2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Data Augmentation for Cardiovascular Time Series Data using WaveNet

1st Robert Feldhans
Machine Learning Group
Bielefeld University
Bielefeld, Germany
rfeldhans@techfak.uni-bielefeld.de

2nd Alexander Schulz
Machine Learning Group
Bielefeld University
Bielefeld, Germany
aschulz@techfak.uni-bielefeld.de

3rd Johannes Kummert
Machine Learning Group
Bielefeld University
Bielefeld, Germany
jkummert@techfak.uni-bielefeld.de

4th Moriz Habigt
Anaesthesiology Clinic
RWTH Aachen University
Faculty of Medicine
Aachen, Germany
secretary@irt.rwth-aachen.de

5th Maike Stemmler
Institute of Automatic Control
RWTH Aachen University
Aachen, Germany
M.Stemmler@irt.rwth-aachen.de

6th Christina Kohler
Institute of Automatic Control
RWTH Aachen University
Aachen, Germany
C.Kohler@irt.rwth-aachen.de

7th Dirk Abel
Institute of Automatic Control
RWTH Aachen University
Aachen, Germany
d.abel@irt.rwth-aachen.de

8th Rolf Rossaint
Anaesthesiology Clinic
RWTH Aachen University
Faculty of Medicine
Aachen, Germany
secretary@irt.rwth-aachen.de

9th Barbara Hammer
Machine Learning Group
Bielefeld University
Bielefeld, Germany
bhammer@techfak.uni-bielefeld.de

***Abstract*—In this work we present a novel approach for generating cardiovascular data using a modified WaveNet architecture. This can enable further research in areas where data is scarce and hard to obtain. By generating additional time series data in a set of animal tests performance of existing models could be improved and more difficult approaches, that require substantial amounts of data, attempted. We validate our approach on a classification task and compare it to similar methods of data augmentation.**

***Index Terms*—Time Series, Data Augmentation, Medical Application**

I. Introduction

In the realm of medical applications, accurate and reliable classification of data is important for diagnosis, treatment, and patient care.

However, the efficacy of classifiers heavily relies on the availability of sizable and representative data sets. Ethical considerations play a pivotal role when creating these sets, i.e. privacy protection for patients and animal testing. The generation of synthetic data offers a more ethical alternative to traditional methods of data expansion, particularly when these expose their subjects to potential harm. Moreover, synthetic data generation can accelerate the data collection process, making it a faster and more cost-effective strategy compared to traditional methods.

Funding by the German Federal Ministry for Education and Research (BMBF) under reference number 01|S21056C is gratefully acknowledged.

One potential application is the development and utilization of left ventricular assist devices (LVADs), as gathering relevant data such as the pressure of a ventricle is invasive and thus difficult and expensive in terms of availability and feasibility [1].

To this end, we propose a modified WaveNet model [2] to generate synthetic medical data, particularly tailored to physiological signals that exhibit sinusoidal attributes, such as the behavior of a heart ventricle. We validate our trained model on a downstream task using data augmentation to show its feasibility.

This paper is structured as follows: We first discuss related work with regard to data synthetisation in general and for medical applications in particular. Then, we describe our method for generating novel data points and the experiments conducted to validate our approach. Lastly we discuss potential ways for further improvement and future work.

II. Related Work

Different architectures of deep learning have been applied to data augmentation problems regularly in the past. Certain architectures have proven themselves to be advantageous in specific domains, such as generative adversarial networks (GANs) [3] and de-noising diffusion models [4] in the domain of image generation [5], [6]; or the transformer architecture [7] in text generation.

Reference [8] gives an overview of commonly used architectures for time series data augmentation models.

Reference [9] employs data augmentation techniques for improving the accuracy of a classifier, focussing on LSTM [10] architectures.

Conventional convolutional neural nets (CNNs) have mostly been used for time series classification [11] which use data augmentation to improve the CNNs performance rather than using the CNN for data augmentation. Nevertheless, there are approaches which use CNNs for time series forecasting [12], which can be applied to data augmentation. There are also approaches with hybrid architectures which focus on combining CNNs and LSTMs [13], [14], or on convolutions and GANs [15].

WaveNet has been used for data augmentation mostly in the domain of speech processing [16]. Another approach employs a modified WaveNet on non-speech applications, stepping outside the time series domain by using non-causal convolutional layers and using WaveNet for predictions, only employing data augmentation techniques for training the model [17].

