# MaeFE: Masked Autoencoders Family of Electrocardiogram for Self-Supervised Pretraining and Transfer Learning

Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, *Senior Member, IEEE*, Hao Wang, *Member, IEEE*, Jin He, *Senior Member, IEEE*, and Qijun Huang

***Abstract*— Electrocardiogram (ECG) is a universal diagnostic tool for heart disease, which can provide data for deep learning. The scarcity of labeled data is a major challenge for medical artificial intelligence diagnosis. Acquiring labeled medical data is time-consuming and high-cost because medical specialists are needed. As a kind of generative self-supervised learning method, a masked autoencoder (MAE) is capable to solve these problems. MAE family of ECG (MaeFE) is proposed in this article. Considering the temporal and spatial features of ECG, MaeFE contains three customized masking modes, including masked time autoencoder (MTAE), masked lead autoencoder (MLAE), and masked lead and time autoencoder (MLTAE). MTAE and MLAE pay greater attention to temporal features and spatial features, respectively. MLTAE is a multihead architecture that combines MTAE and MLAE. In the pretraining stage, ECG signals from the pretrain dataset are divided into patches and partially masked. The encoder transfers unmasked patches to tokens and the decoder reconstructs masked ones. In downstream tasks, the pretrained encoder is utilized as a classifier, which is arrhythmia classification performed in the downstream dataset. The process is the so-called transfer learning. MaeFE outperforms the state-of-the-art self-supervised learning methods, SimCLR, MoCo, CLOCS, and MaskUNet in downstream tasks. MTAE has the best comprehensive performance. Compared to contrastive learning models, MTAE achieves at least a 5.18%, 11.80%, and 3.23% increase in accuracy (Acc), Macro-$F_1$, and area under the curve (AUC), respectively, using the linear probe. It also outperforms other models at 8.99% in Acc, 20.18% in Macro-$F_1$, and 7.13% in AUC using fine-tuning. As another downstream task, experiments on the multilabel classification of arrhythmia are also conducted, which reflects the excellent generalization performance of MaeFE. Depending on experimental results, MaeFE turns out to be efficient and robust in downstream tasks. Overcoming the scarcity of labeled data, MaeFE is better than other self-supervised learning methods and achieves satisfying performance. Consequently, the algorithm in this article is on track of playing a major role in practical applications.**

***Index Terms*— Electrocardiography (ECG), mask autoencoder (MAE), pretraining, self-supervised learning, transfer learning.**

Manuscript received 4 September 2022; revised 1 November 2022; accepted 30 November 2022. Date of publication 12 December 2022; date of current version 13 January 2023. This work was supported by the National Natural Science Foundation of China under Grant 81971702, Grant 62074116, and Grant 61874079. The Associate Editor coordinating the review process was Dr. Mohamad Forouzanfar. *(Corresponding author: Qijun Huang.)*

The authors are with the School of Physics and Technology, Wuhan University, Wuhan, Hubei 430072, China (e-mail: huangqj@whu.edu.cn).

Digital Object Identifier 10.1109/TIM.2022.3228267

## I. Introduction

COMPUTERIZED electrocardiogram (ECG) is a noninvasive method to record the electrical activity of the heart [1]. Due to its convenience and practicality, ECG is commonly used in the diagnosis of cardiovascular diseases (CVD). The ECG is pivotal for diagnosing a wide spectrum of abnormalities, from arrhythmias to acute coronary syndrome [2]. With the rapid development of machine learning, CVD can be diagnosed automatically [3], [4]. However, supervised learning methods require an abundant amount of data labeled by cardiologists. Since labeling ECG data accurately is time-consuming, labeled ECG data are scarce. In addition, cardiologists who label the data have different personal experiences. Thus, it is impossible to label the data perfectly within a limited time and effort. Furthermore, medical data do not necessarily fall into just one category. Diseases may co-occur, which makes labeling data more difficult. Yet, the problems would have less impact on unsupervised approaches, for the sake that unsupervised learning methods can learn characteristics from a vast amount of unlabeled ECG data by pretraining [5]. If enough features have been gained during the pretraining phase, models could perform well in downstream tasks under the condition of insufficient data. By unsupervised deep transfer learning (UDTL) [6], unsupervised approaches could even perform better than fully supervised methods in the situation of an extreme data shortage [7]. Unsupervised learning algorithms offer the possibility of using more data for training. Self-supervised learning belongs to the category of unsupervised learning.

Recently, self-supervised learning has achieved remarkable progress in the field of computer vision (CV) and natural language processing (NLP). Many models that perform well are proposed, including generative pre-training (GPT) [8], [9] and masked autoencoding in BERT [10].

In the area of self-supervised learning, the contrastive learning method and generative learning method are the main [11]. For the first one, these methods define positive views and negative views. Representations of positive views are pushed together during pretraining, while the representations of negative views are pulled apart [12]. Unlike contrastive learning, generative learning focuses on the ability to generate data. By reconstructing data, the model will pay attention to key information, which is critical in the downstream tasks. This is

because masked autoencoder (MAE) learns a rich hidden representation in the process of pretraining [13]. Recently, as a kind of generative model, MAE applied in CV is introduced [13], choosing vision transformer (ViT) as the backbone [14]. Having achieved great success in NLP, autoencoders using transformers are proven to be effective pretraining strategies in vision as well according to recent studies.

Based on the masked image modeling (MIM) framework, MAE is an effective and simple self-supervised representation learning strategy. The core idea of MAE is masking and reconstruction [15]. In the masking step, data are partially masked and put into the encoder, and then, in the reconstruction phase, decouple decoder reconstructs the data. To reconstruct the data, the encoder must acquire enough essential information. The ability to retrieve information determines whether the encoder will perform well in the downstream tasks. Through the encode–decode process, MAE pays greater attention to essential information. Different from pictures, ECG has a more flexible way of data composition. ECG data could be represented as multichannel 1-D data or single-channel sub2-D data. This provides the possibility to further explore the potential of MAE at the time scale and channel scale. Moreover, ECG has 12 synchronized leads, which describe the activity of the heart from different perspectives. This is the unique feature of ECG. Therefore, various masking strategies can be introduced to improve the autoencoder in accordance with diverse combinations of leads and time series. Thus, applying MAE to ECG is valuable for research.

In this article, novel applications of a family of MAEs to ECG are proposed and evaluated. Combining the features of ECG, MAE is optimized and integrated into a family. To obtain temporal and spatial features, ECG was analyzed in 2-D. Self-supervised learning and transfer learning are introduced to overcome the need for large labeled datasets. The contributions of this article are listed as follows.

1) A family of customized masked autoencoders for ECG is introduced, which is called MAE family of ECG (MaeFE). MaeFE divides ECG signals into patches and masks them at time level or lead level. As members of MaeFE, masked time autoencoder (MTAE), and masked lead autoencoder (MLAE) focus on time level and lead level, masked lead and time autoencoder (MLTAE) is a multihead architecture and the combination of MTAE and MLAE. The encoder transfers unmasked patches to tokens and the decoder reconstructs masked ones, which is the pretext task. MTAE, MLAE, and MLTAE are members of MaeFE pretrained with different masking strategies. The members focus on different characteristics of ECG. As a result, the models can be trained without a label. Meanwhile, the synchronization and interaction of the leads can be better considered in this process. As a result, temporal and spatial features are extracted.
2) Two different dimensions are introduced in the pretraining stage. MTAE is trained with 1-D data. In MTAE, ECG signals are split into patches on the time scale. The features of morphology and time will be extracted in this way. Different from MTAE, MLAE focuses on data in the sub2-D. The signal of each lead is regarded as one patch so that MLAE pays more attention to the relationship among different leads. Another model is a multihead model combining MTAE and MLAE, named MLTAE.
3) Transfer learning is introduced after MaeFE is pretrained. The encoder of MaeFE is used as the classifier in the downstream dataset. The downstream task is the classification of single-label arrhythmia and the classification of multilabel arrhythmia. In this stage, the pretrained model is tuned to fit the downstream dataset. In addition, a comprehensive assessment of transfer learning is presented after experiments are conducted. It demonstrates the utter pervasiveness of the algorithm.
4) Comparisons of MaeFE and other self-supervised learning methods are discussed. Depending on experimental results, MaeFE outperforms the state-of-the-art contrastive learning methods, including SimCLR, MoCo, CLOCS, and MaskUNet in downstream tasks. Different backbones are evaluated and discussed as well.
5) According to diverse masking modes, ablation experiments are conducted to find a better member of MaeFE. The masking ratio, the number of masked leads, and the way the models are combined and explored. To some extent, different modes determine which representations that matter, containing spatial and temporal characteristics of ECG. The sensitivity of MaeFE to the backbone is also discussed in this article, mainly including both ViT and ResNet backbone.

## II. Related Work

### A. Masked Autoencoder

Based on autoencoder [16], [17], MAE is a form of more general denoising autoencoder proposed recently [13], [18]. MAE is a simple but effective approach to training models without supervision. Learnable parameters are tuned during reconstructing partially masked data. As a generative learning method, MAE has the potential to achieve performance comparable to comparative learning methods that have dominated vision self-pretraining in recent years. MAE also achieves great transfer learning performance in downstream tasks, and it is superior to supervised pretraining methods [15].

### B. MAE-Based Algorithm

MAE structure is broadly applied in numerous fields. In [19], [20], and [21], MAE is introduced to medical image analysis and unsupervised anomaly detection (UAD). In this way, contextual information is aggregated and anomalies can be identified more accurately. MAE is also drawn into graph neural networks (GNNs) in [22], [23], and [24]. In these cases, features are learned by reconstructing masked nodes. In the field of point clouds, the point cloud is represented as discrete occupancy values, and features are learned by classifying masked object points and sampled noise points [25], [26]. Also, MAE is also utilized in the video, using cube embedding and tube masking [27]. Innumerable new research findings on MAE are being presented. It is bound to bring a boom to the research of MAE-based structure and model.

### C. Self-Supervised Learning Methods

Self-supervised learning is widely studied recently, and it mainly includes contrastive learning and generative learning. In a manner of self-supervised, models are explicitly trained with supervisory signals that are generated from the data itself by leveraging its structure [11].

