Augmentation Strategies for Self-Supervised Representation Learning from Electrocardiograms

Matilda Andersson, Mattias Nilsson
Neko Health
Stockholm, Sweden
matilda@nekohealth.com, matnil@nekohealth.com

Gabrielle Flood, Kalle Åström
Centre for Mathematical Sciences
Lund University
Lund, Sweden
gabrielle.flood@math.lth.se, karl.astrom@math.lth.se

***Abstract*—In this paper, we investigate the effects of different augmentation strategies in self-supervised representation learning from electrocardiograms. Our study examines the impact of random resized crop and time out on downstream performance. We also consider the importance of the signal length. Furthermore, instead of using two augmented copies of the sample as a positive pair, we suggest augmenting only one. The second signal is kept as the original signal. These different augmentation strategies are investigated in the context of pre-training and fine-tuning, following the different self-supervised learning frameworks BYOL, SimCLR, and VICReg. We formulate the downstream task as a multi-label classification task using a public dataset containing ECG recordings and annotations. In our experiments, we demonstrate that self-supervised learning can consistently outperform classical supervised learning when configured correctly. These findings are of particular importance in the medical domain, as the medical labeling process is particularly expensive, and clinical ground truth is often difficult to define. We are hopeful that our findings will be a catalyst for further research into augmentation strategies in self-supervised learning to improve performance in the detection of cardiovascular disease.**

***Index Terms*—self-supervised, representation learning, ECG, electrocardiogram, augmentation, pre-processing**

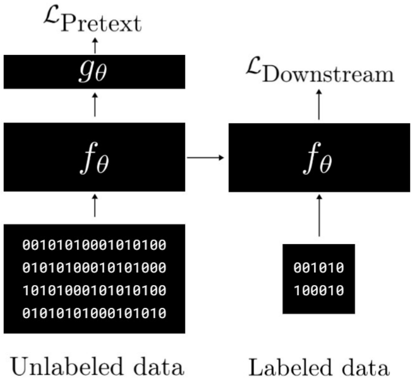

Fig. 1: Image of the self-supervised learning framework. In the self-supervised step, the encoder f_θ and method-specific layers g_θ are trained using unlabeled data. The encoder network is then fine-tuned for the downstream task using labeled data.

I. Introduction

Cardiovascular diseases are the leading cause of death worldwide, increasing yearly. However, many abnormalities in heart cycles can be discovered and treated years before the onset of diseases, using a preventive approach. There have been several attempts to produce automated ECG-based heartbeat classification methods over the last few decades, but their performance is hindered by limited access to high-quality labeled data, restricting their usage to secondary diagnostic purposes. In this regard, the development of self-supervised learning frameworks is important.

The application of machine learning to electrocardiograms (ECG) is an example of the ongoing development of artificial intelligence in cardiovascular medicine. State-of-the-art solutions mainly follow convolutional neural network-based architectures trained in a supervised manner [1]–[4] and offer valuable insights into cardiovascular health and disease detection. Current systems are highly dependent on large amounts of labeled data and further development of clinical AI-based ECG disease detection systems is hindered by its scarcity.

Given recent achievements of self-supervised learning in other fields, present research in the domain of cardiovascular disease detection aims to apply self-supervised learning methods to decouple system performance from the need for excessive amounts of labeled data. Mehari and Strodthoff [5] apply a selection of self-learning methods to 12-lead ECG data and evaluate their representational performance in a multi-label classification task setting. They find an adjusted version of the contrastive predictive coding, CPC [6], and the SimCLR approach to show the highest performance results. Research presented by [7] also considers BYOL and SimCLR approaches for ECG representation learning, but used a very shallow encoder network architecture with only five convolutional layers. Work by [8] highlights the potential for using self-supervised learning methods in ECG representation learning by presenting a careful comparison of the effects of different self-supervised learning methods on linear evaluation and fine-tuning evaluation.