Data augmentation techniques employing deep learning have also already been applied to the medical domain, however, these seem to mostly focus on image data [18], [5], [6]. This is also the case for cardiovascular data [19]. Other methods, such as physical simulation, however, have been applied to time series data [20].

III. Data

The data set we employ consists of cardiovascular data gathered from pigs [1], [21]. It is presented as time series data of several different sensors, ranging from non-invasive to very invasive collection methods. The most invasive of these dimensions are the left ventricular pressure (LVP) and the calibrated left ventricular volume (LV). As such, accurately generating these dimensions is of particular interest. The right ventricular pressure (RVP) is very invasive as well, but less important for the application in left ventricular assist devices. Among the less invasive dimensions are the pulmonary arterial pressure (PaP) and flow (PaQ).

The data is generated by artificially inducing specific heart conditions, also called interventions, in each subject. In each data collection session, only one intervention is induced. The distribution of these interventions across the data collection sessions can be seen in table (I). Intervention 1 corresponds to pre-load reduction. Intervention 2 and 3 are different kinds of after-load increases, intervention 4 are speed-ramps and intervention 10 corresponds to changes in contractility.

Each data collection session consists of five phases, out of which we selectively utilize only two in our analysis: Phase one contains the data from before the induction of the actual intervention. As such, it can be seen as a baseline, which would correspond to normal or healthy behavior of the heart and is consistent across all different kinds of interventions. Phase three on the other hand contains the actual intervention of each data collection session.

TABLE I: Distribution of interventions across data collection sessions and available samples.

Intervention	1	2	3	4	10	Phase 1
# Sessions	158	5	109	3	31	all / 306
# Samples	2111	33	2164	55	1268	2651

We downsample the data from 1khz to 50hz, as there is no meaningful loss of information due to jitters in the data and reducing the size of input data improves the efficiency of the training algorithm. Additionally, we only use a selection of the available dimensions, i.e. the aforementioned LVP, LV, RVP, PaP and PaQ.

For training our modified WaveNet, we split the data into smaller windows. As input for the model we take signals of length 200, i.e. 4 seconds and predict the next 100 timesteps, i.e. 2 seconds. To increase the amount of training data, we split the signal using a sliding window of half the input length, effectively doubling the amount of available training data.

IV. Method

The architecture for our approach is a modified version of the original WaveNet architecture, specifically their global conditioning variant we implemented using PyTorch [22].

The first modification is the change from a single dimension input, such as a typical audio signal, to the multi-dimensional data presented by the medical data set. As such,

$$p(\vec{\mathbf{x}}|\mathbf{h}) = \prod_{t=1}^{T} p(\vec{x_t}|\vec{x_1}, ..., \vec{x_{t-1}}, \mathbf{h}) \tag{1}$$

describes the multi-dimensional conditional distribution $p(\vec{\mathbf{x}}|\mathbf{h})$ of the medical data given condition h (i.e. intervention) and the previous signal points $\vec{x_i}$ (see Eq. (3) [2]).

Likewise, the conditional activation function (see Eq. (2) [2]) is described as

$$\mathbf{z} = \tanh(W_{f,k} * \vec{\mathbf{x}} + V_{f,k}^T \mathbf{h}) \odot \sigma(W_{g,k} * \vec{\mathbf{x}} + V_{g,k}^T \mathbf{h})$$

where $\vec{\mathbf{x}}$ is the n-dimensional waveform, $V_{*,k}$ are learnable linear projections, $W_{*,k}$ are learnable convolution filter, k is the layer index and f and g denote filter and gate respectively. To guarantee the correct interaction between the input signal's dimensions, the convolutional layers in our proposed WaveNet modification convolute all input to all output dimensions.

A. Output Layer Changes

The second modification to the original WaveNet architecture lies in the omission of the final softmax layer. Reference [23] shows that transforming the regression problem into a multi-modal classification problem increases performance, the particular features of this data set make this approach unfeasible.

Adding the softmax layer requires a discrete normalization, which forbids the model from producing data outside of those discrete values. However, this discretization might not be desirable, depending on normalization details. Changes in the distribution of a models input -by extension of each of the models layers' input- forces the model to continuously adapt to these changes, ultimately leading to slower learning and worse model performance [24].