Self-supervised learning is dominated by contrastive learning approaches for a long time, including SimCLR [28], contrastive predictive coding (CPC) [29], momentum contrast (MoCo) [30], and bootstrap your own latent (BYOL) [31]. The purpose of contrast learning is to learn the common features of similar samples and the differences between nonsimilar samples. Similar samples and nonsimilar samples are given a positive view and a negative view, respectively. Representations of positive views are pushed together during pretraining, while the representations of negative views are pulled apart [11]. The model learns the similarity between positive views and the differences between negative views in the process of contrastive learning. Consequently, the model learns the features needed to discriminate the samples. Some of the contrastive learning methods are introduced to ECG. In [32], random crop, time out, and physical noise are utilized as data augmentation methods of SimCLR for ECG. The algorithm is based on instance discrimination and latent forecasting. In [33], CPC is applied to ECG. CPC learns how to infer global structure in the signal, rather than only model complex local relationships. Also, Gopal et al. [12] proposed a physiologically inspired contrastive learning approach that generates views using 3-D augmentations of the 12-lead ECG, whose name is 3KG. Contrastive learning of cardiac signals (CLOCSs) is proposed in [34], which contains contrastive multisegment coding (CMSC), contrastive multilead coding (CMLC), and contrastive multisegment multilead Coding (CMSMLC). CLOCS encourages representations across space, time, and patients to be similar to one another. As for generative learning methods, U-ResNet-based models are proposed in [35]. With the help of UNet, the model learns the features by reconstructing ECG signals.

## III. Materials and Methods

### A. Datasets

*1) China Physiological Signal Challenge in 2018 (CPSC2018) Database:* The CPSC2018 database contains 6877 12-lead ECG recordings lasting from 6 s to just 60 s. The sampling rate of the raw signals is 500 Hz [36]. To reduce computing costs and ensure the availability of data at the same time, the signals are downsampled to 250 Hz according to existing research and previous experience [37], [38]. ECG signals in this database could be divided into nine categories according to diagnosis. The distribution of diagnosis is shown in Table I.

Due to the diversity of signals in CPSC2018, the dataset is suitable to be used as the pretraining dataset and the downstream dataset. Signals in separate categories offer diverse information, which will be learned through pretraining. Knowledge diversity makes pretrained models more powerful in extracting information.

TABLE I
Distribution of Diagnosis of CPSC2018

| Records | Type |
|---|---|
| 918 | Normal (N) |
| 1089 | Atrial fibrillation (AF) |
| 704 | 1st-degree atrioventricular block (I-AVB) |
| 207 | Left bundle branch block (LBBB) |
| 1695 | Right bundle branch block (RBBB) |
| 556 | Premature atrial contraction (PAC) |
| 672 | Premature ventricular contraction (PVC) |
| 825 | ST-segment depression (STD) |
| 202 | ST-segment elevated (STE) |

TABLE II
Distribution of Diagnosis of PTB-XL

| Supercl ass | Description | Label | Train | Valid | Test |
|---|---|---|---|---|---|
| NORM | Normal ECG | Single | 7254 | 916 | 913 |
| | | Multi | 7607 | 964 | 957 |
| MI | Myocardial Infarction | Single | 2048 | 234 | 256 |
| | | Multi | 4389 | 553 | 544 |
| STTC | ST/T Change | Single | 1353 | 172 | 184 |
| | | Multi | 3912 | 523 | 534 |
| CD | Conduction Disturbance | Single | 416 | 64 | 56 |
| | | Multi | 4193 | 498 | 497 |
| HYP | Hypertrophy | Single | 1907 | 256 | 243 |
| | | Multi | 2121 | 263 | 271 |

*2) PTB-XL ECG Dataset:* PTB-XL ECG dataset is a large dataset of 21 837 clinical 12-lead ECGs from 18 885 patients with 10-s lengths. ECG signals in this database are single label or multilabel. The distribution of diagnosis and the description are shown in Table II. Every signal available has two sampling rates, 100 and 500 Hz [39]. The sampling rate of 100 Hz is just so low that a lot of details are missing. Thus, only the signals with 500 Hz are selected as the raw signals in this experiment. To be consistent with the pretraining dataset, ECG signals in this dataset are resampled from 500 to 250 Hz.

According to the recommended method of PTB-XL in [39], the signals are divided into ten folds. The signals in the ninth and tenth fold are used as the validation set and the test set because ECG signals in the two folds are all validated by humans and considered clean in accordance with [39]. The signals in the rest eight folds are employed as the training set. It is worth mentioning that the ECG signals of the training set and test set are from different patients. This mode is known as the interpatient paradigm. From another perspective, the intrapatient paradigm means data in the training set and test set could be from the same patients. Therefore, models trained in the dataset, which comply interpatient paradigm, are more robust and have better generalization capability than the intrapatient paradigm [40]. Single-label classification and multilabel classification are two distinct downstream tasks, which require different methodologies. In single-label classification, ECG signals with multilabel in the dataset are

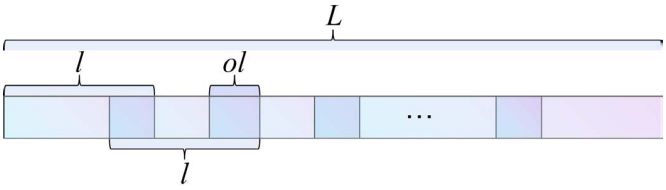



Fig. 1. Method to slice ECG signals of CPSC2018. To reduce serendipity, the length of overlap segments is ensured the same.

removed. Consequently, the numbers of signals in the training set, validation set, and test set are 12 978, 1642, and 1652, respectively. In multilabel classification, all the ECG signals are included.

There are 17 441, 2203, and 2193 ECG signals in the training set, test set, and validation set, respectively. As one of the mainstream methods, single-label classification is more convenient for comparison with different algorithms, so single-label classification is mainly evaluated. Experiments on multilabel ECG signals are conducted as well. Because medical data do not always fall into one category, multilabel classification has more clinical utility.

It is worth noting that there is a domain gap between the pretrain dataset and the downstream dataset in this experiment. As a result, the model pretrained is likely to reach absolute performance [11].

*3) Ningbo First Hospital 12-Lead ECG Open Database:* Ningbo First Hospital 12-lead ECG database (abbreviated as Ningbo database) is a large ECG dataset, containing 34 905 recordings from Ningbo First Hospital. According to clinical relevance, 11 rhythms labeled by certified physicians were merged into four groups (SB, AFIB, GSVT, and SR), SB only included sinus bradycardia, AFIB consisted of atrial fibrillation and atrial flutter (AFL), GSVT contained supraventricular tachycardia, and atrial tachycardia, atrioventricular node reentrant tachycardia, atrioventricular reentrant tachycardia and wandering atrial pacemaker, and SR included sinus rhythm and sinus irregularity [41]. The dataset is utilized as a pretraining dataset in the supplemental experiments.

### B. Data Preprocessing

For the reason that the input signals of MaeFE should be of equal length, data resizing is necessary. ECG signals in PTB-XL and Ningbo database are all regular 10 s, but the ones in CPSC2018 last from 6 to 60 s. Therefore, ECG signals less than 10 s are resampled to 10 s. The resampling was carried out using the fast Fourier transform (FFT) provided by SciPy's resample function. ECG signals that are more than 10 s are sliced according to Fig. 1.

In (1), $\text{Seg}_i$ represents the segment sliced from an ECG signal. The index of the segment is $i$. $L$ and $l$ are the total lengths of the signal and the length of the segment, respectively

$$\text{Seg}_i = [(i \cdot l - i \cdot ol), ((i+1) \cdot l - i \cdot ol)]. \tag{1}$$

$ol$ represents the length of overlap between adjacent segments. In one ECG signal, $L$ and $l$ are constant. Thus, $ol$ is

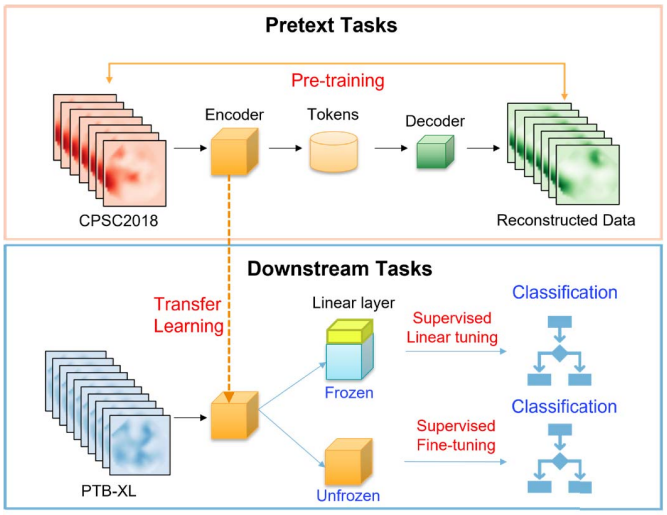



Fig. 2. Whole process of the algorithm, containing pretext tasks and downstream tasks that are reconstruction and classification of arrhythmia in this article.

constant as well. $ol$ is defined as follows:

$$ol = \frac{l\left\lceil \frac{L}{l} \right\rceil - L}{\left\lceil \frac{L}{l} \right\rceil - 1}. \tag{2}$$

In this way, the excess length is spread equally to each segment, and each segment from the same signal has an overlap with equal length. No segment overlaps a lot.

In this case, unequal-length signals are sliced into neat 10 s. There are 2500 sampling points in each segment due to the 250-Hz sampling rate. After slicing and resampling, the CPSC2018 dataset contains 12 002 ECG signals.

All the ECG signals utilized in this article are filtered. Wavelet transform filtering is a great choice for ECG signals. The method in [42] is employed in this work, which is nine-level db6 wavelet decomposition. After $D_1$, $D_2$, and $A_9$ components are removed, the filtered signals are reconstructed by the remaining components.

### C. MAE in ECG

Pretraining and downstream fine-tuning are the two stages of transfer learning. In this work, models are pretrained in a manner of self-supervised and fine-tuned in a supervised manner. Pretraining is generally based on pretext tasks. Pretext tasks are not of genuine interest but are solved only for the true purpose of learning a good data representation [30]. In this article, masking and reconstructing ECG are so-called pretext tasks, and the aim is to enhance the model's ability to classify arrhythmias in downstream tasks. The whole process of the algorithm is shown in Fig. 2.

