Masked Autoencoder-Based Self-Supervised Learning for Electrocardiograms to Detect Left Ventricular Systolic Dysfunction.

Shinnosuke Sawano
sawanos-int@h.u-tokyo.ac.jp
Satoshi Kodera
koderas-int@h.u-tokyo.ac.jp
Hirotoshi Takeuchi
takeuchi-hirotoshi742@g.ecc.u-tokyo.ac.jp
Issei Sukeda
sukeda-issei006@g.ecc.u-tokyo.ac.jp
Susumu Katsushika
katsushikas-int@h.u-tokyo.ac.jp
Issei Komuro
komuro-3im@h.u-tokyo.ac.jp

Department of Cardiovascular Medicine, The University of Tokyo Hospital

Abstract

The generalization of deep neural network algorithms to a broader population is an important challenge in the medical field. In this study, we aimed to apply self-supervised learning using masked autoencoders (MAEs) to improve the performance of deep learning models that detect left ventricular systolic dysfunction (LVSD) from 12-lead electrocardiography data. In our MAE approach, we first mask the vast majority, that is, 75% of the ECG time series. Second, we pretrain a Vision Transformer encoder by inferring the masked part. Our proposed approach enables rich features that generalize well from unlabeled ECG data to be learned. In fact, the reconstructed ECG maintains the relationships among the major ECG components. Transfer performance in the detection of LVSD outperforms the baseline CNN model on external validation datasets and shows promising results for generalization that enables us to use the model for a broader population by solely using ECG data collected in a single medical institution.

1 Introduction

The generalization of deep neural network (DNN) algorithms to a broader population is an important challenge, and this problem can be a major barrier to the social implementation of DNN algorithms in the medical field (1-3). Low generalization performance can occur when the dataset used to train the algorithm is not diverse enough. Improving generalizability requires large amounts of accurately labeled data, however, obtaining such data is often difficult due to labeling costs requiring specialized knowledge and ethical considerations in the medical field. To address this problem, self-supervised learning, which can make more effective use of limited data, is attracting attention in computer-vision tasks. Self-supervised learning using masked autoencoders (MAEs; licensed under an Attribution Non-Commercial 4.0 International license) learns very high-capacity models that generalize well (4). Pretraining with an MAE enables data-hungry models such as Vision Transformers (ViT)-

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Large/Huge (5) to be trained on ImageNet-21K to improve generalization performance. Contrast-learning approaches such as MoCo (6, 7) and SimCLR (8) have also achieved high performance in computer-vision tasks but require image transformation to create a contrast to the original image. In contrast, MAE learns a reusable representation of the input image by masking random patches and reconstructing missing pixels. Therefore, the transformations used in contrast learning are unnecessary. In this study, we aimed to apply self-supervised learning using MAE to the analysis of 12-lead electrocardiograms (ECGs), which is one of the most popular medical time-series data. The deep learning approach to analyzing ECGs has been widely studied in recent years, enabling highly accurate detection of cardiac diseases that were previously difficult to diagnose from ECGs (9-11). However, these approaches require large amounts of labeled electrocardiography data, which are difficult to collect at a single institution. To solve this problem, several methods have been proposed to apply self-supervised pretraining to ECGs (12, 13). However, no previous studies have applied MAE-based pretraining to train a high-capability ViT-Large exclusively for ECG analysis. We focus on detecting left ventricular systolic dysfunction (LVSD) from the ECG. LVSD is a common disease that may increase the risk of sudden death (14, 15). Therefore, early detection of LVSD is desirable. In clinical practice, LVSD is currently diagnosed by echocardiography (16), and diagnosis from ECGs has been difficult. However, it has been reported that LVSD can be diagnosed from an ECG by applying deep learning (9), and the implementation of these deep-learning algorithms in clinical practice is expected.