Normalization can be achieved in a number of ways. Each data collection session could be normalized to its own max and min values. However, this would mean that the signals of some data collection sessions, would no longer be distributed like others, as e.g. an exceptionally high LVP would shift the center of the whole signal severely up. This is especially the case for signals of intervention 10, as they strain the heart of the subjects, which leads to, among other things, higher than normal pressure and lower then normal volume in the heart. In this case, data of phase one would no longer be consistent across data collection sessions, harming performance.

Another approach would be to normalize across all data collection sessions. In this case most of the data would be squished, loosing details of the signal and extreme cases may even lead to computational problems due to rounding errors. Additionally, slight differences in the test subjects and application of the sensors during data collection manifest in distribution changes between data collection sessions. An example of such a distribution difference between two data collection sessions can be seen in Fig. (1). While the minima of both signals are virtually identical (6.53 vs 6.85), the maxima differ substantially (60.5 vs 51.55).

Discretizising the signal and leaving some headroom, i.e. not using the most extreme values is plagued by the same problems outlined above; Even more so, since these extreme values can have a very high false positive rate [23].

To increase the amount of samples, the model is never fed the whole signal, but only windows which could be normalized individually. However, in this case much of the information contained in the signal would be lost. It would, e.g. not be clear if the window is in an area of exceptionally low or high pressure, leading again to poor performance.

We opted therefore to normalize the signals of each data collection session according to the values of its phase one, which consequently worked best in our experiments. While this carries the problem of introducing data from outside the normalization range as input and as target, due to exposing the model to this kind of data during training, those can not be considered as unexpected data. Moreover, all signals of a data collection session are distributed equally to those of all other data collection sessions of the same intervention, especially those signals of phase one. However, introducing data from outside the normalization range does permit us from effectively using discretization.

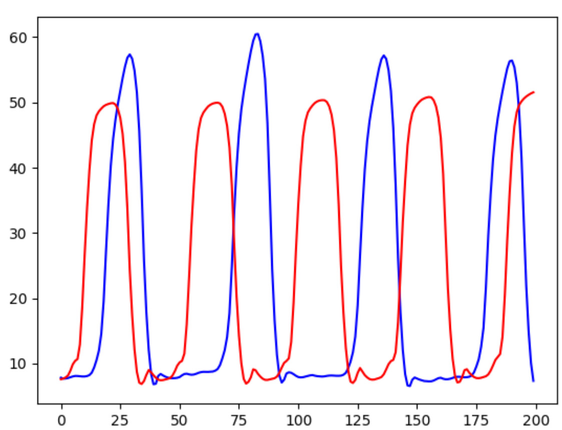


Fig. 1: Example of different value distributions across different data collection sessions. Pictured are two LVP signals of phase one gathered from different subjects as well as intervention 1 (blue) and 10 (red).

B. Loss changes

The third modification to the WaveNet model lies in the loss used for training. Our modified WaveNet model is -in contrast to the original WaveNet- trained using a single Soft-DTW loss [25], as our domain does not provide us with an analogue to the original WaveNets frame position. As the data is collected from living subjects, their heart-rate is not completely constant and may speed up or slow down during a data collection session, resulting in a x-axis shift in the target signal, thus making the Soft-DTW loss suitable to this application.

C. Manual ECG Construction

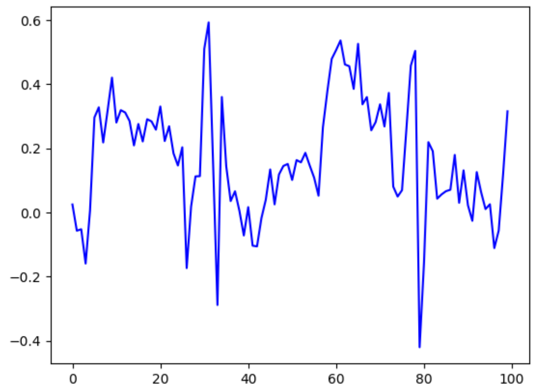
Despite the original data set including heart rate information in the form of an ECG signal, we opted to manually construct such signals in the form of sine waves. The sensors measuring ECG may produce slightly different output in between data collection sessions, based on their placement, application and different specimens. The simpler and less noisy synthetic heart rate produced better results in our tests then the original, slightly noisy data. Our synthetic ECG signal is generated as follows: The signals for LVP and LV are sinoid and tend to have high derivations of the first order around their roots. We thus calculate the roots of the LVP signal and draw a sine wave between every two roots. Such a exemplary manually constructed heartrate signal can be seen in Fig. (2).

