Conditional Diffusion with Label Smoothing for Data Synthesis from Examples with Noisy Labels

Gentry Atkinson, Xiaomin Li, Vangelis Metsis
Department of Computer Science, Texas State University, USA
{gma23, x_l30, vmetsis}@txstate.edu

***Abstract*—Data generation techniques are critical in various fields where obtaining real-world data is difficult or expensive. However, data generators like generative adversarial networks (GANs) and diffusion-based models depend on training models using pre-existing data and labels. When the labels are unreliable, the performance of data generators can suffer. This paper proposes a novel adaptation of Denoising Diffusion Probabilistic Models (DDPM) that employs label smoothing to enhance the reliability of the generated data in the presence of label noise. Label smoothing mitigates the impact of label noise by preventing the model from becoming overconfident in mislabeled instances of data. We demonstrate that DDPM with label smoothing outperforms both conditional and unconditional DDPM in terms of the closeness of the generated data to the original data's distribution, even when the training data contains instances with mislabeled labels.**

***Index Terms*—Data generation techniques, conditional DDPM, label noise, label smoothing.**

I. Introduction

Data generation techniques are essential for machine learning (ML) approaches, enabling the application of ML models to datasets that would otherwise have insufficient samples to train them. Methods for generating signal data include generative adversarial networks (GANs) [17], denoising diffusion probabilistic models (DDPMs) [8], and statistical techniques [10]. Although both GANs and DDPMs can replicate the features of existing data, diffusion-based models have been shown to better replicate all classes of a dataset while GANs can overfit to the majority class [5]. However, the performance of ML models can be severely degraded when some labels in a dataset are misassigned [1]. In this paper, we propose an application of label smoothing to DDPM, which improves its performance in the presence of label noise.

Real-world datasets typically have a mislabeling rate of 3% to 5% [20]. Mislabeled instances of data are assigned a label ($\tilde{Y}$) that does not reflect the true class of the instance (Y), with some likelihood $P(\tilde{Y} \neq Y)$, which may vary by class or within the feature space X. Mislabeled training data have been shown to have a greater impact on the training of ML models than their correctly labeled counterparts [23]. This means that a low occurrence of label noise in training data can have a disproportionate impact on the performance of models trained using the noisy dataset. Both GANs and DDPMs rely on the performance of trained deep learners and are, therefore, susceptible to label noise in their training data.

Signal data, and time-series data in general, can be more challenging to interpret than other domains of data like images and text, making it easier to mislabel during data collection and more difficult to correct the labels later on [1]. Time-series data require special tools for label cleaning [2], making signal data a particularly interesting domain for considering ML approaches related to label noise.

The DDPM trains a denoising model with the label of the instance being de-noised embedded using some positional encoder. Following training, new instances of data are generated by feeding samples of Gaussian noise into the denoising model with a targeted label embedded in the noise. Label smoothing, which augments the values of one-hot encoded labels during training, has been shown to mitigate the harmful effects of label noise by preventing the model from developing overconfidence in the noisy instances of data [14]. It should be pointed out that data noise and label noise are two different types of errors that can occur in a dataset. Data noise refers to errors in the input data itself that can be caused, for example, by faulty measurements. Label noise, on the other hand, refers to errors in the class labels assigned to the input data.

This work has adapted label smoothing to conditional DDPM to make a data generation technique that is robust to the presence of label noise in the training data[1]. We will show that this technique outperforms conditional and unconditional DDPM in the closeness of fit to the distribution of the original data when trained with noisy labels. The contributions of this paper are as follows:

1) A novel integration of label smoothing with DDPM.
2) An adaptation of Fréchet Inception Distance (FID) [11] for use with signal data.
3) A first analysis of the impact of label noise on data generation using DDPM.