*1) Encoder: 1-D-ViT and Sub2-D-ViT:* To complete the pretext tasks, ECG signals are sliced into patches at first, in which unmasked ones will be sent to the encoder. After training, masked patches are reconstructed by the decoder. In this experiment, only the encoder is utilized to produce ECG signal representations for recognition. Therefore, the encoder is the paramount factor that determines MaeFE's performance [13]. For the sake of that, the decoder could be

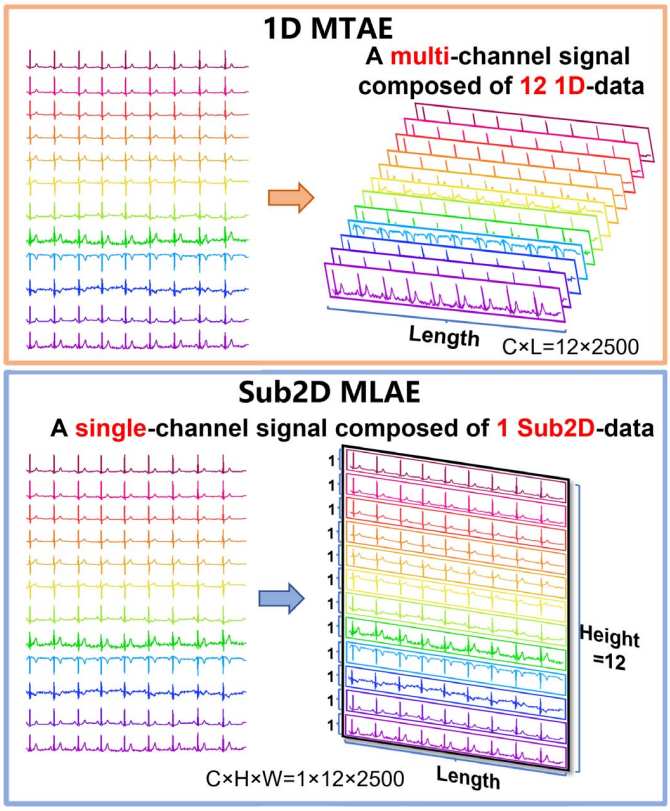



Fig. 3. Input of MTAE is 1-D data with 12 channels. The input of MLAE is sub2-D data with one channel, whose height is equal to 12.

designed lightweight to reduce the cost of training. Moreover, a suitable backbone should be selected for the encoder because the encoder is the essential factor for downstream tasks. ViT is utilized as the backbone of MaeFE in this experiment.

ViT is composed of a patch embedding module, a position embedding module, and a transformer module [14], [43]. ECG signals are sliced into patches first and transferred to tokens through the patch-embedding module. At the same time, the position information is integrated into the model through position embedding so that the reconstructed patches' location information is able to correspond to the original one. Finally, these tokens are sent to the transformer module, which is composed of six transformer blocks stacking. As a result, the data are encoded. ViT can analyze data from different modalities. 1-D-ViT and sub2-D-ViT are designed for diverse members of MaeFE, which will be discussed later.

*2) Members of MaeFE:* As a kind of autoencoder, MAE sends a few unmasked patches of ECG signals to the encoder and reconstructs the value of masked ones by the decoder. Numerous members of MaeFE can be designed according to different masking strategies. Here, three members of MaeFE are introduced.

*a) MTAE:* MTAE is based on 1-D-ViT, and thus, an ECG signal can be represented as a multichannel signal composed of 12 1-D data, as shown in Fig. 3. MTAE masks patches on the time scale. It means that whether the signals are masked or not depends on time. As we know, ECG signals have synchronized 12 leads. Therefore, after the time sequence is sliced into patches, data at the same time of all leads are bound in the same patch. Each patch contains information on the same position of all the 12 leads in the time domain. In other words, masked patches' information of all 12 leads is masked. Defined as (*Lead*, *Length*), the data's size is (12, 2500) and the patches' size is (12, 25). This is because 25 sampling points are the patch length, as shown in Fig. 4. The mask part is replaced by a weak Gaussian noise. As shown in Fig. 4, the shaded part represents the mask, which is aligned according to time. The reason is that each heartbeat has about 200 sampling points, and using 25 as the patch size can help models to focus more on the local features of heartbeats. Therefore, an ECG signal is sliced into 100 patches, each patch containing 25 sampling points. The masking rate stands for the percentage of masked patches. Different masking rates of MTAE directly determine the difficulty of the pretext tasks, and it will affect the robustness of the pretrained model.

*b) MLAE:* Another member is MLAE. In such a case, the leads of each ECG signal are masked. To make the data more applicable to the model, 1-D ECG signals are transferred into sub-2-D ones. The so-called sub-2-D means that an ECG signal of each lead is a 2-D picture with only 1 height. This means that 1-D-signal with 12 channels is turned into sub2-D-pictures with one channel whose height is 12, as shown in Fig. 3. Therefore, the sizes of signals and patches are (1, 12, 2500) and (1, 1, 2500), respectively, which is defined as (*Channel*, *Height*, *Width*), as shown in Fig. 4. Each patch contains all the information for a particular lead. The masked leads are randomly selected for each training step. Only the number of masked leads can be tuned. Therefore, if the number of masked leads is fixed and the epochs are large enough, every possibility will occur in the process of pretraining. In this way, the model can acquire stronger information extraction ability and better generalization performance. To some degree, the number of masked leads is equal to the masking rate in MTAE. MTAE randomly chooses patches to be masked with a fixed masking rate, and MLAE randomly chooses leads to be masked with a fixed number of masked leads.

*c) MLTAE:* MLTAE is an integrated strategy, and it merges the aforementioned MTAE and MLAE. These autoencoders are trained according to two different strategies. MTAE masks all leads of some time, and MLAE masks all data of some leads. The encoders of MTAE and MLAE both output data via multilayer perceptron heads (MLP-head), which is the output module of the ViT-based model. In the pretraining stage, MLP-head is used for feature extracting, and thus, the output of MLP-head is 512-D in this work. In downstream tasks, MLP-head is used for classification, and thus, the output of MLP-head depends on the number of classes, which is 5 in this work.

After pretraining, features are extracted from the encoders' MLP-head. MLP-head is an MLP with a hidden layer in the pretraining phase, which is replaced by a single linear layer in the fine-tuning phase instead. Before being forwarded to the downstream classifier, the features would be fused into multifeatures. That means that features extracted from MTAE and MLAE are concatenated. Because 512-D MTAE features and 512-D MLAE features are merged, the features of MLTAE have 1024-D.

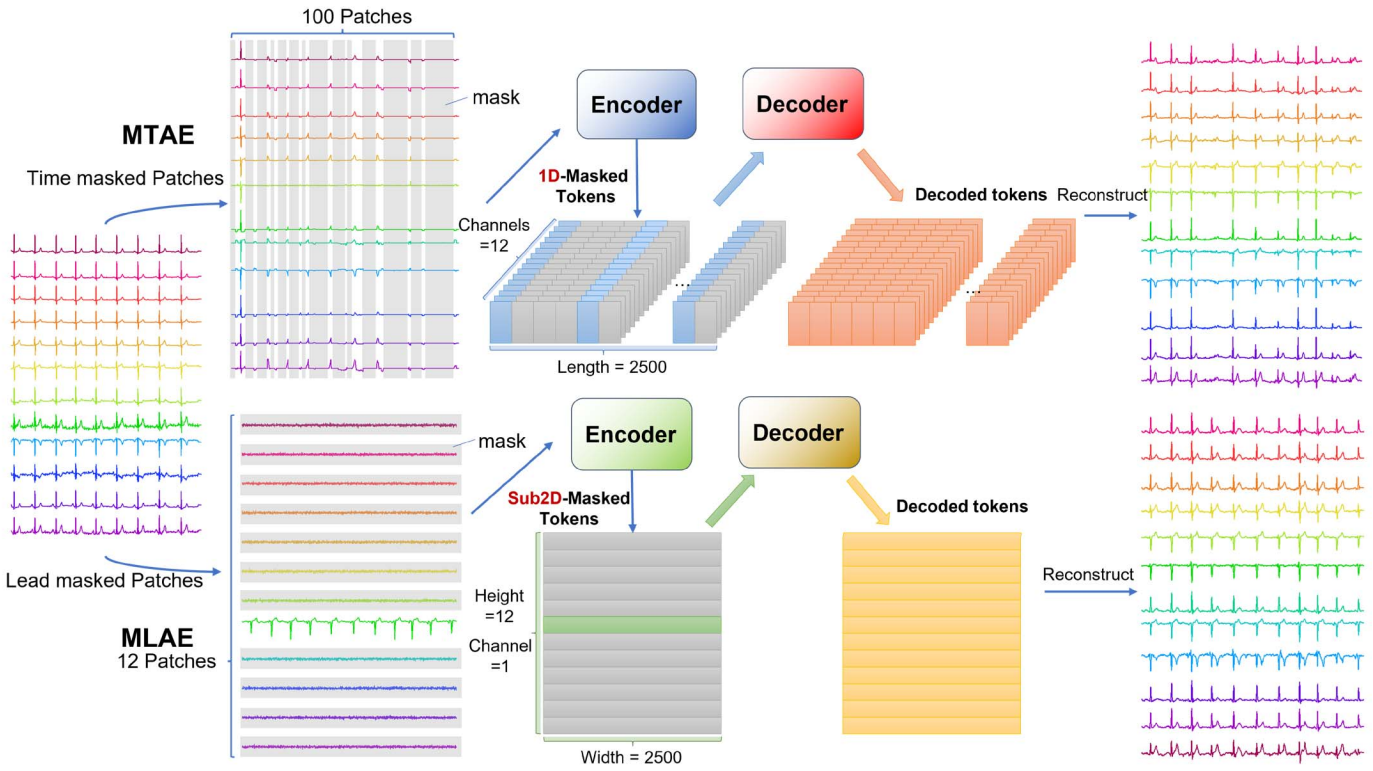



Fig. 4. Pretext tasks with two masking strategies of MaeFE. MTAE transfers partial masked ECG (shaded) into 1-D tokens with 12 channels, and then, masked patches are reconstructed by the decoder. MLAE transfers ECG with some leads masked (shaded) to sub2-D tokens with 8 height and 2500 width, and then, masked patches are reconstructed by the corresponding decoder. Shadow stands for mask.