V. Experiments

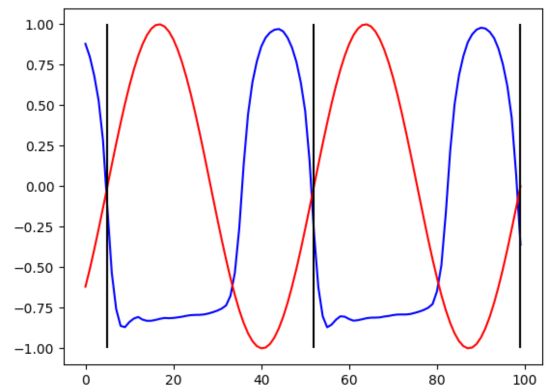
In this section, we will go into some architecture details and then present the evaluation of our approach.

A. Architecture Details

The proposed WaveNet model utilizes ten stacked residual WaveNet blocks. We decided to only generate a subset of the available dimensions, i.e. LVP, LV, RVP, PaP, PaQ as described in chapter (III). This improves the computational efficiency of our approach, while not decreasing the accuracy of the resulting generator, especially for the LVP and LV dimensions. Additionally, we use the intervention as the conditional input, which enables us to produce data of a specific intervention. However, we chose to model the data of phase one as an



(a) Original ECG signal.



(b) Generated ECG signal (red) with the corresponding LVP signal (blue) and identified roots (black).

Fig. 2: Original (left) and manually constructed (right) ECG signal for training of the model.

additional kind of intervention, the reason being two-fold: Firstly to provide more training data, and secondly to serve as baseline data, as all interventions are similar, pathological variations of the phase one data. We performed a hyper-parameter grid-search on the previously mentioned parameters and chose the model which performed best in replicating an unknown test signal for the following experiments.

B. Evaluation with Random Forest

For the evaluation of our approach we conduct a simple classification experiment using random forest (RF) [26]. To this end we train a number of RF classifiers which classify a given input signal of length 200 by their intervention. This is a relevant downstream task as it is not trivial and can be helpful for low powered implanted devices such as LVAD controllers [21]. We specifically use scikit-learns implementation [27] of the RF algorithm and use their default parameters. We chose to employ only the most medically invasive dimensions LVP and LV, and utilize a train-test split of 80% to 20%.

We first train RF on the training set of the original data. Then, we augment the training data set using our modified WaveNet and thus double the available training data, with which we then train another RF, however, we do not interfere with the class imbalance. See Fig. (3) for an example of generated data. We also augment the original data set using SMOTE [28], which is particularly suitable to augment our highly imbalanced original data set; As well as a LSTM model, trained similarly to our modified WaveNet model, based on PyTorch's [22] implementation. We also provide a baseline in form of the unmodified Wavenet, trained with a maximum log likelihood loss and without our proposed output layer changes. We opted to normalize akin to the modified model, truncating the signal when outside the normalization range. Lastly, we trained a CNN utilizing temporal convolution layers. We then test each classifier on the test set of our unaugmented original data. The results of ten repetitions of these experiments can be seen in table (II). The difference between the overall score of the classifier trained on unaugmented data to each of the classifiers trained on modified WaveNet, SMOTE and LSTM augmented data is significant with at most $p < 4.3 \cdot 10^{-4}$, which we calculated using scipy's [29] Welch t-test [30].