II. Related Work

In recent years, there has been a growing interest in using Generative Adversarial Networks (GANs) for time-series and sequential data generation. [3] provided a comprehensive overview of GAN implementations on time-series data and highlighted the advantages of using GAN as a time-series data augmentation tool. GANs can solve data shortage issues by augmenting smaller datasets and generating new, previously unseen data. They can also recover missing or corrupted data, reduce data noise, and protect data privacy by generating

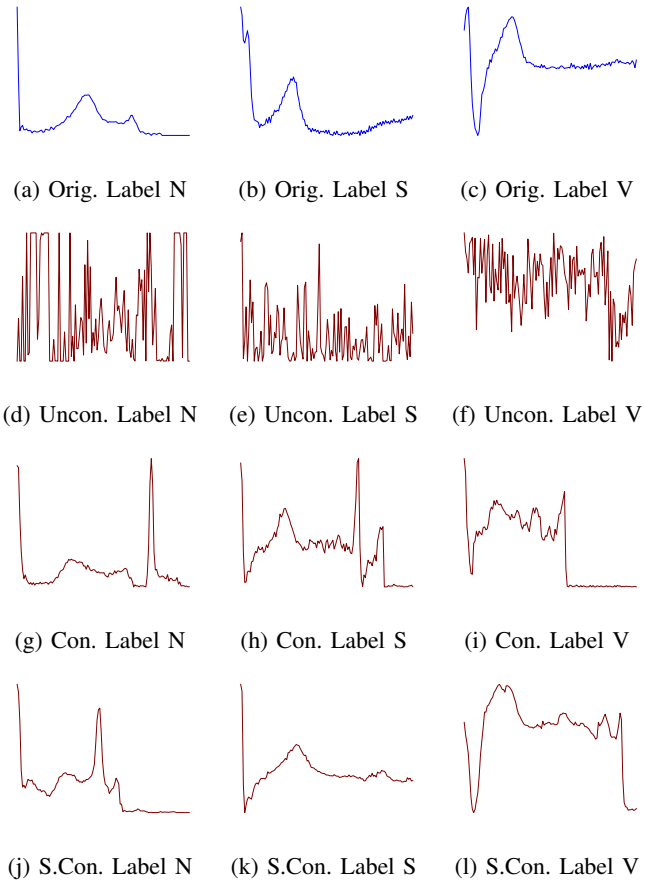
[1]Project source code: https://github.com/imics-lab/SmoothConditionalDiffusion

differentially private datasets that do not contain sensitive information. [3] presented several state-of-the-art GAN models and algorithms for generating time-series data, such as ([18]), RCGAN ([4]), TimeGAN ([26]), and SigCWGAN ([22]), which all use recurrent neural networks (RNNs) as the base architecture of their GAN models. However, RNN-based GAN models face challenges in producing long synthetic sequences that are realistic enough to be useful due to the sequential processing of time steps.

The transformer architecture, which relies on multiple self-attention layers ([24]), has recently become a prevalent deep learning model architecture. Since the transformer was invented to handle long sequences of text data and does not suffer from a vanishing gradient problem, theoretically, a transformer GAN model should perform better than RNN-based models on time-series data. [12] introduced a transformer-based GAN model (TTS-GAN) to generate synthetic time-series data that achieves more realistic synthetic data quality than previous RNN-based GANs. Furthermore, [13] also introduced a conditional transformer GAN model that can generate multi-category multi-variable arbitrary length time series.

It is noteworthy that the diffusion model has recently outperformed GANs in image synthesis, and more data synthesis researchers are starting to employ diffusion models instead of GANs. The utilization of the diffusion model in time-series synthesis has also demonstrated significant improvements, such as in audio synthesis, time-series forecasting, and time-series imputation.

To the best of our knowledge, our work is the first to employ label noise mitigation techniques in the process of signal generation to create a platform that is resilient to the presence of uncertainty in the label of the original data. Label smoothing has been shown to enhance the separation of classes in the learned feature spaces of deep classification models [21] by allowing the model to learn representations that recognize other labels as possibilities. Intuitively, a technique that can alleviate label noise during classification [14] could also be utilized to enhance the separation of classes in generated data.