*3) Pretraining:* 1-D-ViT and sub2-D-ViT are used as encoders of MTAE and MLAE, respectively. Only visible tokens are sent to the encoders in the pretext tasks. The reason why masked tokens are discarded is that these tokens will not exist in downstream tasks. If all of the tokens are utilized, there will be a gap between the pretraining model and downstream tasks. The gap will cause interference and affect the performance [15]. The position information of each patch is recorded through the aforementioned learnable position embedded layer so that masked patches will be reconstructed in the right place.

As for the decoder, it is decoupled with the encoder. The decoder is only used to reconstruct data in the pretraining stage. The real purpose of downstream datasets is not reconstruction, but arrhythmia classification. What is more, the design of the encoder and decoder can be asymmetric because only the encoder is utilized to produce ECG signal representations for recognition. To reduce computing costs and save source and time, a simple and lightweight decoder is designed. The decoder consists of merely one transformer block. In addition, all the tokens are sent to the decoder, including masked tokens and unmasked ones. This is distinct from the encoder. It is worth noting that the masked tokens are not transferred from masked patches. Instead, they are learnable and shareable tokens generated during the training process.

The input data of the encoder are normalized by Z-score to ensure comparability between data. The Z-score, also known as the standard score, is the process of dividing the difference between a score and the mean by the standard deviation. $\text{input}_i$, $\text{input}_i^*$, $\mu$, and $s$ stand for the input, standardized input, mean of input, and standard deviation of input, respectively. The process of the pretext task is shown in Fig. 4

$$\text{input}_i^* = \frac{\text{input}_i - \mu}{s}. \tag{3}$$

To train the pretraining model, mean squared error (MSE) is chosen as the loss function, and the MSE is defined as (4). Parameters in the formula stand for observed value and predicted value

$$\text{MSE} = \frac{1}{M}\sum_{m=1}^{M}\left(\text{observed}_m - \text{predicted}_m\right)^2. \tag{4}$$

MSE is the loss function in the pretraining stage. The loss is calculated by comparing the true value and predicted value of all masked patches. Furthermore, the AdamW optimizer is chosen to optimize MaeFE. MLAE and MTAE are trained according to the two masking strategies mentioned above.

The results of pretext tasks are shown in Fig. 5. As presented in Fig. 5, MTAE reconstructs the masked patches better if the training set and test set are the same. In addition, MTAE pretrained in CPSC2018 reconstructs different kinds of ECG signals in CPSC2018 and PTB-XL databases. It can be seen that there is little difference in the reconstructive performance of different ECG categories.

*4) Transfer Learning:* After being pretrained on the CPSC2018 dataset, the encoder of MaeFE is applied to the PTB-XL database as a classifier. The downstream task is the classification of arrhythmia, which is a five-category classification job.

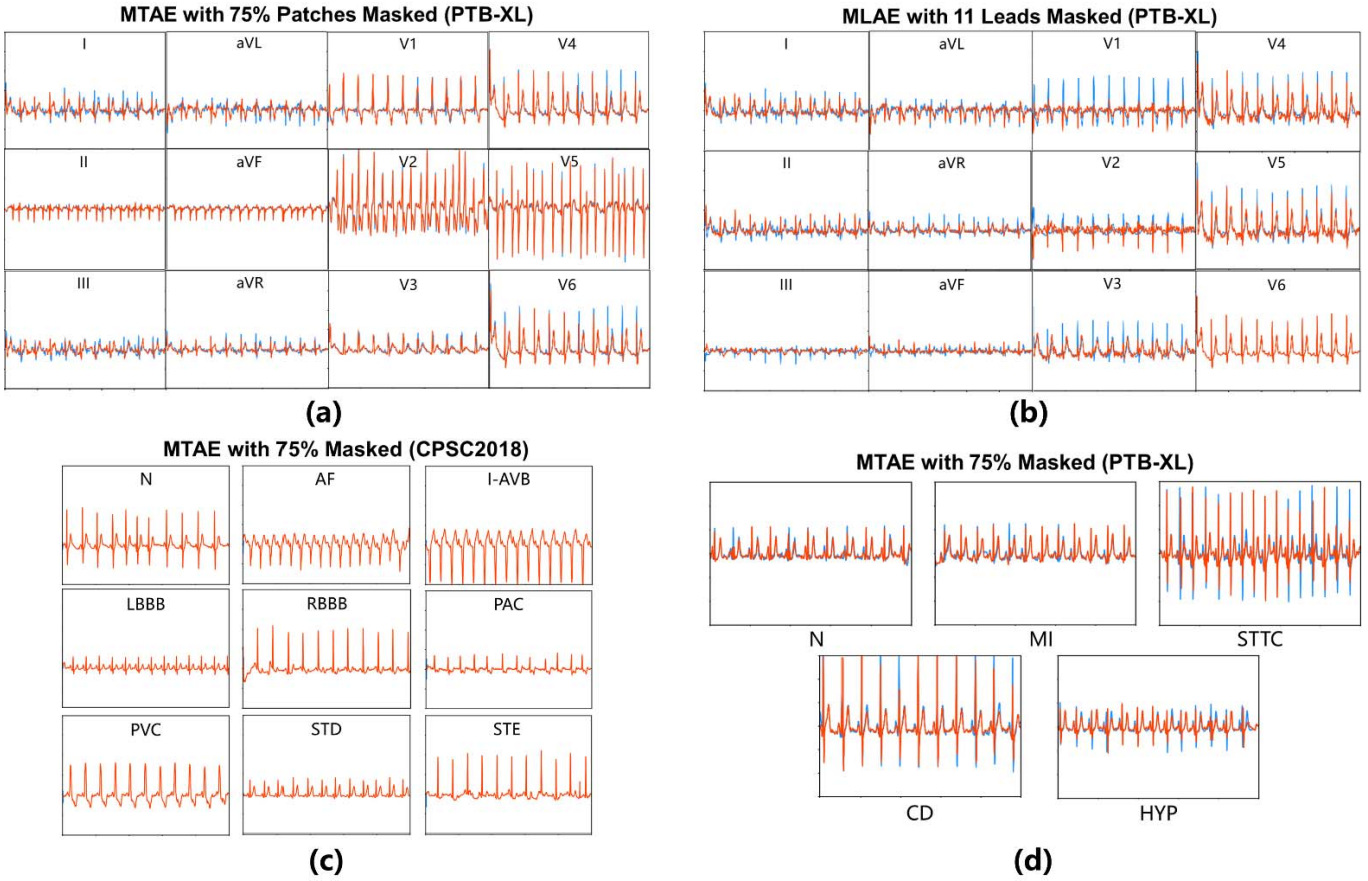



Fig. 5. Reconstructions of an ECG signal. All of the 12 leads of one ECG are reconstructed. Original signals (blue line) and reconstructed signals (orange line) are presented. All the models are pretrained on the CPSC2018 database. (a) Reconstruction of an ECG signal from PTB-XL by MTAE with 75% patches masked. (b) Reconstruction of an ECG signal from PTB-XL by MLAE with 11 leads masked. (c) Reconstruction of different types of ECG signals from CPSC2018, only lead II is shown. (d) Reconstruction of different types of ECG signals from PTB-XL, only lead II is shown.

To make the encoder of MaeFE suitable for the target task, tuning in the downstream dataset is introduced. The tuning methods can be divided into linear probe and fine-tuning. Fine-tuning adjusts all the parameters of the pretrained model, without any layers frozen. Thus, the performance of the fine-tuned model is more dependent on the backbone itself, rather than pretraining. While the linear probe freezes all the weight of the model except the last fully connected layer, the output nodes in the last fully connected layer are modified to fit downstream tasks, which stand for the number of categories. Therefore, a linear probe characterizes the performance of pretraining more accurately and directly. In a way, the encoder of MAE is utilized as a feature extractor. The obtained features are linked to what the encoder learns in the pretraining stage. The fully connected layer acts as a linear classifier that tunes the model to fit the downstream dataset. The reason is that the linear classifier has weak learning ability, and the results can better represent the performance of the pretraining model. Therefore, the effect of pretraining is mainly evaluated by the linear probe. The process of transfer learning and downstream tasks is shown in Fig. 6.

*5) Training and Evaluation Protocol:* As for the training and evaluation protocol of downstream tasks, AdamW is invoked as the optimizer, and cross-entropy (CE) is employed as the loss function.

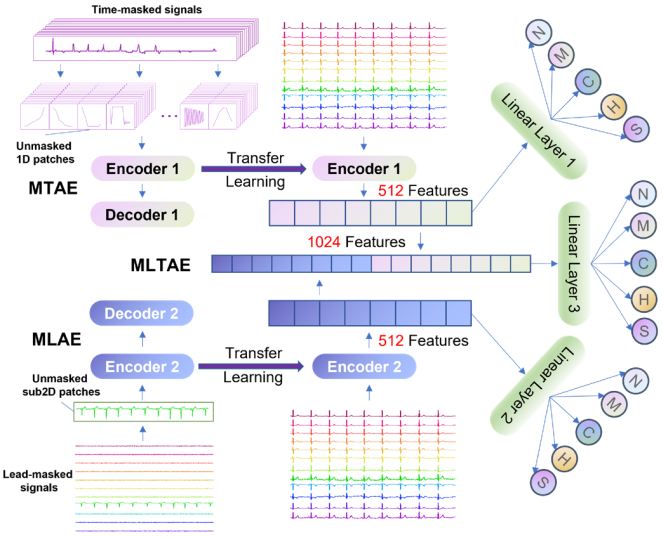



Fig. 6. Process of encoding with ViT and transfer learning. Models are trained independently using MLAE and MTAE. As linear classifiers, encoders of MTAE and MLAE classify data in the downstream dataset. MLTAE concatenates the features extracted by MLAE and MTAE.

