13]. However, the former approach tends to introduce the issue of ECG amplitude baseline drift, whereas in the latter approach, information is lost during the signal recovery process from the magnitude of the spectrograms [12], [14].

To address the aforementioned limitations in existing ECG generation methods, this paper introduces a novel spectrogram generation framework called *ECG-DPM*. This framework is inspired by the recently emerged approach in generative modelling literature known as diffusion probabilistic models (DPMs) [15], [16]. Its core idea for data generation involves initially learning the underlying data distribution by iteratively adding noise to an initial sample and subsequently generating samples by removing noise from any given noisy sample. Compared to GANs and VAEs, DPMs have demonstrated

This work was partially supported by the National Natural Science Foundation of China (Nos. 62272044 and 62227807), the National Key R&D Program of China (No. 2023YFC2506804), the Beijing Natural Science Foundation (No. L243034), the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects (Nos. 2021ZD0201900 and 2021ZD0200601), the European Research Council Project 833296 (EAR), and the Teli Young Fellow Program from the Beijing Institute of Technology, China. (*Lujundong Li and Tong Xia contributed equally to this work. Corresponding authors: Kun Qian and Bin Hu.*)

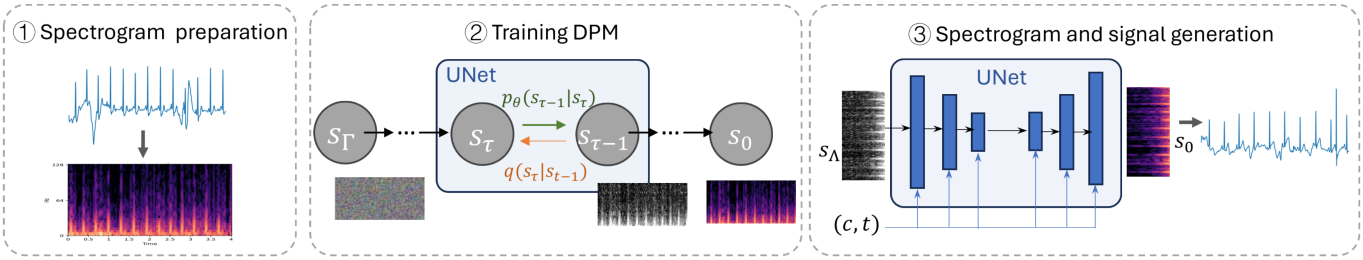

Fig. 1. ECG-DPM framework overview. The three-channel spectrograms are fed into a DPM underpinned by a UNet architecture. New spectrograms can be generated from the model and thus, ECGs signals can be reconstructed from the spectrograms.

superior efficiency and robustness in various applications [17]. Building upon DPMs, for the complete recovery of the ECG signals from spectrograms, we utilise the real part, the imaginary part, and the magnitude of the spectrogram as a three-channel input to fit the model parameters.

The proposed *ECG-DPM* was trained on a publicly available single-lead ECG database [18], enabling us to generate an unlimited number of ECGs. Visualisation demonstrates that the generated ECG signals, representing both normal and abnormal classes, exhibit high quality and diversity. Additionally, we conducted a quantitative comparison of the fidelity of our generated ECGs with two state-of-the-art (SOTA) baselines. Our extensive results reveal that our generated ECG signals can be accurately classified by previously developed CVD diagnostic models (outperforming the SOTA by $12.5 \sim 19.7\%$), exhibit a strong similarity to the original ECGs (outperforming SOTA by 37.4%), and maintain stability for long signal duration (outperforming the SOTA by 62.5%).

II. Methods

This section introduces our proposed ECG-DPM framework. As illustrated in Fig. 1, ECG-DPM consists of there primary modules, which are elaborated on in the following.