(a) Orig. Label N (b) Orig. Label S (c) Orig. Label V

(d) Uncon. Label N (e) Uncon. Label S (f) Uncon. Label V

(g) Con. Label N (h) Con. Label S (i) Con. Label V

(j) S.Con. Label N (k) S.Con. Label S (l) S.Con. Label V

Fig. 1: Waveforms of sample signals from the MIT-BIH, [19] dataset. An original signal has been plotted in blue, and below, generated signals from each diffusion model are plotted in red. Every column shares a label, but otherwise, is not intended to be visually similar. The column labels are Normal (N), Atrial Premature (S), and Ventricular Premature (V). while the rows are Unconditional, Conditional, and Smoothed Conditional Diffusion.

III. Methodology

A. Data Generation

Denoising diffusion probabilistic models (DDPMs) generate new data samples that simulate training samples by learning to iteratively add noise and then remove it from the example data. The denoising model learns to estimate the added noise between time steps as noise is added. For a noise model that increases the noise in a data sample from time step t to $t+1$, the denoising model learns to estimate a noise pattern that would take the sample from time step $t+1$ back down to t. Since the added noise follows a Gaussian distribution, and the sum of any Gaussian distribution is also a Gaussian distribution, a sample can be placed at time step t during training with iterative noise addition [8].

Conditional DDPM incorporates the signal's label during training and sampling using a positional embedding function. In this project, we have utilized Sinusoidal Positional Embedding. The supplied label to the embedding function can be categorical, one-hot encoded, or smoothed.

Unconditional diffusion modeling can still be leveraged to generate new data samples targeting specific labels from the example dataset by employing an ensemble of diffusion models, with each model being dedicated to a particular class. However, this reduces the availability of training data for each instance and badly degrades the performance of the denoising model. For a balanced dataset of n instances in c classes, each model will only train on n/c instances. In unbalanced datasets, it is possible that one denoising model might only train on a few instances, making a very poor data generator.

Label smoothing is a regularization technique that augments the label of each instance following the assumption that every instance will have some likelihood of having been sampled from each class. Label smoothing has been shown to improve

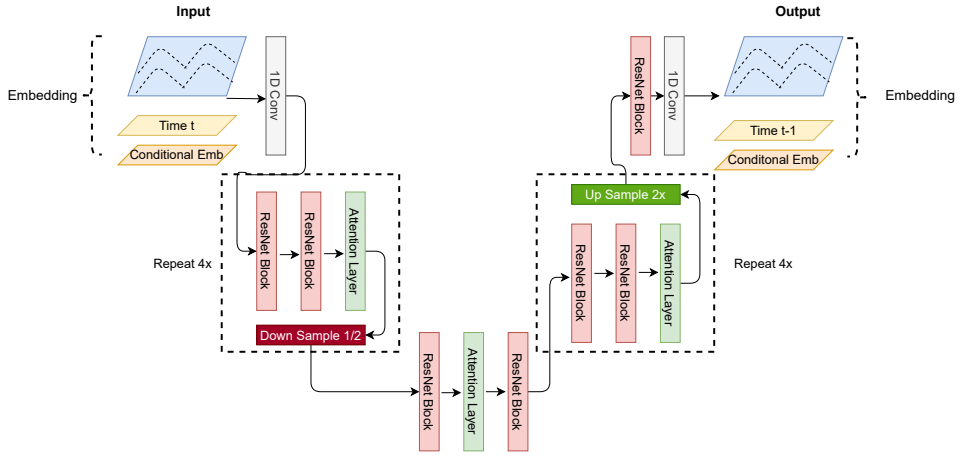


Fig. 2: A 1D-Unet Architecture was used as the denoising model in all three DDPMs. This model is fitted to a noise distribution during training and is used to synthesize new examples from random input afterwards.