Accuracy (Acc), macro-$F_1$ ($F_1$), and area under the curve (AUC) are introduced as metrics of evaluation. Among these metrics, Acc is a more common one, which is a direct

reflection of the inferred results. Macro-$F_1$ and AUC are more balanced, which minimizes the impact of data imbalance. TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. *Classes* means the number of subclasses, which are arrythmia in this work

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{Macro } F_1 = \frac{1}{\text{classes}} \sum_{i}^{\text{classes}} 2 \times \frac{\text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i} \tag{6}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

Macro-$F_1$ is the harmonic mean of precision and recall, it is more intuitive to reflect the precision and recall and can find the balance between them. AUC stands for the area under the receiver operating characteristic (ROC) curve. Original AUC is a metric for binary classifiers. To make it available for multiclass classifiers, one-versus-one AUC is introduced, which computes the average AUC of all possible pairwise combinations of classes. By the definition of these two metrics, Macro-$F_1$ is a visual representation of a model's performance and can provide a more comprehensive evaluation of a model. While AUC describes the discrimination of a model, it means that AUC focuses more on the ability to make a distinction between positive and negative samples, i.e., the overall performance, but ignores the differentiation between positive and negative samples. Therefore, considering the above three metrics together is beneficial to better evaluate the models. In general, higher Acc, Macro-$F_1$, and AUC represent better performance.

## IV. Experiments and Results

There are 71 kinds of diagnostic labels in PTB-XL, which were aggregated into five classes, including normal (NORM), myocardial infarction (MI), ST/T change (STTC), conduction disturbance (CD), and hypertrophy (HYP). The downstream tasks focus on the classification of five-category.

### A. Masking Rate in MTAE

An ablation experiment is carried out to find a suitable masking rate for MTAE. The hyperparameter is tuned in the process. In the experiment, masking rates are set from 10% to 90% in the phase of pretraining. In the downstream tasks, the pretrained encoder is utilized as the feature extractor. The encoder extracts 512 characteristics from each signal of the downstream database. These features are delivered to the linear classifier and classified into five categories. In this process, the model is tuned with CE loss. According to the experimental results, it can be seen that 75% of patches masked are the better choice due to their great behaviors in Acc, $F_1$, and AUC in agreement with Fig. 7.

In the pretraining stage, a higher masking rate makes it more difficult to reconstruct the ECG signals because more information is missing. As shown in the figure, models pretrained by tougher pretext tasks have more chances to obtain better

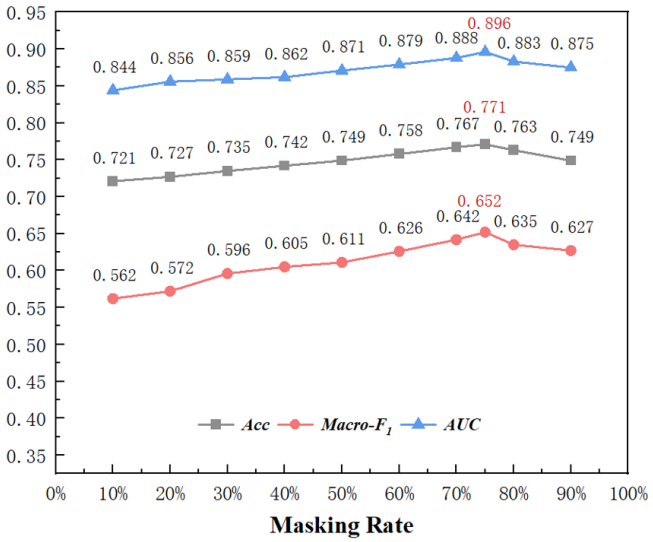



Fig. 7. Results of MTAE with different mask ratios in downstream tasks. Masking rates utilized are from 10% to 90%, and the evaluating metrics are Acc, Macro-$F_1$, and AUC.

performance from overall perspectives, which is dependent on the nature of unsupervised learning. To accomplish tougher pretext tasks, models should make effort to be equipped with greater capabilities to extract key information. The capability is both needed in pretext tasks and downstream tasks. As a result, the model will perform well in the downstream tasks. However, models' behaviors are not always positively correlated with the masking rate. A higher masking rate brings more training time and higher risks of training failure. For the sake of inadequate information is available, models have difficulty in finding potential features and connections in limited information. Thus, a large masking rate might lead to the failure of pretraining, and it might bring a bottleneck in learning as well. With a too high masking rate, the pretraining can only converge at locations where the loss is enormous or even not converge. In this situation, models can only learn very basic knowledge and cannot achieve successful reconstructions. In addition, characteristics extracted by these models cannot represent specific classes. Thus, downstream tasks will fail as well. In [13], 75% is considered the best masking rate for vision due to its good performance in both linear probe and fine-tuning. The results are consistent with the experimental results of MTAE.

However, hidden information among ECG patches is not as fertile as that among pictures because ECG signals are 1-D. Furthermore, there is less intuitive information among ECG patches, such as the edges of an object. These factors enhance the difficulty of reconstructing ECG signals. The reason why 75% masked MTAE still has good performance is that ECG signals are cyclical. Synchronized segments in different leads are concatenated as one patch. Thus, masked patches can be reconstructed easier because the only variable parameter is time. Due to the periodicity of ECG signals in the time domain, some missing information could be inferred by relating unmasked patches, as shown in Fig. 8. Moreover, the gap among ECG signals is smaller than that among the objects in the pictures. It means that many basic and general

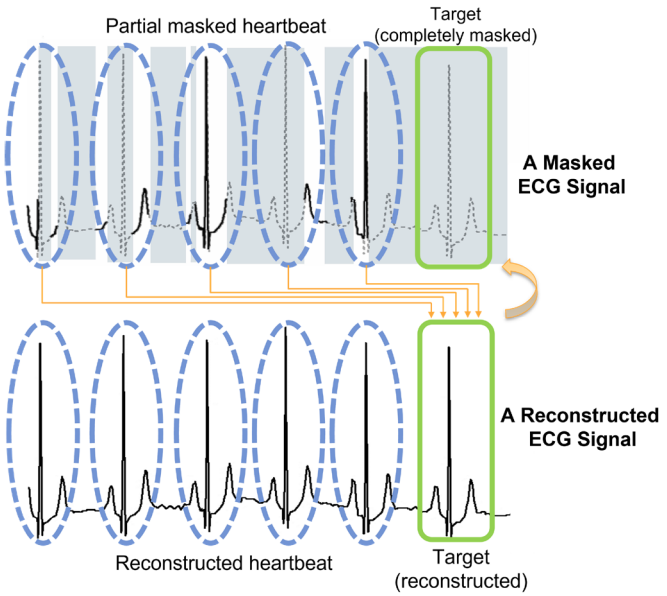



Fig. 8. Example that reconstructing a completely masked heartbeat using MTAE. Much useful information can be acquired from other partial masked heartbeat so that the difficulty of the reconstructing mission will be reduced. Shadow stands for mask.

features could be learned even if there are merely a few characteristics available. The aforementioned reasons reduce the difficulty of reconstruction. Therefore, MTAE can still achieve a nice performance with 75% of patches masked. However, recognition of deeper features requires a more robust feature extraction capability, which is why a high enough masking ratio is still needed. To sum up, an acceptably sharp masking rate of 75% is advisable in this experiment.

### *B. Number of Masked Leads in MLAE*

In the experiment of MLAE, the number of masked leads is set from 1 to 11. After rebuilding the signals, the pretrained encoder of MLAE with distinct making schemes is evaluated. The output of the encoder is also 512-D, representing 512 characteristics for an ECG signal. These features will be classified into five categories through the fully connected layer as well.

It is noted that MLAE pretrained with only one lead unmasked performs better, no matter in Acc, Macro-$F_1$, or AUC. The results are presented in Fig. 9.

As shown in the figure, MLAE does not perform as well as MTAE in this downstream task. Consistent with the aforementioned viewpoint, the experimental results demonstrate that only pretext tasks hard enough can train a powerful model. Although MLAE masks 11/12 (91.67%) of the signal, pretext tasks of MLAE are not as difficult as MTAE's pretext tasks because much information on different leads appears correlatively. As a result, a large amount characteristics can be recognized easily for the reason that one complete and precise signal is given. Furthermore, 12 leads are not entirely independent. Due to the component of ECG, data redundancy exists. Some leads can be mathematically calculated by other leads. As shown in (9)–(12), in which all the parameters are leads of ECG

$$\mathrm{II} = \mathrm{I} + \mathrm{III} \tag{9}$$

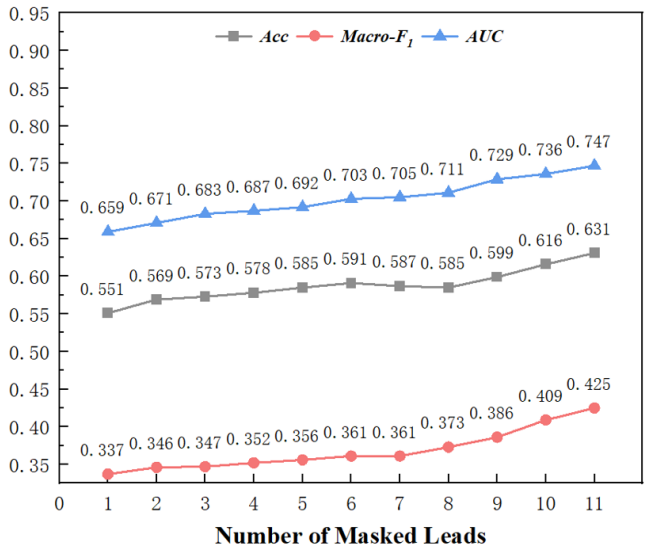



Fig. 9. Results of MLAE with a different number of masked leads in downstream tasks. The numbers of masked leads are from 1 to 11, and the whole number is 12.

$$\mathrm{aVR} = -\frac{(\mathrm{I} + \mathrm{II})}{2} \tag{10}$$

$$\mathrm{aVL} = \frac{(\mathrm{I} - \mathrm{III})}{2} \tag{11}$$

$$\mathrm{aVF} = \frac{(\mathrm{II} + \mathrm{III})}{2}. \tag{12}$$

Without MLAE, the value of signals in some leads could be calculated by signals in other leads. Thus, the nonindependence of leads reduces the difficulty of reconstruction. However, 12-lead ECG signals are still utilized in this work rather than eight-lead ones. There are three main reasons for it. First, although the mathematical calculations are simple, the mapping relationships among the leads are valuable and will make a positive impact. Second, situations that occur in eight-lead MLAE all occur in 12-lead MLAE. Due to the random selection of leads and enough epochs, these situations play sufficient roles. Third, compared to eight-lead MLAE, 12-lead ones offer more direct data without mathematical transformation. In addition, because lead masking is random, 12-lead MLAE has more experiments and contains more situations. Therefore, 12-lead MLAE can better excavate the relationship between leads than eight-lead MLAE. More characteristics might be extracted by 12-lead MLAE.