2 Method

2.1 Datasets

To build and evaluate the models, we collected electrocardiographic data from eight academic medical institutions. We created two datasets for training. The first is a small dataset (Dataset1) comprising 37,456 ECGs collected at institution 1, and the second is a large dataset (Dataset2) comprising 126,203 ECGs collected from institutions 1, 2, and 3 (Tables 1 and 2). LVSD was assessed by echocardiography and defined as an ejection fraction of less than 40%. Patients who had LVSD were labeled as positive, and the rest were labeled as negative. Tables 1 and 2 show the details of the datasets. Dataset1 was divided into Train1 (containing train1 and valid1) and Test1, and Dataset2 was divided into Train2 (containing train2 and valid2) and Test2. Note that to avoid data leakage, all ECGs from one patient were assigned to the same split. All ECG data from Dataset1 and Dataset2 were used separately for the self-supervised pretraining, in which the labels were not used. The Train1 or Train2 dataset was then used for the downstream task, in which a ViT-Large model was trained to detect LVSD by supervised learning. In addition, ECGs from the five external institutions that were not used for training were prepared as an external validation dataset, as detailed in Appendix A, to test the performance of the model. The prevalence of LVSD was also included in Appendix A. This study was conducted with the approval of the University of Tokyo's Institutional Review Board and in a manner such that the data could not be used to identify individuals.

2.2 Self-Supervised Pretraining

The first step of our approach was to pretrain an efficient ViT-Large encoder by self-supervised learning using Dataset1 and Dataset2. In this study, we used the MAE algorithm. In an MAE for image data (e.g., 224×224), the input is divided into 16×16 patches. These patches are randomly masked and the missing pixels are then reconstructed. This process creates an encoder that has learned a useful representation (4). Raw data from each ECG examination record were represented as a 12×5000 matrix of ECG voltage, in which the first dimension was a spatial dimension (each column represented one lead) and the second dimension was the temporal dimension (each row represented a specific time point). To take advantage of the interrelationship of the 12 ECG leads, the patch size was changed to 1×250. As a result, the ECG information per patch was 0.5 s. For self-supervised learning, we trained the ViT-Large encoder for 1600 epochs with a batch size of 1024. Implementation details followed those of the previous study (4).

2.3 Model Architecture

We use a ViT-Large architecture to take advantage of its strong generalization capabilities. Although ViT-Huge could have been used, ViT-Large was selected for usability considerations. The encoder

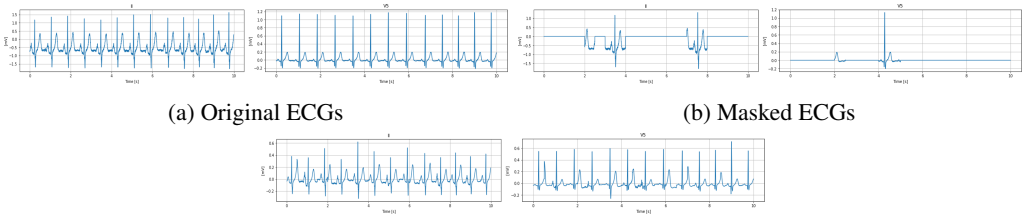
Table 1: Details of Dataset 1. Small dataset from a single institution.

	Split		Case	Control	Unique patients	Over all
Dataset1	Train1	train1	2,160 (8%)	23,992 (92%)	16,187	26,152
		valid1	461 (8%)	5,203 (92%)	3,478	5,664
	Test1		493 (8%)	5,147 (92%)	3,489	5,640
Over all			3,114	34,342	23,154	37,456

Table 2: Details of Dataset 2. Large dataset from three institutions.

	Split		Case	Control	Unique patients	Over all
Dataset2	Train2	train2	7,881 (9%)	80,599 (91%)	4,8598	88,480
		valid2	1,601 (8%)	17,490 (92%)	10,432	19,091
	Test2		1,722 (9%)	16,910 (91%)	10,275	18,632
Over all			11,204	114,999	69,305	126,203

takes a 12×5000 ECG matrix as input, and finally outputs a 1024-dimensional feature vector. In the self-supervised pretraining, ECGs after normalization were input to the encoder using MAE. In the downstream task, a normalization layer and a fully connected layer were added to the encoder to detect LVSD. As a baseline, a CNN architecture was used because a previous study showed that this architecture can achieve high performance in detecting LVSD (9). The encoder consists of six temporal convolution blocks, one spatial convolution block, and one fully connected layer. The encoder takes a 12×5000 ECG matrix as input, and finally outputs a 128-dimensional feature vector. In the downstream task, two fully connected layers and a sigmoid layer were added to the encoder to detect LVSD. The binary cross-entropy loss was minimized. To obtain a qualitative sense of our reconstruction task, see Figure 1 and Appendix C.