the generalization of models [21] and is closely related to the broader family of loss-correction techniques [14]. Smoothed labels are calculated by counting the frequency of each label in the dataset. Each label is assumed to have a likelihood equal to its frequency times some scaling factor α, with the exception of the assigned label which is given likelihood $1-\alpha* frequency$. For a label y with scaling factor α and frequencies f_0 to f_n for n classes, the smoothed label is computed as:

$$< y_0, ..., y_n >= \begin{cases} 1 - (\alpha * f_i) \text{ if } i = y \\ \alpha * f_i \text{ otherwise} \end{cases} \quad (1)$$

Our technique integrates label smoothing into the reverse diffusion process, which learn to predict the noise model that has been introduced into examples during training, or from samples of Gaussian noise during sampling. Under conditional DDPM, the label is embeded with the signal as an input to the denoising model. We smooth the label before embedding, preventing the denoising model from developing overconfidence in noisy labels.

B. Experimental Design

To demonstrate the advantages of Smooth Conditional Diffusion, three diffusion models were prepared, and their outputs were compared. The investigated data generation models were: unconditional DDPM, conditional DDPM, and smoothed conditional DDPM. The denoising model utilized for all three approaches was a 1D-UNet model [9] with eight ResNet blocks. The complete architecture of the diffusion models is illustrated in Figure 2. During the training of the model and for data generation, 1,000 time-steps of noise addition were employed. The denoising models were fitted using 150 training epochs. For smooth conditional diffusion, an α value of 0.1 was used.

These three diffusion models were trained on various datasets of signal data: one synthetic, one for arrhythmia detection, and two for human activity recognition. Following training, each of the three diffusion models was used to generate additional instances of data for each dataset, equivalent to the size of the original training dataset.

The synthetic dataset was generated using the techniques described in [10]. This technique emulates the characteristics of real-world signal datasets but cannot be adjusted to examples of data and cannot be utilized to imitate samples of a specific dataset. Instead, this technique is useful for generating data for training other techniques without any chance of mislabeling. A total of 5,001 instances of artificial data were generated in five classes.

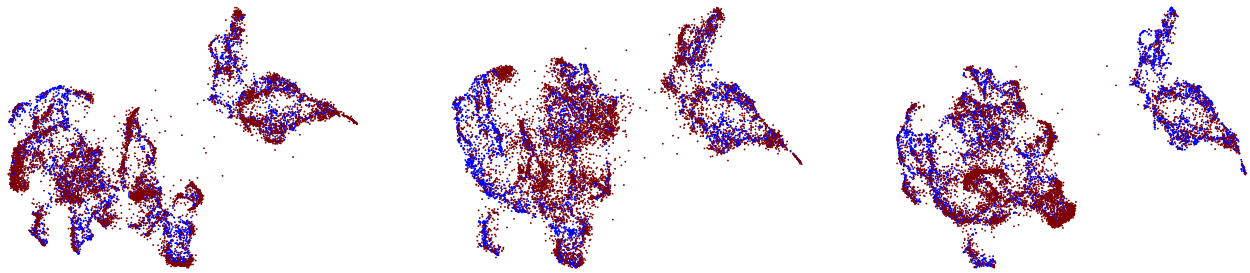
The MIT-BIH dataset [19] was obtained from the Beth Israel Hospital between 1975 and 1979 from 47 subjects using electrocardiography at a rate of 360 samples per second. Two cardiologists manually labeled the dataset as containing healthy heartbeats or presenting one of four forms of arrhythmia.

The UniMiB SHAR dataset [16] contains accelerometer data gathered from 30 subjects executing one of nine common activities using commercial smartphones. The phone was carried in the right hip pocket of the subjects. The data were segmented into three-second windows centered around signal peaks.

The TWristAR dataset [7] was gathered at Texas State University in 2022. Three subjects wore an Empatica E4 wristband on their left wrist while executing one of six physical activities. Acceleration data was collected at a frequency of 32 Hz. Video data with timestamps was provided alongside the accelerometer data to confirm the labels.