From another perspective, the ultimate goal of this experiment is the classification of arrhythmia. Classifying arrhythmia relies more on temporal characteristics and morphological features. MTAE is competent to learn features between patches of a different time and pay more attention to morphological features, whereas MLAE focuses more on the relationship among leads. Information about the relationship reflects the location of cardiac activity or the lesion. Diagnosing cardiac activities depends more on temporal and morphological characteristics [44]. With less temporal and morphological information, MLAE can hardly gain superior performance in arrhythmia classification, but MLAE will be suitable for downstream tasks that rely more on the relationship among leads. For example, downstream tasks can be determining the

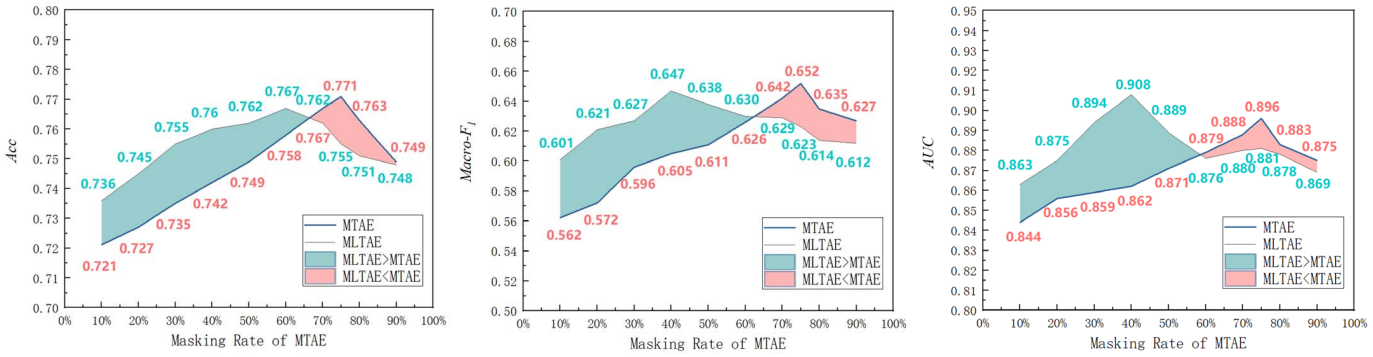



Fig. 10. Comparison of MLTAE and MTAE. Two cases in each chart are MTAE with different masking ratios, and MLTAE, which combines MLAE with 11 leads masked and MTAE with different masking ratios. The masking ratios chosen are from 10% to 90%.

TABLE III
BINARY PREDICTION OF AMI AND IMI

| DATASET | MODE | $F_1$ | ACC | AUC |
|---|---|---|---|---|
| PTB-XL | Supervised | 0.321 | 0.473 | 0.500 |
| | Fine-tuned (MLAE) | **0.476** | **0.525** | **0.574** |

location where the lesion occurred, identifying the site of MI, and distinguishing whether the disease is ventricular or atrial. To prove that MLAE is operative in some tasks, a binary prediction on anterior MI (AMI) and inferior MI (IMI) is done. The results are presented in Table III.

The main difference between AMI and IMI is that the site of onset is different. AMI occurs in the anterior left ventricular wall of the heart, while IMI occurs in the lower left ventricular demonstrating that MLAE can use spatial information for classification. Therefore, MLAE plays a functional role in some specific tasks. Note that the poor performance of the supervised model is due to small data volume, data imbalance, and large backbone. AMI and IMI are merely subclasses of MI in the PTB-XL, which has only 354 AMI samples and 2685 IMI samples. In this case, overfitting is easily induced, and it makes the performance not so good.

### C. Combination of MLAE and MTAE

MLTAE has a multihead structure that combines the features extracted from the encoders of MTAE and MLAE. The features extracted from MTAE and MLAE are both 512-D, and thus, the fusion features are 1024-D. In linear evaluation, the 1024-D features are utilized as the input. Depending on previous experimental results, MTAE performs better in this downstream task. Therefore, merely, the best MLAE is supposed to be combined with various MTAEs. Consequently, MTAE with different mask ratios and MLAE with 11 leads masked are chosen. The results are shown in Fig. 10.

From Fig. 10, it can be seen that the fusion model performs better than the pure MTAE in making a rate less than approximately 60%. The performance of MTAE is better when the masking rate is higher. In addition, different from MTAE, MLTAE achieves its best performance when the masking rate of MTAE is 40%, rather than 75%. It is worth mentioning that MLTAE achieves the highest AUC in linear evaluation, which is 0.908. Also, MTAE outperforms MLTAE in max Macro-$F_1$, which is 0.657.

Intuitively, MLTAE should perform better than MTAE, but the latter achieves the highest Macro-$F_1$. Such results are explained by an incomplete match between the features of MLTAE. In detail, to avoid overfitting and enhance generalization, layer normalization (LN) is introduced. LN indicates that all features of the same sample are normalized. Consequently, the relationships between samples are removed. For the sake that features extracted from MLAE are not absolutely accurate, they will influence features extracted from MLAE by LN. In other words, if the two parts of the features interfere with each other more than their mutual contributions, it is possible that the MLTAE does not perform better than the pure MTAE. The interference brought by LN, and the contributions are supplementary information from MLAE.

### D. Ablation Experiments on ResNet

To investigate the sensitivity of MaeFE to the backbone, ablation experiments on another architecture are conducted. The previous experiments were based on a ViT-based backbone. As one of the representatives of CNN structure, ResNet can be utilized as the backbone of MAE as well. Due to the difference in the way CNN and transformer analyze data, the entire process needs to be redesigned. After ECG signals are sliced into patches, masked patches are replaced by faint Gaussian noise. ResNet is selected as the encoder. Different from ViT, all the patches are directed to the encoders, rather than unmasked ones. Decoders are composed of five deconvolutional layers, without residual blocks. Considering that MTAE performs better in linear evaluation, it is chosen as the structure used in ablation experiments. The results are shown in Fig. 11.

As shown in Fig. 11, the overall trend of MTAE with ResNet is the same as MTAE with ViT. They both point out a conclusion that pretraining with the right level of difficulty is more efficient. ResNet obtains the best performance at a masking rate of 60%. The difference in best masking rate is due to differences in architecture and masking way, and they affect the difficulty of pretraining to varying degrees.

### E. Experiments on Other Datasets

To evaluate the model more comprehensively, another dataset named the Ningbo database is added to the

TABLE IV
EXPERIMENTAL RESULTS OF MAEFE IN DIFFERENT DATASETS

| Pre-train Dataset | CPSC2018 | | | Ningbo | | | PTB-XL | | | Ningbo | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Downstream Dataset | PTB-XL | | | | | | CPSC2018 | | | | | |
| | *Acc* | $F_1$ | *AUC* | *Acc* | $F_1$ | *AUC* | *Acc* | $F_1$ | *AUC* | *Acc* | $F_1$ | *AUC* |
| Only Pretrained (MLAE) | 0.557 | 0.142 | 0.544 | 0.557 | 0.142 | 0.544 | 0.233 | 0.042 | 0.536 | 0.233 | 0.042 | 0.536 |
| Only Pretrained (MTAE) | 0.553 | 0.142 | 0.508 | 0.553 | 0.142 | 0.508 | 0.233 | 0.042 | 0.662 | 0.233 | 0.042 | 0.662 |
| Supervised (MLAE) | 0.582 | 0.394 | 0.704 | 0.582 | 0.394 | 0.704 | 0.314 | 0.244 | 0.688 | 0.314 | 0.244 | 0.688 |
| Supervised (MTAE) | 0.716 | 0.587 | 0.865 | 0.716 | 0.587 | 0.865 | 0.513 | 0.465 | 0.830 | 0.513 | 0.465 | 0.830 |
| *Linear Probe* | | | | | | | | | | | | |
| MLAE | 0.631 | 0.425 | 0.747 | 0.581 | 0.356 | 0.713 | 0.238 | 0.259 | 0.671 | 0.242 | 0.224 | 0.665 |
| MTAE | **0.771** | **0.652** | 0.896 | **0.765** | 0.625 | **0.896** | **0.457** | **0.430** | **0.820** | **0.424** | **0.412** | **0.815** |
| MLTAE | 0.767 | 0.647 | **0.908** | 0.761 | **0.635** | 0.887 | 0.411 | 0.386 | 0.788 | 0.410 | 0.382 | 0.790 |
| *Fine-tune* | | | | | | | | | | | | |
| MLAE | 0.603 | 0.420 | 0.734 | 0.573 | 0.381 | 0.707 | 0.348 | 0.304 | 0.722 | 0.323 | 0.289 | 0.705 |
| MTAE | **0.783** | **0.682** | **0.919** | **0.775** | **0.668** | **0.906** | **0.744** | **0.716** | **0.945** | **0.730** | **0.710** | **0.936** |

The table includes results of MaeFE with different pre-train dataset and downstream dataset. Linear probe and Fine-tune are introduced in the downstream tasks.

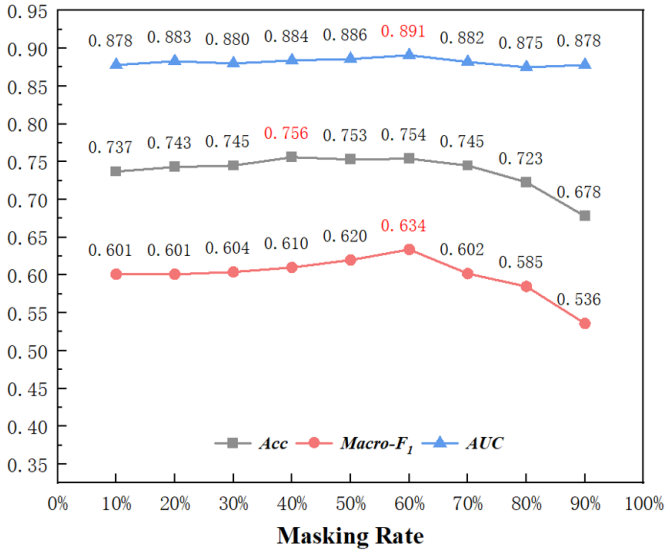



Fig. 11. Results of MTAE based on ResNet with different masking rates using linear probe.

experiments. The Ningbo dataset has a larger amount of data, and the domains of the data are much different from CPSC2018 and PTB-XL. Therefore, the introduction of the Ningbo dataset can better verify the universality of MaeFE. MaeFE's sensitive to data domains and data volume can also be studied. In addition, the experimental scheme is enriched, which can further verify the universality of MaeFE. In this experiment, MaeFE is pretrained on CPSC2018, Ningbo, and PTB-XL databases. In the downstream tasks, CSPC2018 and PTB-XL are chosen as the dataset for tuning. The results are shown in Table IV. In Table IV, "only pretrained models" denotes a pretrained model without any fine-tuning in the downstream tasks. "Supervised models" denotes the model with the same architecture, which is trained in a conventional supervised manner. All the backbones of the models in Table IV are ViT.