A. Spectrogram preparation

As a variation of the Fourier Transform, STFT provides a time-dependent representation of the frequency components of a signal. Given an ECG signal $x(t)$, STFT is defined by the energy coefficient at any time t and frequency f. To easy the computing, it is common to use the discrete STFT [19]: $x(t)$ obtained by sampling frequency f_s will be segmented into overlapped short windows (i.e., each segment has a duration T containing N data points with $N = T \cdot f_s$). The frequencies are considered by frequency bins, which are evenly spaced between 0 Hz and the Nyquist frequency (i.e., half of f_s) [20]. Mathematically, the discrete STFT can be expressed as,

$$X[n,k] = \sum_{m=0}^{N-1} x[n+m] \cdot w[m] \cdot e^{-j2\pi km/N}, \qquad (1)$$

where $w[m]$ is the value of the window function at segment m, and $e^{-j2\pi km/N}$ represents the complex sinusoidal basis function at frequency bin k for segment m.

Under certain constrains (e.g., constant overlap-add compliant) [12], the signal can be reconstructed from the discrete STFT by[1],

$$x(t) = \sum_n \sum_k X[n,k] \cdot w[t-nT] \cdot e^{j2\pi kt/T}. \qquad (2)$$

For spectrogram based deep learning models, the *power spectrogram* (i.e., the magnitude of the complex STFT coefficient) is widely used [21], [22], as it provides information about the distribution of power (or energy) across different frequency components and time intervals in a signal. However, for signal generation purpose, it should be noted that the signal $x(t)$ cannot be recovered from merely the magnitude of $X[n,k]$. This suggests that even if a generative approach can generate perfect power spectrograms, we can hardly reconstruct the signals from the spectrograms. To this end, we propose to use the three-channel spectrogram $S[n,k]$ which preserves all the information that is needed for signal reconstruction. Specifically, $S[n,k]$ consists of the real channel, imaginary channel, and magnitude channel, as below,

$$\begin{aligned} S_{\mathbb{R}}[n,k] &= \sum_{m=0}^{N-1} x[n+m] \cdot w[m] \cdot \cos(-2\pi km/N), \\ S_{\mathbb{I}}[n,k] &= \sum_{m=0}^{N-1} x[n+m] \cdot w[m] \cdot \sin(-2\pi km/N), \\ S_{\mathbb{M}}[n,k] &= \sum_{m=0}^{N-1} x[n+m] \cdot w[m]. \end{aligned} \qquad (3)$$

Like the RGB channels of an image, those three channels for a spectrogram are correlated with each other, as they are bounded by the phase $\phi = -2\pi km/N$ and the magnitudes. Now, S tends out as a real-valued spectrogram, which can be fit into deep learning models. $S \in \mathbb{R}^{3\times L\times K}$, where L is the total number of temporal segments, and K is the total number of frequency bins. Fig. 2 displays an example of a three channel spectrogram.

B. Diffusion probabilistic model training

Diffusion probabilistic models (DPMs) [15], [17] generate samples by learning the target data's underlining distribution.

[1] $e^{j2\pi kt/T}$ a complex-valued exponential term, but when we sum up all the contributions over k for each time n, the result is a real-valued $x(t)$.

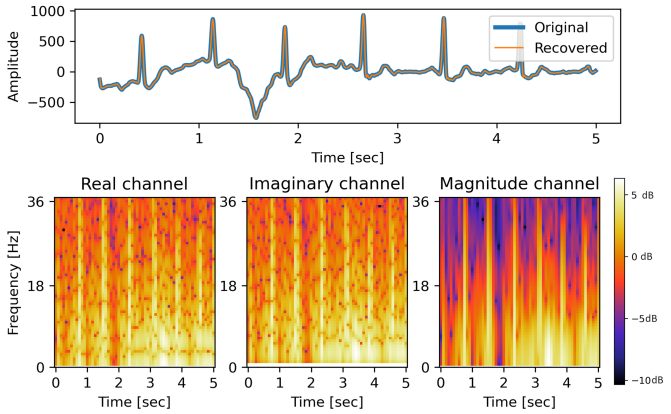

Fig. 2. Example of a three-channel spectrogram. For visualisation purpose, we display the logarithm of the absolute value of the coefficient, but we fit the original coefficient into the model. These three channels are highly related but contain different information.