C. Discussion of Results

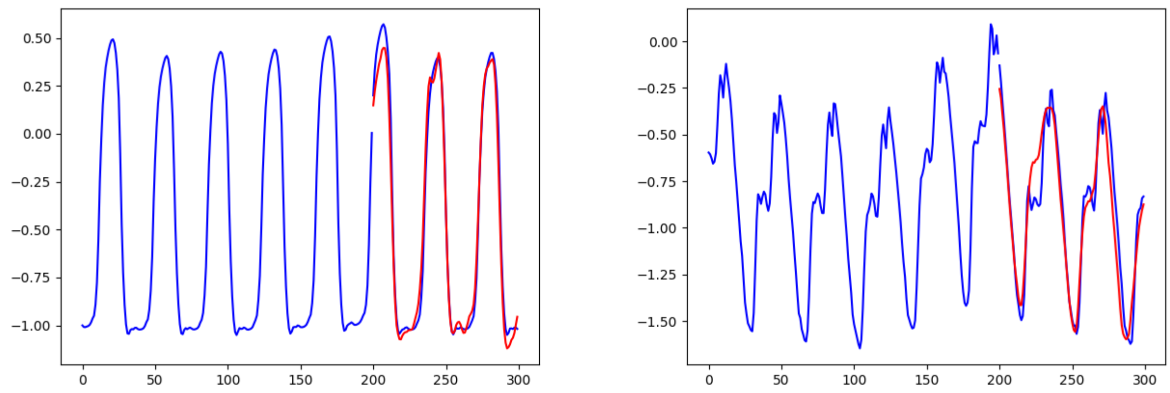
Due to the heavy class imbalance, the classifier trained only on the original data performs poorly on interventions 2 and 4. Consequently, while its overall accuracy is quite high, its average class accuracy suffers.

While the classifier trained on SMOTE-augmented data performs worse over all data, it improves the performance on the underrepresented interventions 2 and 4. This would be expected due to the oversampling methods employed by SMOTE.

While the classifier trained on LSTM's data augmentation performs slightly worse over all data than the baseline of no augmentation, the generated data is capable of increasing performance on intervention 10. Considering that half the data it is trained on was generated, this classifier is still nearly as capable of classifying original data. As such, the data generated by the LSTM seems to only impede the classifier slightly.

The classifier trained on data augmented with the original WaveNet improves the accuracy on the underrepresented interventions 2 and 4, leading to an improved average class accuracy compared to no augmentation whatsoever, roughly on the same level as the SMOTE augmentation.

The classifier trained on CNN-augmented data especially improves on the underrepresented interventions 2 and 4, but shows a decrease in performance on the other interventions as well as overall performance. However, it achieves a much higher consistency in average class accuracy when compared to the classifier trained on unaugmented data.



(a) LVP Dimension

(b) LV Dimension

Fig. 3: Example of a generated signal (red) in contrast to the original signal (blue). The first 200 data points are the models input, while the last 100 are unknown to it and included here as a reference.

TABLE II: Results of the classification experiment, by class i.e. intervention and method used for data augmentation.

	no augmentation	SMOTE	LSTM	original Wavenet	CNN	modified Wavenet
Intervention 1	0.9118 ±0.0052	0.8935 ±0.0065	0.9013 ±0.0054	0.89 ±0.0081	0.7936 ±0.0082	**0.9222 ±0.0115**
Intervention 2	0.1556 ±0.0544	0.2 ±0.0667	0.0888 ±0.0444	0.3 ±0.0999	0.2143 ±0.0714	**0.4 ±0.3000**
Intervention 3	0.9359 ±0.0022	0.9243 ±0.0049	0.9243 ±0.0023	0.9359 ±0.0029	0.8448 ±0.0111	**0.9667 ±0.0069**
Intervention 4	0.0 ±0.0	0.0833 ±0.0833	0.0 ±0.0	0.0083 ±0.025	0.3250 ±0.1017	**0.82 ±0.14**
Intervention 10	0.9474 ±0.0068	0.9380 ±0.0061	0.9623 ±0.0055	0.9115 ±0.0054	0.8056 ±0.0099	**0.9647 ±0.0126**
Phase 1	**0.9996 ±0.0008**	0.9847 ±0.0020	0.9983 ±0.002	0.9802 ±0.0019	0.7969 ±0.0143	0.9720 ±0.0079
Overall	0.9414 ±0.0018	0.9289 ±0.0033	0.9372 ±0.0023	0.9256 ±0.0028	0.8055 ±0.0045	**0.9518 ±0.0046**
Average	0.6584 ±0.4138	0.6707 ±0.3765	0.6458 ±0.427	0.6709 ±0.3760	0.6300 ±0.2574	**0.8409 ±0.2040**

While the WaveNet-augmented classifier performs worse on data of phase one, it improves on all other interventions and in consequence performs better over all data compared to the classifier trained on unaugmented data. Additionally, it performs better than the baseline classifiers trained on unaugmented, SMOTE-, LSTM- and CNN-augmented data in almost all test cases and considered metrics. In particular, the proposed WaveNet data augmentation combines an increase in average class accuracy with a CNN-like boost in consistency.