The results demonstrate more consistent performance from MTAE. Peak MLTAE performance sometimes outperforms peak MTAE. This result illustrates that another phenomenon large pretraining datasets do not necessarily achieve the best pretraining performance. There are 34905 ECG records in the Ningbo database, which is larger than CPSC2018 and PTB-XL. While MaeFE pretrained on CPSC2018 and PTB-XL both outperform the pretrained model on Ningbo. This is caused by the differences in data sampling methods and data domains. Therefore, selecting the right pretrained dataset can lead to better models at a limited cost.

Compared to models trained in a manner of supervised learning, fine-tuned MaeFE has a better performance in Acc, Macro-$F_1$, and AUC in all cases. It is shown that fine-tuning is effective and can avoid the model falling into some local optima. In linear evaluation, different downstream data sets lead to different conclusions. When downstream tasks are conducted on the PTB-XL database, linear-tuned MaeFE outperforms the supervised one, while the supervised ones perform better if experiments are conducted on the CPSC2018 database, as shown in Table IV. This is due to the distribution and diagnosis of downstream datasets. If the features required for the downstream classification tasks can be learned in the pretraining phase, then linear-tuned models will perform better; otherwise, some of the key information for classification is specific to the downstream dataset. In this case, supervised ones will perform better because all layers of the models should be tuned rather than merely the last fully connected layer.

### *F. Partial Fine-Tuning*

Experiments of fine-tuning are done as well. As another approach, fine-tuning means unfreezing more layers or blocks of pretrained models than the linear probe, even all the layers or blocks. In this way, models would fit downstream datasets better due to more supervised factors. Supervised factors introduce some prior knowledge, so the results would be better for fine-tuning. Because fine-tuning is closer to supervised learning, the results rely more on the performance of the model itself rather than the effect of pretraining. Considering that the linear probe can represent the effect of pretraining

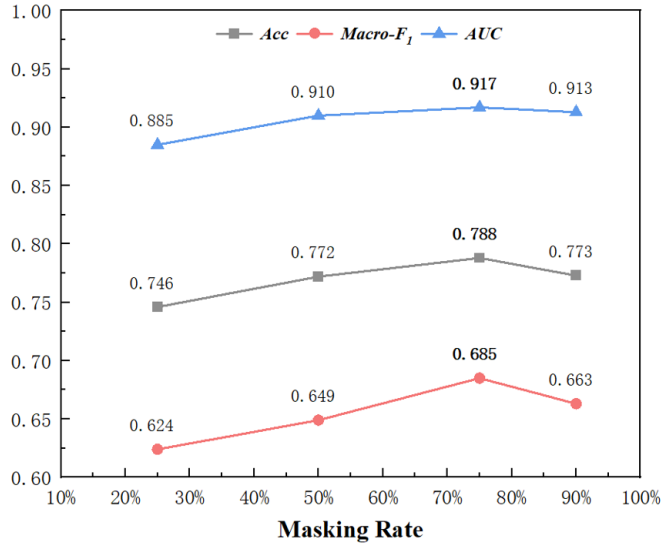



Fig. 12. Results of MTAE with different masking rates using partial fine-tuning.

more accurately, only MTAE is fine-tuned due to its better performance in the linear probe. Unlike linear probe, fine-tuning not only tunes the last fully connected layer but also the entire model.

An attempt was done to unfreeze all layers and blocks. However, it is strange to find that the results are terrible, even worse than the consequences of the linear probe. It could be noted that the dataset used in pretext tasks is not as large as the vision dataset such as ImageNet. Moreover, the similarity among ECGs is higher than that among pictures, which is likely to trigger overfitting. For the sake of that, partial fine-tuning is introduced. Only the last transformer block and the last fully connected layer are unfrozen [45]. In this way, performance will be improved and overfitting will be avoided meanwhile. The results are presented in Fig. 12.

As shown in Fig. 12, the best masking rate is still 75%, whereas the gaps among MTAE with various masking rates are narrowed, especially the gap of AUC. The more layers or blocks are unfrozen, the closer fine-tuning is to supervised learning. Fully supervised learning has the problem of overfitting; therefore, partial unfreezing can better utilize the model's ability to extract information. Moreover, due to the presence of pretraining, overfitting problems can be alleviated. Thus, the model is more likely to avoid trapping local optimums in downstream tasks.

### G. Multilabel Classification

To demonstrate the generalization ability of MaeFE, another downstream task is introduced, which is the classification of multilabel arrhythmia. Medical data do not always fall into just one category, and they are usually multilabel data. According to Table II, more data are discarded in the case of single-label classification. Moreover, in practice, whether clinical ECG data are multilabel data cannot be known in advance. Therefore, single-label classification algorithms have some clinical limitations, and multilabel classification algorithms are more suitable for clinical utility. The experiment is performed on PTB-XL because there are 71 kinds of statements in this

TABLE V
EXPERIMENTAL RESULTS OF MULTILABEL CLASSIFICATION ON PTB-XL

| Mode | Method | Pre-train Dataset | *AUC* |
|---|---|---|---|
| Only Pretrained | MLAE | CPSC2018 | 0.506 |
| | MTAE | CPSC2018 | 0.615 |
| Supervised | MLAE | - | 0.686 |
| | MTAE | - | 0.828 |
| Linear Probe | MLAE | CPSC2018 | 0.668 |
| | MTAE | CPSC2018 | **0.875** |
| | MLTAE | CPSC2018 | 0.862 |
| | MLAE | Ningbo | 0.670 |
| | MTAE | Ningbo | 0.861 |
| | MLTAE | Ningbo | 0.856 |
| Fine-tune | MLAE | CPSC2018 | 0.745 |
| | MTAE | CPSC2018 | **0.886** |
| | MLAE | Ningbo | 0.730 |
| | MTAE | Ningbo | 0.879 |

The table contains the results of multi-label classification of arrythmia on PTB-XL database.

database, and the labels could coexist in one ECG signal. In other words, ECG signals in PTB-XL could be multilabel data. MaeFE is pretrained on CPSC2018 and Ningbo databases.

In this experiment, the output data are a probability of length 71, distributed between 0 and 1. The threshold for determining whether the prediction is true or not is 50%. This means that if the probability is greater than 0.5, this category is predicted to be true. The results are presented in Table V. In Table V, the "only pretrained model" denotes a pretrained model without any fine-tuning, and the "supervised model" denotes a model trained in a conventional supervised manner. All the backbones of the models in Table V are ViT.

As the results present, MTAE performs better in multilabel classification, and it outperforms only pretrained MTAE by 42.3% in AUC. MTAE and MLTAE perform better than supervised ones. This suggests that the model's ability to generalize and prevent overfitting is enhanced by pretraining.

### H. Comparisons With State-of-the-Art Algorithms

To show the advantages of MaeFE more intuitively, experiments of related work are conducted as well. Comparative experiments mainly include some contrastive learning methods since they are the mainstream methods of self-supervised learning so far. SimCLR and MoCo are two kinds of dominant contrastive learning methods. Referring to the implementation of SimCLR in ECG in [32], random crop, time out, and physical noise are used as transforms to train contrastive models [46]. ViT and ResNet are chosen as the backbone of SimCLR. Similar transforms are adopted in MoCo [30]. In addition, momentum and queen are introduced in MoCo to address the problems of capacity and consistency. Data augmentations are cores of contrastive learning. Thus, after the data augmentation methods are adapted to ECG, the methodology will be unaffected by different modalities. In other words, SimCLR and MoCo in this experiment are customized for ECG. For comparison with self-supervised pretraining algorithms dedicated to ECG, CLOCS and MaskUNet are

TABLE VI
EXPERIMENTAL RESULTS OF RELATED WORK AND MAEFE IN LINEAR EVALUATION

| Method | Backbone | Linear Probe | | |
|---|---|---|---|---|
| | | *Acc* | $F_1$ | *AUC* |
| SimCLR (rrc, to)[32] | ResNet | 0.672 | 0.520 | 0.841 |
| SimCLR (phy.)[32] | ResNet | 0.651 | 0.501 | 0.835 |
| MoCo[30] | ResNet | 0.695 | 0.534 | 0.866 |
| SimCLR (rrc, to)[32] | 1D-ViT | 0.711 | 0.533 | 0.856 |
| SimCLR (phy.)[32] | 1D-ViT | 0.633 | 0.412 | 0.742 |
| MoCo[30] | 1D-ViT | 0.721 | 0.561 | 0.848 |
| MaskUNet[35] | UNet | 0.733 | **0.583** | **0.868** |
| CMSC[34] | ResNet | **0.738** | 0.574 | 0.866 |
| CMLC[34] | ResNet | 0.700 | 0.543 | 0.834 |
| CMSMLC[34] | ResNet | 0.706 | 0.523 | 0.841 |
| MTAE | ResNet | 0.754 | 0.634 | 0.891 |
| MLAE | Sub2D-ViT | 0.631 | 0.425 | 0.747 |
| MTAE | 1D-ViT | **0.771** | **0.652** | 0.896 |
| MLTAE | 1Sub2D-ViT | 0.767 | 0.647 | **0.908** |

The table contains the results of pre-trained MaeFE, pre-trained contrastive learning methods using linear probe.