This is learnt by a gradual reverse process of adding noise, which recovers the less noisy value in each step.

In the forward process of DPMs, Gaussian noise is gradually added to the initial clean observation $s_0 \in \mathbb{R}^{3\times L\times K}$, until $s_\Gamma \in \mathbb{R}^{3\times L\times K}$ becomes a random noise after Γ steps. The process can be formulated by a Markov chain,

$$q(s_{1:\Gamma}|s_0) = \prod_{\tau=1}^{\Gamma} q(s_\tau|s_{\tau-1}), \tag{4}$$

where $q(s_\tau|s_{\tau-1})$ is a known Gaussian distribution parameterised by β_τ. On the contrary, the reverse denoising process is unknown, and DPMs aim to learn a deep learning model θ to approximate the reverse distribution $p_\theta(s_{\tau-1}|s_\tau)$ for any time step τ. The parameter θ can be optimised by minimising the negative log-likelihood via a variational bound,

$$\min_{\lim\theta} \mathbb{E}_q \leq \min_{\lim\theta} \mathbb{E}_q[-\log p(s_\Gamma) - \sum_{\tau=1}^{\Gamma} \frac{p_\theta(s_{\tau-1}|s_\tau)}{q(s_\tau|s_{\tau-1})}]. \tag{5}$$

For implementation, we choose the UNet architecture [23], which comprises an encoder and a decoder as the model θ to reduce the noise for a given diffusion step. The encoder progressively reduces the resolution of map features, while the decoder employs up-sampling gradually to restore feature maps to the original shape. The class label c for the ECG sample and the diffusion step τ are also concatenated with the feature maps; hereby, the output of the model is denoted by $\epsilon_\theta(s_\tau, c, \tau)$. By applying a re-parameterisation trick [24] for Equ. (5), the optimisation for θ can be efficiently achieved by Algorithm 1 (line 1-6).

C. Spectrogram and signal generation

Once the model θ is trained, we can generate ECGs from a noisy spectrogram s'_Λ. s'_Λ is obtained by adding Λ steps of noise into a randomly picked clear spectrogram s_0. Then, we employ the trained UNet to gradually remove the noise from s_Λ, ending up with a clear and new ECG spectrogram s'_0. The detailed implementation is illustrated in Algorithm 1 (line 8-13).

Algorithm 1: Training and Sampling

Input: Noise level β_τ, so $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha_\tau} = 1 - \prod_{\tau=1}^{\Gamma} \beta_\tau$.

1 **Training (repeat until converged):**
2 Sample s_0 from the dataset
3 $\tau \sim$ Uniform($\{1, 2, .., \Gamma\}$)
4 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{I} \in \mathbb{R}^{3\times L\times K}$
5 Take gradient descent step on:
6 $\nabla_\theta \mathcal{L}_{MSE}(\epsilon, \epsilon_\theta(\sqrt{\bar{\alpha_\tau}} s_0 + \sqrt{1-\bar{\alpha_\tau}}\epsilon, c, \tau))$
7
Sampling:
8 $s'_\Lambda \leftarrow q(s_{1:\Lambda}|s_0)$
9 $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{I} \in \mathbb{R}^{3\times L\times K}$
10 **for** $\tau = \Lambda, .., 1$ **do**
11 $s'_{\tau-1} = \frac{1}{\sqrt{\alpha_\tau}}(s'_\tau - \frac{1-\alpha_\tau}{\sqrt{1-\bar{\alpha\tau}}})\epsilon_\theta(s'_\tau, c, \tau) + \sqrt{\beta_\tau} z$
12 **end**
13 Return $S \leftarrow s'_0$