In this paper, we present an assessment of self-supervised representation learning on 12-lead clinical ECG data to examine the importance of pre-processing and augmentation for self-supervised learning methods applied to ECG signals. The main contribution of this paper is a new augmentation strategy for self-supervised learning for cardiovascular disease detection. Specifically, we show that the performance on a

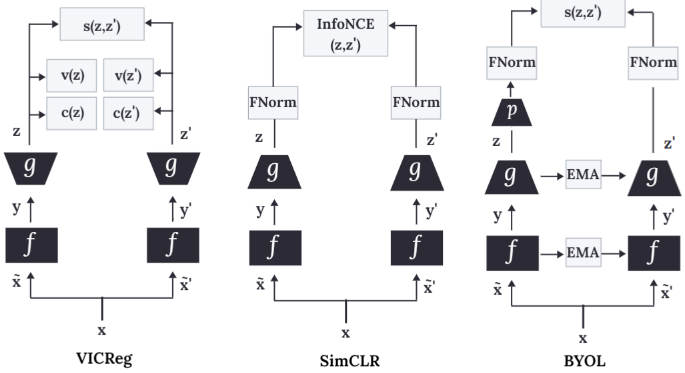

Fig. 2: Sample **x**, transformed into positive pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ is encoded by f into **y**, **y**$'$. These are projected by g to lower dimension representations **z**, **z**$'$ in SimCLR, BYOL, or expanded in VICReg. The loss applies to **z**, **z**$'$. FNorm and EMA stand for feature-normalized embeddings and exponential moving average, respectively. InfoNCE is the loss function in [16] and **c**, **v**, and **s** are covariance, variance, and similarity functions.

multi-label classification ECG task depends heavily on how the self-supervised learning setup is configured and that improved performance over classical supervised learning can be obtained by 1) increased input signal segment length; 2) data augmentation applied to only one of the two signal paths; and 3) tuned strength of the applied data augmentation.

II. Self-Supervised Representation Learning

Self-supervised learning [9], was first introduced to the field of natural language processing, where self-supervised trained models such as BERT [10] entailed significant performance improvements without the increasing need for labeled data. State-of-the-art self-supervised learning methods form representations through joint-embedding architectures, called Siamese networks. Representations are learnt by maximizing agreement between embeddings of different augmented copies of the same data example, also referred to as views, via a loss function in the latent space [11]–[14]. In general, data views are constructed by applying different augmentations (modifications) to the input data.

In self-supervised learning, we usually distinguish between the pretext task and the downstream task as displayed in Figure 1. A pretext task is a self-supervised learning problem in which the model constructs feature representations from unlabeled data inputs. By forcing a model to solve a deliberately designed pretext task it learns to extract task-agnostic feature representations of the data. Commonly, the unsupervised learning phase is followed by the supervised fine-tuning process. This is often the primary task to be solved and uses few labeled data samples [15].

In the context of self-supervised learning on ECG signals for the detection of cardiovascular diseases, and specifically in this paper, the downstream task is formulated as a multi-label classification problem targeting various diagnostic statements, such as normal or abnormal QRS complex and arrhythmia.

TABLE I: Size of the pre-training and fine-tuning datasets.

# ECG recordings in	Pre-training	Fine-tuning
Training set	34763	17441
Validation set	14899	2193
Test set	-	2203

Furthermore, in this paper, we use both contrastive and non-contrastive learning methods. Contrastive self-supervised learning methods are based on the idea of instance discrimination. Instead of predicting the exact class of a data sample, the objective is to predict whether pairs of inputs belong to the same class or not. Specifically, contrastive learning has recently become a dominant component in computer vision with self-supervised learning methods such as SimCLR, MoCo, and CPC [6], [11], [16] being developed. Non-contrastive learning methods, unlike contrastive methods, learn non-trivial representations using only positive sample pairs. Non-contrastive approaches include methods such as BYOL and VICReg [14], [17]. Instead of explicitly defining negative samples, they introduce asymmetry in the network architecture to solve the representation problem. BYOL uses two neural networks to learn, the online and the target network, while VICReg introduces instance contrasting in the loss function.