The datasets were subjected to symmetric label noise, with a uniform 5% mislabeling rate. Labels were randomly selected and reassigned with equal probability to any other label. This type of label noise occurs frequently in real-world applications [1] and was chosen for its ease of introduction at exact mislabeling rates across multiple types of signal data.

To demonstrate that the diffusion models learned to reproduce the distribution of the example data, we adapted the



(a) Unconditional (b) Conditional (c) Smooth Conditional

Fig. 3: 2D projections of original UniMiB SHAR data plotted in blue against the output of three DDPM data generators in red. The greater overlap in the Smooth Conditional plot shows that the generated data is closer to the distribution of the examples.

Fréchet Inception Distance (FID) [6], a common metric for evaluating synthetic image data. FID measures the distance between the feature distributions learned from the example data and those from synthesized data with the Inception v3 model [25] for generating features to compute FID for consistency with image generation. For this project, we have selected a pre-trained 1D feature extractor that is appropriate for signal processing.

Previous work has shown that FID adaptations are valid even when substituting the pre-trained model with another one that is more appropriate for a different data domain, such as audio [11]. Following this concept, we selected a published feature learner that is suitable for time-series data collected from wearable sensors. Wave2Vec [27] was developed for speech recognition and audio processing, but its feature extractor's sampling rate can be adjusted to an appropriate range for arbitrary signals. Our team's work has demonstrated the suitability of Wave2Vec for ECG and accelerometer data, and it may also be applicable to a broad range of signals for comparing signal generation techniques using FID.

We computed the FID of each diffusion model by comparing the Wave2Vec features learned on the example data to those learned from the synthesized data. Additionally, we prepared UMAP [15] projections by overlaying the example and synthetic data in two dimensions. These projections show how closely each generator's output aligns with the distribution of its example data.

IV. Results

Our findings demonstrate that label smoothing enhances the capability of conditional DDPM to generate new data samples that closely conform to the distribution of the example data, even when some instances in the example data are mislabeled.

As shown in Figure 1, all three DDPM models generate acceptable signal data. The ECG signals depicted in this figure share the same label, and the synthetic instances (in red) have been overlaid on an example from the original data. All six signals in the figure share the same label. It can be observed that the unconditional DDPM has overfitted to mislabeled instances in the example data and is reproducing features that do not correspond to the depicted class.

In Figure 3, we present 2-dimensional projections of the UniMiB dataset plotted against synthetic data generated by three DDPM approaches. UMAP is used to learn a lower-dimensional projection while preserving the relative distances of each instance in the projection. Since instances that are close in these projections are also close in the original feature space, we can infer the structure of the raw data by observing the plots. Data points generated using Smooth Conditional DDPM overlap more extensively with the real example data (plotted in blue), indicating that the distribution of the two samplings of data is closer. These figures can be challenging to interpret, but a data generator can be considered effective if the synthetic points in the plot (in red) are generally situated on or near the blue points, and very few areas of blue points remain uncovered.

Table I displays the Fréchet Inception Distance (FID) of the output of each diffusion model. Lower FID values indicate that the output of a generator is closer to the distribution of the example data. The two conditional DDPMs consistently outperform the unconditional DDPM. Moreover, the FID of the output of the conditional DDPM is, on average across all four datasets, 3.5% lower than the unconditional DDPM, while the output of the smooth conditional DDPM is, on average, 14.6% closer to the example data. These results demonstrate that label smoothing improves the performance of conditional DDPM in the presence of label noise by preventing the denoising model from overfitting to mislabeled instances in the example data.

It is noteworthy that the unconditional DDPM takes substantially longer to train since an ensemble of denoising models is trained following the approach described in Section III. One denoising model is trained for each class represented in the example dataset, making the difference noticeable in the five to nine class data used in this project. However, larger numbers of classes could create increasingly unwieldy ensembles of denoising models when using unconditional DDPM.