The generated instance S is a three-channel spectrogram as defined by Equ. (3). We recover the ECG signal from the real and imaginary channels by inverse the STFT according to Equ. (2). Formally, the reconstruction is formulated by,

$$x(t) = \sum_{n=1}^{L}\sum_{k=1}^{K}(S_{\mathbb{R}}[n,k] + jS_{\mathbb{I}}[n,k]) \cdot w[t-nT]. \tag{6}$$

III. Experiments

A. Experimental set-up

Dataset. We trained the model using the dataset from the PhysioNet/CinC Challenge 2017 [18]. The dataset consists of single-lead ECG recordings collected via clinical devices. The training set comprises 8 528 recordings, with duration ranging from 9 to slightly over 60 seconds. All records were sampled at a rate of 300 Hz and were annotated into normal, Atrial Fibrillation (AF), and other categories. For our training, we used data from both normal and AF categories.

Since the original signals vary in length, we uniformly trimmed them to the first 5 seconds and the first 30 seconds (signals shorter than 30 seconds were extended by repetitive padding). We then converted the signals into spectrograms using STFT with a Hanning window function (w), a segmentation length of $N = 50$, an overlapping length of 25, and FFT points set to $K = 512$ [25]. Observing that the energy is mainly distributed in the low-frequency band of the spectrograms, to expedite model training, we retained only the first 64 frequency bins of the spectrograms. This is equivalent to applying a low-pass filtering technique to the original signals with cutoff frequencies set at 37.4 Hz [26]. Consequently, the resulting spectrogram has dimensions of $\mathbb{R}^{3\times 61\times 64}$.

Training parameters. The UNet consists of four down-sampling layers, a middle layer, and four up-sampling layers.

Three ResNet blocks are introduced before each sampling layer. Attention blocks are added at the last down-sampling layer and the first up-sampling layer. The middle layer consists of two ResNet blocks and an attention block. Channel multiplications are set to $\{1, 2, 4, 8\}$, respectively, and sampling layers are implemented with convolution blocks. We implemented our UNet with reference to [15], [16], [27]. For the diffusion process, we used $\Gamma = 1000$ and β_τ uniformly distributed in $[0.0001, 0.2]$ to convert the spectrogram into a random noise[2]. To optimise the parameters, the AdamW optimiser with a weight decay of $1e-6$ and a starting learning rate of $5e-5$ was utilised. The batch size was set to 64. All experiments were conducted by Python on three RTX 4090 GPUs.

Baselines. We compared our model with the diffusion model that takes the time series data directly as input, which is called ***1D-Diffusion*** method. The structure of the UNet for this baseline is similar to that of ECG-DPM, with the same channel multiplication of $\{1, 2, 4, 8\}$. This baseline utilises one-dimensional (1D) conventional layers, specifically four 1D-ResNet blocks in each layer to model time series. We also employed another ECG generation baseline based on GAN, namely ***ECG-GAN*** [28]. This method consists of a generator which converts given random noise into ECG signals via a Long Short-Term Memory (LSTM) [29] network and a discriminator which employs a similar LSTM to predict whether the input is an ECG or noise. We used the official implementation from [28]. In this model, we trained two models for the normal and AF class, separately, since the LSTM layer cannot explicitly decode the class label.

B. Experimental results

Visualisation of the synthetic ECGs. We first demonstrate that ECG signals can be recovered from our defined and preprocessed spectrograms. Fig. 2 illustrates the spectrogram of the provided signal. It can be observed that by retaining only the low-frequency band up to 37.4 Hz, the signal can be successfully reconstructed from the real and imaginary parts of the spectrogram (as described in Equ. (6)). Given that the typical heart rate falls within the range of 60 to 100 beats per minute, a cutoff frequency of 37.4 Hz strikes a good balance between learning complexity and effectiveness.