This paper examines the importance and effects of different augmentation strategies using SimCLR, BYOL, and VICReg, which we present with an architectural overview in Figure 2. SimCLR and BYOL are investigated due to their demonstrated potential in the work conducted by [5] and [8], while the exploration of VICReg is motivated by its novelty, simplicity, and theoretical transparency.

III. Datasets

All ECG recordings used in this paper have been obtained from publicly accessible databases. All datasets consist of short-duration (7-10 seconds) standard 12-lead ECG recordings. The datasets also contain some metadata, but in this paper, only the ECG has been used. A collection of three datasets were used for the pre-training process: CinC2020 [18], Ribeiro [19], and Zheng [20]. Fine-tuning was carried out exclusively on the PTB-XL dataset [21]. Table I shows the number of recordings used for each section. All ECG recordings and associated annotations are obtained from in-clinic exams conducted by clinical cardiologists. As each ECG recording is annotated with a subset of the 71 labels in the PTB-XL dataset [21], these labels form the multi-label classification downstream task. The labels cover a wide variety of diagnostic, form, and rhythm statements that can be used for a comprehensive evaluation of ECG analysis algorithms.

IV. Augmentation Strategies

In self-supervised learning, the choice of views controls the information captured in the representation. Self-supervised techniques encourage representations to discard information regarding the augmentations applied to the input data, thereby becoming invariant to the set of chosen augmentations. The pre-training phase serves to decouple the correlations of irrelevant features between the representations of positive samples. Recent research, though not extensively studied, has shown that the ideal transformations to use are indeed dependent on the downstream task [22]. Throughout the experiments in this paper, the data transformation module is constructed to sequentially apply random resized crop augmentation followed by time out augmentation, following the experiments carried out by [5].

A. Random Resized Crop

Random resized crop cuts a random contiguous segment of the signal and rescales it to its original size. A crop parameter p is sampled uniformly from the range (l, m). In our experiments, we define *original augmentation strength* as $(l, m) = (0.5, 1.0)$, which corresponds to the values suggested by [5]. Additionally, we extend the experimental space by further investigating the effect of various augmentation strengths. In order to do this, we add another augmentation setting which we refer to as *stronger augmentation* and is defined as $(l, m) = (0.3, 1.0)$. This means that we crop a portion p (p being between 50% and 100%, or 30% and 100%, respectively) from the signal and upsample to the original size.

B. Time Out

The temporal specific transformation, time out augmentation [23], sets a random contiguous segment of the signal to zero. The range of the cutout window is determined by the parameters (t_l, t_u), from which the time out parameter t is uniformly sampled. The parameter describes how much of the original signal will be set to zero. Throughout our experiments, the augmentation setting referred to as *original augmentation strength* sets the parameters $(t_l, t_u) = (0.0, 0.5)$ and follows the implementation described by [5]. Thereby a stochastically chosen window with a maximum length of 50% of the original signal is set to zero. As with the random resized crop, we aim to increase our understanding of the augmentation strength's effect on downstream performance. Thus, we define a *stronger augmentation*, to allow the time out parameter t to take on values in the range $(t_l, t_u) = (0.2, 0.6)$.

V. Single Augmentation in Pairs

The first phase of the learning pipeline consists of a stochastic data transformation module, as described in Section IV. This module randomly transforms any given data sample to produce two correlated views of this same data instance. Current state-of-the-art approaches define a positive pair as two augmented copies of a data sample [11], [13], [14], [16], [17], making this the current approach for defining positive pairs. Through our work, we refer to our implementation of this approach as *double augmentation*. With the aim of making meaningful contributions to the field of research, we investigate how different formulations of positive pairs affect the downstream performance of the learned representations. To this end, we introduce an alternative approach that we refer to as *single augmentation*. Here, positive pairs are defined such that one augmented copy is paired with the original signal, giving a positive pair with only one transformed view, retaining more original signal information.