(a) Original ECGs (b) Masked ECGs

(c) Reconstructed ECGs

Figure 1: Example of the reconstruction process in II and V5 lead

2.4 Evaluation experiments

We conducted two experiments: one with our proposed approach and one with the baseline (without pretraining) using four sets of Nvidia Tesla A100 80 GB graphics processing units (Nvidia Corporation, Santa Clara, USA). First, we conducted self-supervised pretraining on the ViT-Large encoders for each pretraining dataset: Dataset1 and Dataset2. Then, we fine-tuned the model on each downstream dataset (Train1 or Train2) using the pretrained ViT-Large encoders. Second, the baseline was trained, omitting the self-supervised pretraining step. We directly trained the models on the downstream datasets using supervised learning. The CNN model was trained with randomly initialized model weights, whereas the ViT-Large model was trained with the model weights pretrained by MAE using ImageNet-21K. Table 4 in Appendix A shows the correspondence of the datasets used in each experiment. In all experiments, the Adam optimizer was used. Initial learning rates of 1e-4, 3e-5, and 1e-5 were tested in each experiment; the model weights with the lowest validation loss were saved and later used for evaluation on the test set. Models fine-tuned on Train1 (consisting of data from institution 1) were evaluated on Test1 and the data from the remaining seven institutions, whereas the

models fine-tuned on Train2 (consisting of data from institutions 1, 2, and 3) were evaluated on Test2 and all data from the remaining five institutions. Appendix B presents an overview of our experiment.

3 Results and Discussion

We analyze the fine-tuning results of the self-supervised ViT-Large. For the models fine-tuned on Train1, the model pretrained using Dataset1 performed better, with area under the receiver operating characteristic curve (AUROC) results of 0.96, 0.94, 0.92, 0.92, 0.95, 0.96, and 0.94 compared to the baseline CNN model, which obtained AUROC results of 0.95, 0.88, 0.89, 0.87, 0.92, 0.94, and 0.90 for the Test1 and institution 2–8 cohorts. Moreover, the model pretrained using Dataset2 obtained the highest AUROC results of 0.97, 0.95, 0.93, 0.92, 0.95, 0.97, and 0.94 (Table 3). The proposed method performs better than the baseline CNN, indicating that the ViT encoder can learn feature representations that are effective in detecting LVSD through self-supervised pretraining. The results also show that higher model performance can be obtained by performing self-supervised learning with MAE using a large dataset. Regarding the models fine-tuned with Train2, the model pretrained using Dataset1 obtained AUROC results of 0.95, 0.92, 0.95, 0.97, and 0.94, which are higher than those of the baseline CNN, which obtained AUROC results of 0.93, 0.90, 0.93, 0.95, and 0.92 for the Test2 and institution 4–8 cohorts. The model pretrained using Dataset2 performed the best, with AUROC result of 0.96, 0.94, 0.96, 0.98, and 0.95. Furthermore, the baseline ViT-Large pretrained by MAE using ImageNet-21K had the lowest AUROC results of 0.91, 0.88, 0.91, 0.94, 0.90 for the Test 2, institutions 4–8 cohorts. The baseline ViT-Large performed worse than the baseline CNN model (table 3). These data show that the performance of the pretrained ViT-Large can be further improved using a larger training dataset. By contrast, the low accuracy of the ViT-Large pretrained on ImageNet-21K indicates that, although fine-tuning was performed using similarly large training data, the training dataset used in this study was not sufficient to train the ViT-Large. Therefore, we believe that our proposed method helped train ViT-Large without using additional labeled data and instead making effective use of limited data. Furthermore, as shown in the results in Table 3, ViT-Large using the model pretrained with self-supervised learning yields higher AUROC scores with external validation data than do the baseline models, indicating that generalizability was improved. We highlight a comparison of the performance of ViT-Large pretrained using Dataset1 and fine-tuned on Train1 (bolded text in Table 3) and the baseline CNN model fine-tuned on Train2 (underlined text in Table 3). ViT-Large, which was pretrained and fine-tuned using a small training dataset from a single institution, obtained higher AUROC results on the external validation dataset and better generalization performance than the baseline CNN model fine-tuned on a large training dataset collected from three institutions. The total number of labels used for our proposed self-supervised approach is only 31,816 labels in Train1. In contrast, the baseline CNN model used 107,571 labels in Train2. In summary, the proposed method achieves higher AUROC results than the baseline CNN with only 30% labels and improves generalization performance.