After confirming the suitability of the spectrograms for ECG signals, we proceeded to train our ECG-DPM model and generate spectrograms using Algorithm 1. In Fig. 3, we present examples of both original and generated ECG signals. It is evident that our ECG-DPM generates ECGs that differ from the original signals. The generated normal ECGs exhibit regular heartbeats but show various artifacts, while the generated AF ECGs tend to display irregular rhythms. This observation also underscores the capability of ECG-DPM to capture the class-conditional data distribution, enabling the generation of an unlimited number of new samples.

Quantitative comparison. In addition to visualisation, we also conducted experiments to quantitatively demonstrate the superiority of our model compared to baselines. The following measurements are reported to showcase the fidelity of the generated ECGs. For all generative methods, we generated 150 normal samples and 150 AF samples to derive metrics. As a reference, we also report the metrics on the original validation set containing 150 normal and 50 AF samples.

[2]We adapted the implementation from https://github.com/openai/improved-diffusion

TABLE I
QUANTITATIVE PERFORMANCE COMPARISON WITH BASELINES (WITH ECG DURATION FIXED AT 5 SECONDS). THE ARROW INDICATES THE OPTIMAL DIRECTION OF THE METRICS.

	Precision↑	Recall↑	FID↓	Slope↓
Raw validation set	0.768	0.884	0.000	9.86e-4
ECG-GAN	0.461	0.786	34.328	1.54e-2
1D-Diffusion ($\Lambda = 200$)	0.585	0.640	36.211	1.69e-3
1D-Diffusion ($\Lambda = 100$)	0.621	0.857	21.305	1.77e-3
ECG-DPM ($\Lambda = 200$)	0.563	0.692	28.203	2.79e-4
ECG-DPM ($\Lambda = 100$)	0.712	0.840	11.135	9.54e-4

TABLE II
QUANTITATIVE PERFORMANCE COMPARISON WITH BASELINES (WITH ECG DURATION SET TO 30 SECONDS). Δ REPRESENTS THE RELATIVE IMPROVEMENT OF OUR *ECG-DPM* OVER *1D-Diffusion*.

	Precision↑	Recall↑	FID↓	Slope↓
Raw validation set	0.852	0.910	0.000	9.86e-4
ECG-GAN	0.501	0.415	45.623	1.01e-3
1D-Diffusion ($\Lambda = 100$)	0.527	0.583	38.844	2.35e-3
ECG-DPM ($\Lambda = 100$)	0.593	0.698	24.312	8.82e-4
Δ	12.5%	19.7%	37.4%	62.5%

topsep=1pt, itemsep=0pt, leftmargin=10pt

- **Being correctly classified**. We employed an ECG classifier trained by the same databases to classify the synthetic ECGs [30]. The performance is quantified by *Precision* and *Recall* by treating normal as class 0 and AF as class 1.
- **Be similarly distributed with the training ECGs.** To quantify the distance of the distribution between the original ECG set and the generated set ECG, we employed Fréchet Inception Distance (FID) [31]. To obtain FID, we derived embeddings for the power spectrogram of ECGs via a pre-trained InceptionV3 [32], [33], and then calculate FID as $FID = ||\mu_g - \mu_o||^2 + Tr(\Sigma_g + \Sigma_o - 2\sqrt{\Sigma_g \Sigma_o})$, where $\mu \in \mathcal{R}^{2048}$ and $\Sigma \in \mathcal{R}^{2048 \times 2048}$ denote the mean and covariance for embeddings. The subscript g and o denote the generated set and original set, respectively.
- **Being stable**. We notice that the baselines that generated ECGs as time series face the problem of ECG amplitude drift, while our spectrogram-based method does not have this issue. To verify this, we fit the baseline amplitude of the ECGs using the least squares, and report the averaged slope.