VI. Length of ECG Signals

For a healthy heart with a typical heart rate of 70 to 75 beats per minute, each cardiac cycle, or heartbeat, lasts for more or less 0.8 seconds [24]. Therefore, 2.5 seconds of an ECG signal as used in [5] entails about three complete cardiac cycles. This may not contain sufficient information for a classifier, given that some cardiovascular diseases, such as arrhythmia, can only be detected sporadically. To investigate the impact of ECG signal length on model performance, we extend the experiments by also introducing 10-second long ECG signals.

VII. Experimental Setup

All our adopted self-supervised learning frameworks follow the same procedure, starting with the pre-processing of data. During the self-supervised pre-text task, data augmentation is applied to each input sample to form multiple views of the data. The data views are then passed through an encoder network and mapped to a latent representation space. Following this step, the representation will either be projected into a lower dimensional space or expanded into a higher dimensional space, depending on the method currently in use. During the last step of the pre-text task, a loss function is minimized in this final representation space. After the pre-text task, the encoder network is fine-tuned using labeled data without augmentations and evaluated on a multi-label classification task. This is the downstream task of our procedure.

A. Data Preparation and Augmentation

The self-supervised learning pre-text task begins with the data augmentation module. An ECG signal $\mathbf{x}$ is uniformly sampled from the dataset and when performing a double augmentation, two data transformations τ and τ' are sampled from the transformation distribution $\mathcal{T}$. Each data transformation $\tau \sim \mathcal{T}, \tau' \sim \mathcal{T}$ is applied to the ECG signal $\mathbf{x}$, producing two different signal views, $\tilde{\mathbf{x}} = \tau(\mathbf{x}), \tilde{\mathbf{x}}' = \tau'(\mathbf{x})$ that form a positive pair with double augmentation. For some of the experiments, positive pairs are defined using the single augmentation strategy where one augmented copy $\tilde{\mathbf{x}} = \tau(\mathbf{x})$ is paired with the original signal $\mathbf{x}$, giving a positive pair as $(\tilde{\mathbf{x}}, \mathbf{x})$. The transformations are stochastic resizing crops of the signal, followed by time out augmentation, as described in Section IV. Following the work of [5] and [25], the ECG recordings used throughout these experiments are restricted to ECG data at a sampling rate of 100 Hz. Signals are segmented into windows of length T, where $T = 250$ or $T = 1000$ depending on whether the used signal length is 2.5 seconds or 10 seconds.

B. Self-Supervised Training

Following current state-of-the-art methods [5] the ResNet-50 architecture [26] is chosen as the encoder network for all three self-supervised learning approaches implemented in this work. For the Siamese network architectures used in SimCLR [16] and VICReg [14], the encoders on both branches share the same set of weights, while BYOL [17] updates the encoder weights of the target branch according to a moving average of the online branch. The architectural implementations of the three methods follow those described in their original papers but are adjusted to one-dimensional inputs. Apart from the number of training epochs, which we set to 2000, and, batch size which is set to either 2048 or 512 depending on ECG signal length, our training regime and network hyperparameters also follow the methods' original implementations.

C. Multi-Label Classification Task

For the downstream task, we use a 1-dimensional ResNet-50 model extended with a fully connected classification layer for the 71 labels. In the case of the pre-trained models, this is equivalent to adding a classification layer on top of the ResNet-50 encoder. After fine-tuning, we compare their downstream performance with that of a model that is randomly initialized but architecturally identical and trained using labeled data. For both fine-tuning and supervised training, we employ a standard binary cross-entropy loss. During fine-tuning of the pre-trained models, and supervised training for the baseline model, a constant learning rate of 0.008 is used to optimize a binary cross-entropy loss. Furthermore, we use an AdamW optimizer in combination with a weight decay regularization of 0.001 [27]. When training the network using ECG recordings of 10 seconds in length, the batch size is set to 512, whereas a batch size of 2048 is used for networks trained with ECG recordings of length 2.5 seconds. Addressing class imbalance, model performance is measured using macro-AUC, as described in [25], computed from the 71 labels on the most fine-grained level in PTB-XL [21].