Table 3: AUROC of each model evaluated on test datasets. Our proposed methods were used to pretrain a ViT on Dataset1 and Dataset2. A CNN and ViT pretrained on ImageNet are the baseline models.

Model	CNN	ViT	ViT	CNN	ViT	ViT	ViT
Pre-training	—	Dataset1	Dataset2	—	ImageNet	Dataset1	Dataset2
Fine-tuning	Train1	Train1	Train1	Train2	Train2	Train2	Train2
Test1	0.95	0.96	0.97	—	—	—	—
Test2	—	—	—	0.93	0.91	0.95	0.96
Institution2	0.88	0.94	0.95	—	—	—	—
Institution3	0.89	0.94	0.95	—	—	—	—
Institution4	0.87	**0.92**	0.93	<u>0.90</u>	0.88	0.92	0.94
Institution5	0.92	**0.95**	0.95	<u>0.93</u>	0.91	0.95	0.96
Institution6	0.94	**0.96**	0.97	<u>0.95</u>	0.94	0.97	0.98
Institution7	0.90	**0.94**	0.94	<u>0.92</u>	0.90	0.94	0.95
Institution8	0.89	**0.93**	0.94	<u>0.92</u>	0.89	0.94	0.95

4 Broader Impact

Although our approach allowed us to develop a deep learning model with high generalization performance that efficiently uses few labels and detects LVSD, more detailed studies are needed to determine whether our method can be widely used for other ECG-based tasks. Based on this study, we aim to publish a powerful ECG pretraining model that is capable of multitasking. Our goal is to contribute to research for the social implementation of deep learning models to analyze ECGs soon. To the best of our knowledge, our work has no potential negative impacts.

References

[1] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15(11):e1002683.

[2] Goto S, Solanki D, John JE, Yagi R, Homilius M, Ichihara G, et al. Multinational Federated Learning Approach to Train ECG and Echocardiogram Models for Hypertrophic Cardiomyopathy Detection. Circulation. 2022:101161circulationaha121058696.

[3] Abrams C. Google's effort to prevent blindness shows AI challenges. Wall Street J. January 26, 2019. 2019.

[4] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R, editors. Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022.

[5] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.

[6] He K, Fan H, Wu Y, Xie S, Girshick R, editors. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020.

[7] Chen X, Fan H, Girshick R, He K. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:200304297. 2020.

[8] Chen T, Kornblith S, Norouzi M, Hinton G, editors. A simple framework for contrastive learning of visual representations. International conference on machine learning; 2020: PMLR.

[9] Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019;25(1):70-4.

[10] Sawano S, Kodera S, Katsushika S, Nakamoto M, Ninomiya K, Shinohara H, et al. Deep learning model to detect significant aortic regurgitation using electrocardiography. J Cardiol. 2022;79(3):334-41.

[11] Attia ZI, Harmon DM, Behr ER, Friedman PA. Application of artificial intelligence to the electrocardiogram. Eur Heart J. 2021;42(46):4717-30.

[12] Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ECG data. Computers in Biology and Medicine. 2022;141:105114.

[13] Sarkar P, Etemad A. Self-supervised ECG representation learning for emotion recognition. IEEE Transactions on Affective Computing. 2020.

[14] Wang TJ, Levy D, Benjamin EJ, Vasan RS. The epidemiology of "asymptomatic" left ventricular systolic dysfunction: implications for screening. Ann Intern Med. 2003;138(11):907-16.

[15] Wang TJ, Evans JC, Benjamin EJ, Levy D, LeRoy EC, Vasan RS. Natural history of asymptomatic left ventricular systolic dysfunction in the community. Circulation. 2003;108(8):977-82.