The results are summarised in TABLEs I and II. It is evident that regardless of generating ECGs at different lengths (either

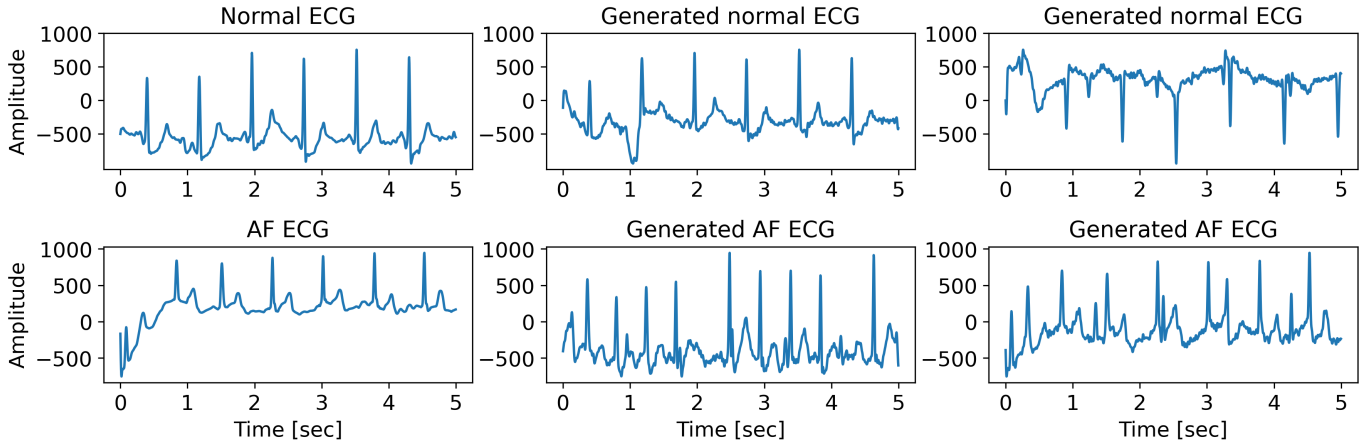

Fig. 3. Examples of the raw and generated ECGs for normal and AF classes, where we set $\Lambda = 100$. The generated samples are visibly different from the original signals that are used for model training.

5 seconds or 30 seconds) and under varying noise levels ($\Lambda =$ 100 or 200), our ECG-DPM consistently outperforms both 1D-Diffusion and ECG-GAN across all metrics. Notably, when generating longer ECGs, i.e., 30 seconds, with a noise level of $\Lambda = 100$, our ECG-DPM surpasses 1D-Diffusion by a margin ranging from 12.5% to 62.5%, as detailed in TABLE II. This highlights the high fidelity and stability of the ECGs generated by our model.

While there is a noticeable performance gap between the generated ECGs and the original validation set, our model excels in generating an unlimited number of ECGs as data augmentations while maintaining acceptable data quality. By comparing TABLE I and TABLE II, we also observe that generating longer ECGs (e.g., 30 seconds) presents a greater challenge compared to shorter ones (e.g., 5 seconds). This underscores the need for more extensive training data to enhance the performance of generative models.

IV. Conclusions

In this paper, we explored a spectrogram-based diffusion probabilistic model for generating ECGs. Our proposed method, ECG-DPM, has demonstrated superior performance compared to other state-of-the-art baselines. By utilizing a novel multi-channel spectrogram-based approach, our model ensures high fidelity in the generated ECGs and effectively captures both magnitude and phase information. Extensive experiments show that ECG-DPM outperforms existing methods by a significant margin, achieving improvements ranging from 12.5% to 62.5% in various metrics. These findings highlight the potential of ECG-DPM in generating high-quality, diverse ECG signals that can be used for augmenting training datasets and improving machine learning models for cardiovascular diagnosis.

In the future, we aim to extend our work to generate multi-lead ECGs, which would provide a more comprehensive representation of the heart's electrical activity. Additionally, we plan to explore the generation of other types of bio-signals, such as electroencephalograms (EEGs) and electromyograms (EMGs), leveraging the flexibility of our diffusion probabilistic model framework. These extensions will further validate the robustness and applicability of our approach across different types of sensory data.