The model selected for evaluation is the one that during training obtained the highest macro-AUC score when evaluated on the validation data. Reported metrics are the respective test set score of this selected model. Moreover, five runs of fine-tuning were performed on the same set of data, each with stochastically sampled data augmentations.

VIII. Results

The macro-AUC scores for the three different models (SimCLR, BYOL, and VICReg) using different pre-processing and augmentation strategies are presented in Table II. The results are evaluated after fine-tuning using 100% of the labeled data. The table presents results for all combinations of signal length, double/single augmentations, and augmentation strength previously presented, where stronger augmentation refers to stronger both in the sense of random resized crop and time out. The presented values are averaged over the five different fine-tuned models and the standard deviation of these are presented in parenthesis. The asterisk refers to results that follow the same strategy as [5]. The best results are achieved on pre-training on BYOL with 10-second long signals, and single, stronger augmentation.

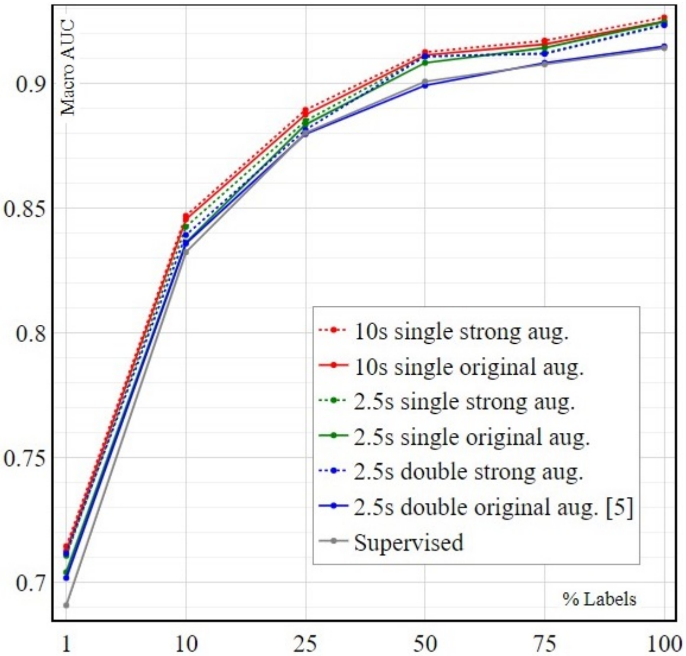

Fig. 3: Performance as a function of % labeled fine-tuning data for models pre-trained using the BYOL method. Note the non-linearity of the x-axis.

As BYOL showed the most promising results we also present the macro-AUC scores as a function of how much of the fine-tuning data is used, for the different augmentation strategies, in Figure 3. The 2.5s double original aug. refers to the previously published augmentation strategies in [5] (same as asterix in Table II). The plot shows that all our augmentations outperform both this and the supervised network at all levels of labeled data, while [5] only performs better than the supervised network for 1% and 10%. Again, we see that both stronger augmentation, single augmentation, and increased signal length improve the performance, with 10s single strong aug. performing best. The 10s double original and strong aug. have been left out of the plot as they performed similarly to their corresponding versions with single augmentation. However, it is still important to note that it improves the performance to use single augmentation instead of double for the 2.5-second signals. This shows that the performance still can be boosted even if only shorter signals are available.

IX. Conclusions

In this study, we present an assessment of augmentation strategies for self-supervised representation learning on 12-lead clinical ECG data. Although self-supervised algorithms have been applied successfully in other data domains, the ECG signal is of a different data modality on which the applications of self-supervised learning have not yet been extensively examined. Published experiments mostly follow frameworks presented in the computer vision domain, though direct policy adoption could lead to weaker generalization on downstream tasks. With this in mind, we implement

TABLE II: Macro-AUC scores for SimCLR, BYOL, and VICReg using different augmentation strategies. Mean and std. macro-AUC are computed on the test data over five evaluation runs, fine-tuned for 100 epochs on 100% of the labeled data. A supervised network (no pretraining) obtained 0.9157(.0037) with 2.5-second signals and 0.9165(.0032) with 10-second signals.