VI. Conclusion

In this paper, we proposed a modified WaveNet architecture for generating cardiovascular time series data and described challenges and modification choices for such a domain. We then evaluated our approach on a domain-relevant downstream task with real-world data, showing it to improve upon data augmentation methods established in the time series domain.

Future work should focus on truly continuous data generation. While it is generally possible to generate continuous signals with our approach by simply feeding the generated signal back into our model, it is currently not possible to replicate complete data collection sessions accurately due to the lack of data containing transitional periods between phase one and each intervention. Addressing this limitation could increase the performance of LVAD controllers and potentially reduce detection time of cardiac events.

The proposed WaveNet architecture might also be used to facilitate training and testing LVAD controllers as mentioned in the motivation for our evaluation task.

References

[1] M. A. Habigt, J. Gesenhues, M. Ketelhut, M. Hein, P. Duschner, R. Rossaint, and M. Mechelinck, "In vivo evaluation of two adaptive starling-like control algorithms for left ventricular assist devices," *Biomedical Engineering / Biomedizinische Technik*, vol. 66, no. 3, pp. 257–266, 2021. [Online]. Available: https://doi.org/10.1515/bmt-2020-0248

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[4] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8857–8868.

[5] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "Gan-based synthetic brain mr image generation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 734–738.

[6] F. Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn, "Medical diffusion: Denoising diffusion probabilistic models for 3d medical image generation," 2023.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[8] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting covid-19 time-series data: A comparative study," *Chaos, Solitons & Fractals*, vol. 140, p. 110121, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096007792030518X

[9] K. Bandara, H. Hewamalage, Y. Liu, Y. Kang, and C. Bergmeir, "Improving the accuracy of global forecasting models using time series data augmentation," *CoRR*, vol. abs/2008.02663, 2020. [Online]. Available: https://arxiv.org/abs/2008.02663

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

[11] A. Le Guennec, S. Malinowski, and R. Tavenard, "Data Augmentation for Time Series Classification using Convolutional Neural Networks," in *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, Riva Del Garda, Italy, Sep. 2016. [Online]. Available: https://shs.hal.science/halshs-01357973

[12] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *CoRR*, vol. abs/1803.01271, 2018. [Online]. Available: http://arxiv.org/abs/1803.01271

[13] K. Deng, T. Fei, X. Huang, and Y. Peng, "Irc-gan: Introspective recurrent convolutional gan for text-to-video generation." in *IJCAI*, 2019, pp. 2216–2222.

[14] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[15] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional gans," *arXiv preprint arXiv:1706.02633*, 2017.

[16] J. Wang, S. Kim, and Y. Lee, "Speech augmentation using wavenet in speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6770–6774.

[17] S. Racheal, Y. Liu, and A. Ayodeji, "Improved wavenet for pressurized water reactor accident prediction," *Annals of Nuclear Energy*, vol. 181, p. 109519, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306454922005497

[18] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[19] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *Simulation and Synthesis in Medical Imaging*, S. A. Tsaftaris, A. Gooya, A. F. Frangi, and J. L. Prince, Eds. Cham: Springer International Publishing, 2017, pp. 3–13.

[20] D. Mancini and P. C. Colombo, "Left ventricular assist devices: A rapidly evolving alternative to transplant," *Journal of the American College of Cardiology*, vol. 65, no. 23, pp. 2542–2555, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0735109715020665

[21] M. A. Habigt, M. Hein, J. Gesenhues, D. Abel, R. Rossaint, and M. Mechelinck, "In vivo evaluation of a novel control algorithm for left ventricular assist devices based upon ventricular stroke work," *ASAIO Journal*, vol. 69, no. 1, pp. 86–95, 2023.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[23] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," 2016.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[25] M. Cuturi and M. Blondel, "Soft-DTW: a differentiable loss function for time-series," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 894–903. [Online]. Available: https://proceedings.mlr.press/v70/cuturi17a.html

[26] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[29] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[30] B. L. WELCH, "THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 01 1947. [Online]. Available: https://doi.org/10.1093/biomet/34.1-2.28