The ability to generate realistic bio-signals has several important applications. Firstly, it can significantly enhance data privacy during data sharing by generating synthetic datasets that can be used for research and development without compromising patient confidentiality, thereby preventing attacks from individual identification methods [34], [35]. Secondly, the generated data can contribute to the training of large-scale models, providing abundant and diverse training examples that improve model generalization and performance [36]–[38]. Additionally, given the popularity and power of multi-modal large language models (MM-LLMs) [39], integrating our generative model into these frameworks as a decoder will enable more applications, such as combining textual clinical notes for bio-signal generation [40], [41].

References

[1] K. Mc Namara, H. Alzubaidi, and J. K. Jackson, "Cardiovascular disease as a leading cause of death: how are pharmacists getting involved?" *Integrated Pharmacy Research and Practice*, pp. 1–11, 2019.

[2] K. Qian, B. Hu, Y. Yamamoto, and B. W. Schuller, "The voice of the body: Why ai should listen to it and an archive," *Cyborg and Bionic Systems*, vol. 4, p. 0005, 2023.

[3] B. Pyakillya, N. Kazachenko, and N. Mikhailovsky, "Deep learning for ecg classification," in *Journal of Physics: Conference Series*, vol. 913. Madrid, Spain: IOP Publishing, 2017, p. 012004.

[4] X. Liu, H. Wang, Z. Li, and L. Qin, "Deep learning in ecg diagnosis: A review," *Knowledge-Based Systems*, vol. 227, p. 107187, 2021.

[5] L. El Bouny, M. Khalil, and A. Adib, "Ecg heartbeat classification based on multi-scale wavelet convolutional neural networks," in *Proc. ICASSP*. Barcelona, Spain: IEEE, 2020, pp. 3212–3216.

[6] K. Qian, Z. Zhang, Y. Yamamoto, and B. W. Schuller, "Artificial intelligence internet of things for the elderly: From assisted living to health-care monitoring," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 78–88, 2021.

[7] K. Qian, X. Li, H. Li, S. Li, W. Li, Z. Ning, S. Yu, L. Hou, G. Tang, J. Lu *et al.*, "Computer audition for healthcare: Opportunities and challenges," *Frontiers in Digital Health*, vol. 2, p. 5, 2020.

[8] H. Chung, J. Kim, J.-m. Kwon, K.-H. Jeon, M. S. Lee, and E. Choi, "Text-to-ecg: 12-lead electrocardiogram synthesis conditioned on clinical text reports," in *Proc. ICASSP*. Rhodes Island, Greece: IEEE, 2023, pp. 1–5.

[9] E. Adib, F. Afghah, and J. J. Prevost, "Synthetic ecg signal generation using generative neural networks," *arXiv preprint arXiv:2112.03268*, 2021.

[10] Y. Xia, W. Wang, and K. Wang, "Ecg signal generation based on conditional generative models," *Biomedical Signal Processing and Control*, vol. 82, p. 104587, 2023.

[11] Y. Gu, Q. Chen, K. Liu, L. Xie, and C. Kang, "Gan-based model for residential load generation considering typical consumption patterns," in *Proc. ISGT-Europe*. Bucharest, Romania: IEEE, 2019, pp. 1–5.

[12] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[13] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, vol. 33, Virtual Conference, 2020, pp. 17 022–17 033.

[14] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*. Virtual Conference: IEEE, 2020, pp. 6199–6203.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, vol. 33, Virtual Conference, 2020, pp. 6840–6851.

[16] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. ICML*. Virtual Conference: PMLR, 2021, pp. 8162–8171.

[17] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[18] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. Johnson, and R. G. Mark, "Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017," in *Proc. CinC*. Rennes, France: IEEE, 2017, pp. 1–4.

[19] K. Qian, "Automatic general audio signal classification," Ph.D. dissertation, Technische Universität München, 2018.

[20] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.