	2.5 sec				10 sec			
	Double		Single		Double		Single	
Method	Orig.	Strong	Orig.	Strong	Orig.	Strong	Orig.	Strong
SimCLR	.9150(.0035)*	.9184(.0053)	.9175(.0047)	.9198(.0055)	.9224(.0030)	.9254(.0032)	.9246(.0021)	.9258(.0081)
BYOL	.9139(.0047)*	.9257(.0022)	.9162(.0019)	.9270(.0015)	.9263(.0037)	.9269(.0036)	.9270(.0050)	**.9304(.0063)**
VICReg	.9153(.0033)*	.9152(.0097)	.9195(.0022)	.9179(.0067)	.9154(.0035)	.9186(.0037)	.9168(.0023)	.9263(.0025)

and analyze the effect of different augmentation strategies on three major self-supervised learning methods: SimCLR, BYOL, and VICReg. Among the insights obtained, the most crucial led to further insights into the importance of data pre-processing and augmentation for improved performance results on the downstream task. By increasing the length of the ECG signal, downstream performance results are improved for most methods. Also, augmenting only a single copy in the positive pairs has a positive effect. Combining this single augmentation and increased signal length with a stronger data augmentation strategy, the self-supervised pre-trained models performed better than previous methods using shorter signals and double, less strong, augmentation. Our models also outperformed their supervised counterparts in all evaluation settings. This highlights that defining a suitable augmentation protocol is crucial for improved performance results on downstream tasks related to cardiovascular health. This study's findings add valuable insights into the importance of formulating an optimal strategy for self-supervised ECG representation learning.

References

[1] Z. Li, D. Zhou, L. Wan, J. Li, and W. Mou, "Heartbeat classification using deep residual convolutional neural network from 2-lead electrocardiogram," *Journal of Electrocardiology*, vol. 58, pp. 105–112, 2020.

[2] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Scientific reports*, vol. 10, no. 1, pp. 1–17, 2020.

[3] J. He, "Automated heart arrhythmia detection from electrocardiographic data," Ph.D. dissertation, Victoria University, 2020.

[4] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nature Reviews Cardiology*, vol. 18, no. 7, pp. 465–478, 2021.

[5] T. Mehari and N. Strodthoff, "Self-supervised representation learning from 12-lead ecg data," *Computers in Biology and Medicine*, vol. 141, p. 105114, 2022.

[6] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[7] D. Kiyasseh, T. Zhu, and D. A. Clifton, "Clocs: Contrastive learning of cardiac signals across space, time, and patients," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5606–5615.

[8] H. Liu, Z. Zhao, and Q. She, "Self-supervised ecg pre-training," *Biomedical Signal Processing and Control*, vol. 70, p. 103010, 2021.

[9] Y. LeCun and I. Misra, "Self-supervised learning: The dark matter of intelligence," Mar 2021. [Online]. Available: https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[12] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[13] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.

[14] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[15] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1476–1485.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[18] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi *et al.*, "Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020," *Physiological measurement*, vol. 41, no. 12, p. 124003, 2020.

[19] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr *et al.*, "Automatic diagnosis of the 12-lead ecg using a deep neural network," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[20] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Scientific Data*, vol. 7, no. 1, pp. 1–8, 2020.

[21] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific data*, vol. 7, no. 1, pp. 1–15, 2020.

[22] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 6827–6839, 2020.

[23] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[24] B. J. Gersh, *Mayo Clinic Heart Book: The ultimate guide to heart health*. W. Morrow, 2000, p. 6–8.

[25] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ecg analysis: Benchmarks and insights from ptb-xl," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2020.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.