[16] Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr. 2015;28(1):1-39.e14.

Appendix

A Dataset Details

Table 4: Data splits used for each institution. (%) indicates the prevalence at each institution.

	Institution1		Institution2		Institution3	
	Case	Control	Case	Control	Case	Control
Train	2,160(8%)	23,992(92%)	3,003(9%)	30,270(91%)	2,718(9%)	26,337(91%)
Valid	461(8%)	5,203(92%)	526(7%)	6,616(93%)	614(10%)	5,671(90%)
Test	493(8%)	5,147(92%)	658(9%)	6,280(91%)	571(9%)	5,483(91%)

	Institution4		Institution5		Institution6	
	Case	Control	Case	Control	Case	Control
Train	—	—	—	—	—	—
Valid	—	—	—	—	—	—
Test	18,952(87%)	2,724(13%)	3,925(95%)	199(5%)	12,759(96%)	595(4%)

	Institution7		Institution8	
	Case	Control	Case	Control
Train	—	—	—	
Valid	—	—	—	—
Test	24,208(89%)	2,947(11%)	31,660(91%)	3,181(9%)

B Overview of Our Experiment

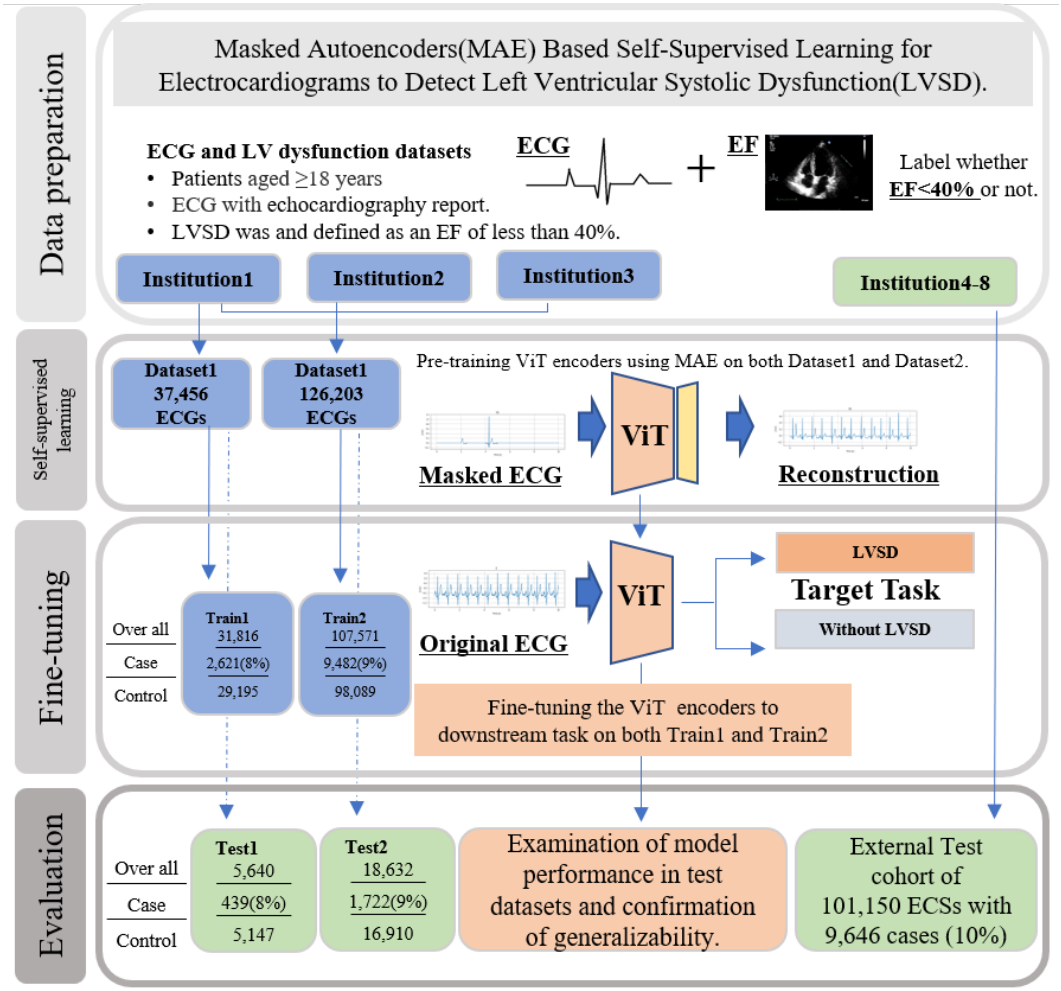
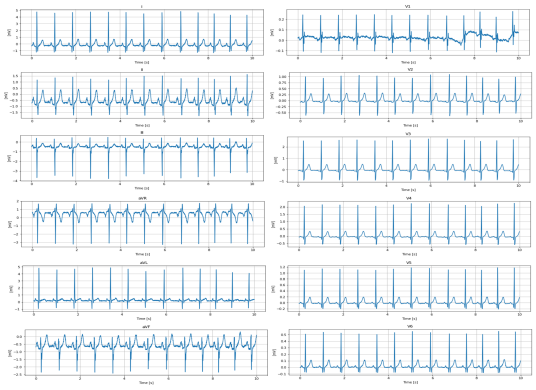