[21] Z. Ahmad, A. Tabassum, L. Guan, and N. Khan, "Ecg heart-beat classification using multimodal image fusion," in *Proc. ICASSP*. Toronto, Ontario, Canada: IEEE, 2021, pp. 1330–1334.

[22] F. N. Hatamian, N. Ravikumar, S. Vesal, F. P. Kemeth, M. Struck, and A. Maier, "The effect of data augmentation on classification of atrial fibrillation in short single-lead ecg signals using deep neural networks," in *Proc. ICASSP*. Virtual Conference: IEEE, 2020, pp. 1264–1268.

[23] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021.

[24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[25] M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time fourier transform domain," *IEEE Circuits and Systems Magazine*, vol. 16, no. 3, pp. 45–77, 2016.

[26] T.-K. Hon, S. R. Subramaniam, A. Georgakis, and S. Alty, "Stft-based denoising of elastograms," in *Proc. ICASSP*. Prague, Czech Republic: IEEE, 2011, pp. 677–680.

[27] H. Zhang, K. Qian, L. Shen, L. Li, K. Xu, and B. Hu, "From noise to sound: Audio synthesis via diffusion models," DCASE2023 Challenge, Tampere, Finland, Tech. Rep., 2023.

[28] A. M. Delaney, E. Brophy, and T. E. Ward, "Synthesis of realistic ecg using generative adversarial networks," *arXiv preprint arXiv:1909.09150*, 2019.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] Z. Xiong, M. K. Stiles, and J. Zhao, "Robust ecg signal classification for detection of atrial fibrillation using a novel neural network," in *Proc. CinC*. Rennes, France: IEEE, 2017, pp. 1–4.

[31] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[32] J. Ji, L. Zhu, H. Zhang, K. Qian, K. Xu, Z. Song, B. Hu, B. W. Schuller, and Y. Yamamoto, "Weight light, hear right: Heart sound classification with a low-complexity model," in *Proc. EUSIPCO*. Lyon, France: IEEE, 2024, pp. 1–5.

[33] X. Qiu, L. Zhu, Z. Song, Z. Chen, H. Zhang, K. Qian, Y. Zhang, B. Hu, Y. Yamamoto, and B. W. Schuller, "Study selectively: An adaptive knowledge distillation based on a voting network for heart sound classification," in *Proc. INTERSPEECH*, Kos Island, Greece, 2024, pp. 1–5.

[34] B.-H. Kim and J.-Y. Pyun, "Ecg identification for personal authentication using lstm-based deep recurrent neural networks," *Sensors*, vol. 20, no. 11, p. 3069, 2020.

[35] D. Jyotishi and S. Dandapat, "An lstm-based model for person identification using ecg signal," *IEEE Sensors Letters*, vol. 4, no. 8, pp. 1–4, 2020.

[36] Y. Zeng, Y. Feng, R. Ma, Z. Wang, R. Yan, C. Shi, and D. Zhao, "Scale up event extraction learning via automatic training data generation," in *Proc. AAAI*, vol. 32, no. 1, New Orleans, Louisiana, USA, 2018.

[37] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. A. Mann, "Improving robustness using generated data," in *Proc. NeurIPS*, vol. 34, Virtual Conference, 2021, pp. 4218–4233.

[38] V. Thambawita, P. Salehi, S. A. Sheshkal, S. A. Hicks, H. L. Hammer, S. Parasa, T. d. Lange, P. Halvorsen, and M. A. Riegler, "Singan-seg: Synthetic training data generation for medical image segmentation," *PloS one*, vol. 17, no. 5, p. e0267976, 2022.

[39] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," *arXiv preprint arXiv:2401.13601*, 2024.

[40] Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue, "Meta-transformer: A unified framework for multimodal learning," *arXiv preprint arXiv:2307.10802*, 2023.

[41] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," in *Proc. AAAI*, vol. 38, no. 3, Vancouver, Canada, 2024, pp. 2256–2264.