Dataset	Uncon. FID	Con. FID	S. Con. FID	Delta Con. FID	Delta S. Con. FID
Synthetic	1.51	1.60	1.39	+6.0%	-7.9%
MIT-BIH	7.85	6.91	6.82	-12.0%	-13.1%
UniMiB	0.81	0.85	0.62	+4.9%	-2.3%
TWristAW	2.53	2.20	2.18	-13.0%	-13.8%
Average	3.18	2.89	2.75	-3.5%	-14.6%

TABLE I: The FID of the output of each diffusion model on each dataset. This value represents the Wasserstein distance of the output of the Wave2Vec feature learner for features from the example data and the synthetic data. Smaller values show that the output of a generator is closer to the distribution of the original data. The two right columns show the percent change from the FID of the unconditional diffusion model. The last row shows the average across all datasets.

The additional difficulty of training an ensemble of denoising models makes conditional DDPM a preferable choice to unconditional when class-targeted data generation is needed. It is safe to assume that some label noise will exist in real-world data used as examples for the generator. Time series data, in particular, is susceptible to label noise [1], making label smoothing a broadly applicable tool for generating signals. Table I demonstrates that conditional DDPM with label smoothing produces synthetic labels in a distribution that is closer to that of the original data than conditional DDPM without label smoothing.

V. Conclusion

This work has demonstrated a new data generation technique based on conditional DDPM, which incorporates label smoothing to improve the resilience of the generation technique to label noise. We have shown that this new technique generates new instances of data that more closely fit the distribution of the original example data than existing techniques. Finally, we have shown that FID can be made appropriate for use with signal data by substituting a pre-trained feature learner from the time-series domain.

References

[1] Gentry Atkinson and Vangelis Metsis. A survey of methods for detection and correction of noisy labels in time series data. In *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17*, pages 479–493. Springer, 2021.

[2] Gentry Atkinson and Vangelis Metsis. Tsar: a time series assisted relabeling tool for reducing label noise. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 203–209, 2021.

[3] Eoin Brophy, Zhengwei Wang, Qi She, and Tomas Ward. Generative adversarial networks in time series: A survey and taxonomy. *arXiv preprint arXiv:2107.11098*, 2021.

[4] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[5] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[7] Lee B. Hinkle, Gentry Atkinson, and Vangelis Metsis. Twristar - wristband activity recognition, January 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[9] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[10] Mohammed Waleed Kadous. Learning comprehensible descriptions of multivariate time series. In *ICML*, volume 454, page 463, 1999.

[11] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, pages 2350–2354, 2019.

[12] Xiaomin Li, Vangelis Metsis, Huangyingrui Wang, and Anne Hee Hiong Ngu. Tts-gan: A transformer-based time-series generative adversarial network. *arXiv preprint arXiv:2202.02691*, 2022.

[13] Xiaomin Li, Anne Hee Hiong Ngu, and Vangelis Metsis. Tts-cgan: A transformer time-series conditional gan for biosignal data augmentation. *arXiv preprint arXiv:2206.13676*, 2022.

[14] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.

[15] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[16] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences*, 7(10):1101, 2017.

[17] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[18] Olof Mogren. C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*, 2016.

[19] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

[20] Nicolas M Müller and Karla Markert. Identifying mislabeled instances in classification datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[21] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[22] Hao Ni, Lukasz Szpruch, Magnus Wiese, Shujian Liao, and Baoren Xiao. Conditional sig-wasserstein gans for time series generation. *arXiv preprint arXiv:2006.05421*, 2020.

[23] Ki Nohyun, Hoyong Choi, and Hye Won Chung. Data valuation without training of a model. In *International Conference on Learning Representations*.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[25] Xiaoling Xia, Cui Xu, and Bing Nan. Inception-v3 for flower classification. In *2017 2nd international conference on image, vision and computing (ICIVC)*, pages 783–787. IEEE, 2017.

[26] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

[27] Ye Yuan, Guangxu Xun, Qiuling Suo, Kebin Jia, and Aidong Zhang. Wave2vec: Learning deep representations for biosignals. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1159–1164. IEEE, 2017.