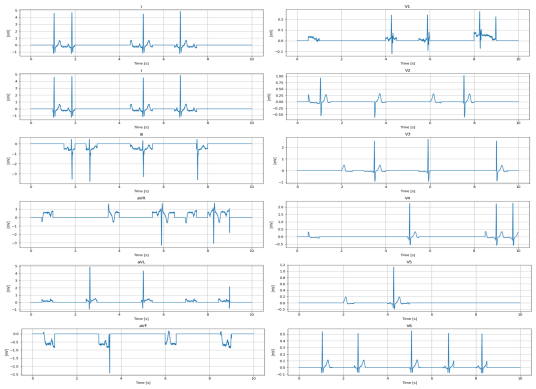
Figure 2: We collected 12-lead electrocardiographic data from eight academic medical institutions and labeled left ventricular systolic dysfunction(LVSD). LVSD was assessed by echocardiography and defined as an ejection fraction of less than 40%. We created two datasets for training. The first is a small dataset (Dataset1) collected at institution 1, and the second is a large dataset (Dataset2) comprising collected from institutions 1, 2, and 3. Dataset1 was divided into Train1 and Test1, and Dataset2 was divided into Train2 and Test2. All ECG data from Dataset1 and Dataset2 were used separately for the self-supervised pretraining, in which the labels were not used. The Train1 or Train2 dataset was then used for the downstream task, in which a ViT-Large model was trained to detect LVSD from 12-lead electrocardiographic data by supervised learning. In addition, ECGs from the external institutions that were not used for training were evaluated as an external validation dataset.

C Qualitatively Understanding of the ECG Features Learned in the Self-Supervised Pretraining

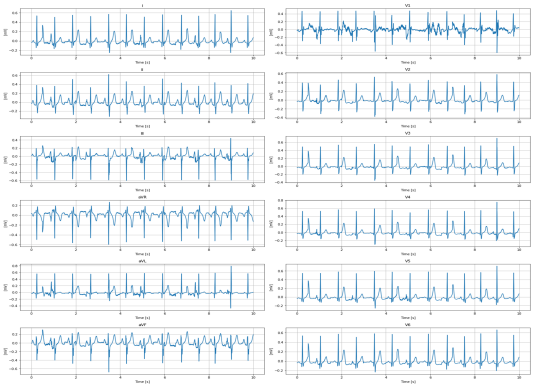
We qualitatively assessed the validity of the ECGs reconstructed by the physician. As shown in Figure 2, the reconstructions maintained the relationships among the major ECG components such as the P wave, QRS wave, and T wave. In addition, there was no significant discrepancy in the information between the II, III, and aVF leads (with information about the lower wall of the heart) and the V1, V2, V3, V4, and V5 leads (with information about the anterior wall), which indicate the position information of the ECG.



(a) Original ECGs



(b) Masked ECGs



(c) Reconstructed ECGs

Figure 3: Example of the reconstruction process of a 12-lead ECG