TABLE VII
EXPERIMENTAL RESULTS OF RELATED WORK AND OUR ALGORITHMS IN PARTIAL FINE-TUNING

| Method | Mask | Partial Fine-tuning | | |
|---|---|---|---|---|
| | | *Acc* | $F_1$ | *AUC* |
| SimCLR (rrc, to)[32] | ResNet | 0.761 | 0.646 | 0.857 |
| SimCLR (phy.) [32] | ResNet | 0.779 | 0.654 | 0.883 |
| MoCo[30] | ResNet | **0.782** | 0.629 | **0.904** |
| SimCLR (rrc, to)[32] | 1D-ViT | 0.716 | 0.566 | 0.814 |
| SimCLR (phy.)[32] | 1D-ViT | 0.694 | 0.535 | 0.768 |
| MoCo[30] | 1D-ViT | 0.723 | 0.570 | 0.856 |
| MaskUNet[35] | UNet | 0.767 | 0.639 | 0.885 |
| CMSC[34] | ResNet | 0.780 | **0.666** | 0.891 |
| CMLC[34] | ResNet | 0.781 | 0.637 | 0.883 |
| CMSMLC[34] | ResNet | 0.772 | 0.648 | 0.901 |
| MLAE | Sub2D-ViT | 0.589 | 0.407 | 0.715 |
| MTAE | 1D-ViT | **0.788** | **0.685** | **0.917** |

The table includes results of MaeFE and dominant contrastive learning methods using fine-tuning. ViT and ResNet are used as backbone of SimCLR and MoCo.

TABLE VIII
TRAINING TIME OF MAEFE

| MODE | METHOD | TRAINING TIME PER EPOCH | MAX EPOCH | MAX TIME |
|---|---|---|---|---|
| Pre-training | MTAE | 28s | 9000 | 2.9d |
| | MLAE | 8s | 6000 | 13.3h |
| Linear probe | MTAE<br>MLAE | 0.16s | 2000 | 5.3m |
| Fine-tuning | MTAE | 23s | 2000 | 12.8h |

The table describes the max training time of MaeFE with a masking rate. Training time per epoch and the max epoch are contained as well.

introduced. CLOCS are contrastive learning methods for ECG, including CMSC, CMLC, and CMSMLC [34]. Positive pairwise in CMSC is segments from the same ECG, and it is different leads from the same ECG in CMLC. In CMSMLC, positive pairwise is composed of single-lead segments of ECG from the same person. CLOCS performs data augmentation on the temporal and spatial features of ECG. Thus, the comparison of CLOCS and MaeFE makes the work more convincing. In addition, as a generative self-supervised method designed for ECG, MaskUNet employs ResNet to UNet and pretrained models with 30% masked ECG signals. A comparison of MaskUNet and MaeFE is more intuitive in convincing because they are both generative self-supervised methods.

The results of the linear evaluation are presented in Table VI. In linear evaluation, MaskUNet performs better than other state-of-the-art self-supervised learning algorithms in Macro-$F_1$ and AUC. CMSC performs better than other contrastive methods in Acc, Macro-$F_1$, and AUC. As a generative learning method, MTAE outperforms MaskUNet by 5.18% in Acc, 11.8% in Macro-$F_1$, and 3.23% in AUC. Generative tasks that are difficult enough and the applications of patches help MaeFE to have stronger classification capabilities. MTAE outperforms CMSC 4.47% in Acc, 13.59% in Macro-$F_1$, and 3.46% in AUC. In addition, MLTAE achieves the highest AUC, which is 0.908. Also, MTAE gets the highest Acc and Macro-$F_1$, which are 0.771 and 0.654, respectively.

From another perspective, in fine-tuning, CMSC outperforms other contrastive learning methods and MaskUNet in Macro-$F_1$. MoCo based on ResNet acquires the highest Acc and AUC among contrastive learning methods. MTAE outperforms the contrastive learning model using the same backbone by 8.99% in Acc, 20.18% in Macro-$F_1$, and 7.13% in AUC, in fine-tuning, as shown in Table VII. Yet, MTAE merely outperforms contrastive learning models using ResNet as backbone by 0.77% in Acc, 2.85% in Macro-$F_1$, and 1.44% in AUC. It is due to the transformer-based model's large size and the lack of inductive bias [19], such as translation equivariance and locality. Therefore, transformer-based models will not generalize well if it trained on insufficient amounts of data [14]. Even so, MTAE can still gain an overall lead, which demonstrates the advantages of unsupervised pretraining based on MAE. In fine-tuning experiments, MTAE achieves the highest Acc, Macro-$F_1$, and AUC, which are 0.788, 0.685, and 0.917, respectively. In conclusion, the sufficiently difficult pretraining task and the optimization for ECG characteristics help MAE to obtain a stronger ability to extract information.

### *I. Hardware and Schedule*

The experiments are conducted on a PC with an Nvidia RTX3060 GPU, an Intel i5-12400 CPU, and 16-GB DDR5 memory. All the models are trained and evaluated by using Pytorch (v1.9.0), CUDA (v11.4), and cuDNN (v8.2.4).

Table VIII shows the max training time of MaeFE with a masking rate. The training time of one epoch and the number of epochs to train one model are contained. For MTAE, each training epoch took about 28 s. MTAE with 90% masked is the most time-consuming one, needing 9000 epochs (2.9 days). For MLAE, training time per epoch is almost 8 s. MLAE with 11 leads masked consumes more epochs, which is 6000 epochs (13.3 h).

In the linear probe, it spends 0.16 s per epoch. All the models can be tuned within 2000 epochs (5.3 min) in linear evaluation. In partial fine-tuning of MTAE, the number is 23 s per epoch. Two Thousand epochs are enough.

## V. Discussion and Conclusion

In this work, a family of MAE with various masking modes applied to ECG classification is put forward, named MaeFE. MaeFE provides customized MAEs for ECG and abstracts autoencoders into a family. It can be noted that MaeFE is superior to the mainstream contrastive learning approaches in the experiments, which have dominated vision self-pretraining in recent years. MaeFE also outperforms other generative self-supervised learning methods such as [35]. The outstanding performance of MaeFE is attributed to variable masking methods, discussion of different masking rates, and optimizations according to ECG characteristics. Considered the more comprehensive member of MaeFE, MTAE achieves a better performance in Acc, Macro-$F_1$, and AUC in both linear probe and fine-tuning. It demonstrates the robust performance of the algorithm using MAE. Achieving the highest AUC, MLTAE has the best discriminatory ability, while MLAE might be more suitable for downstream tasks that focus on the relationship between leads and the location of the disease. Experiments of multilabel classification and MAE based on ResNet are conducted as well. The results reflect MaeFE's robust generalization capability to additional downstream tasks and other structures. In addition, the applications of MaeFE in different pretrain datasets and downstream datasets demonstrate MaeFE's versatility.

Due to the lack of labeled clinical data, unsupervised learning is a more suitable way with a large amount of unlabeled data available. Unsupervised pretraining could effectively speed up the fitting process in downstream tasks and increase the robustness of the algorithm [47].

As for the backbone of MaeFE, ResNet and ViT are introduced in the experiments. ResNet is based on CNN, while ViT is based on the transformer. The comparison of the CNN-based model and the transformer-based models a hot topic recently. Inductive biases give CNN more attention to details, but local connectivity can lead to a loss of global context. Based on self-attention, the transformer contextually up weights the relevance of certain information [19]. In other words, CNN focuses on local information, and transformer concentrates on the global context. To take advantage of both CNN and transformer, models, such as BoTNet [48], CvT [49], and CoAtNet [50], are proposed. It is reasonable to believe that the combination of CNN and transformer may lead to marvelous results in MaeFE in further research.

In the future, the MAE-based algorithm might be optimized by pretraining in larger databases. In addition, MaeFE is probably advanced if a stronger backbone is adopted. Furthermore, innumerable ways of combining MTAE and MLAE are worth trying, which have promising potential to promote the algorithm. Alternative masking strategies could be explored as well, and it is the potential to bring new members to MaeFE.

## References


[1] M. Holm, "Non-invasive assessment of the atrial cycle length during atrial fibrillation in man: Introducing, validating and illustrating a new ECG method," *Cardiovascular Res.*, vol. 38, no. 1, pp. 69–81, Apr. 1998.

[2] Y. H. Awni et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, pp. 65–69, Jan. 2019.

[3] B. M. Mathunjwa, Y.-T. Lin, C.-H. Lin, M. F. Abbod, and J.-S. Shieh, "ECG arrhythmia classification by using a recurrence plot and convolutional neural network," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102262.

[4] M. Hammad, A. M. Iliyasu, A. Subasi, E. S. L. Ho, and A. A. A. El-Latif, "A multitier deep learning model for arrhythmia detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[5] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6707–6717.

[6] Z. Zhao et al., "Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–28, 2021.

[7] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost unsupervised learning for dense crowd counting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 8868–8875.

[8] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[9] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pretraining*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/languageunsupervised/language understanding paper

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 1–16.

[11] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2020.

[12] B. Gopal, R. Han, G. Raghupathi, A. Ng, G. Tison, and P. Rajpurkar, "3KG: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations," in *Proc. Mach. Learn. Health*, 2021, pp. 156–167.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[14] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[15] S. Cao, P. Xu, and D. A. Clifton, "How to understand masked autoencoders," 2022, *arXiv:2202.03670*.

[16] X. Chen et al., "Deep autoencoder imaging method for electrical impedance tomography," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[17] B. Hou, J. Yang, P. Wang, and R. Yan, "LSTM-based auto-encoder model for ECG arrhythmias classification," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1232–1240, Apr. 2020.

[18] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[19] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image analysis," 2022, *arXiv:2203.05573*.

[20] Y. Tian et al., "Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder," 2022, *arXiv:2203.11725*.

[21] Y. Luo, Z. Chen, and X. Gao, "Self-distillation augmented masked autoencoders for histopathological image classification," 2022, *arXiv:2203.16983*.

[22] S. Zhang, H. Chen, H. Yang, X. Sun, P. S. Yu, and G. Xu, "Graph masked autoencoders with transformers," 2022, *arXiv:2202.08391*.

[23] A. Khajenezhad, S. A. Osia, M. Karimian, and H. Beigy, "Gransformer: Transformer-based graph generation," 2022, *arXiv:2203.13655*.

[24] Q. Tan, N. Liu, X. Huang, R. Chen, S.-H. Choi, and X. Hu, "MGAE: Masked autoencoders for self-supervised learning on graphs," 2022, *arXiv:2201.02534*.

[25] H. Liu, M. Cai, and Y. J. Lee, "Masked discrimination for self-supervised learning on point clouds," 2022, *arXiv:2203.11183*.

[26] Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," 2022, *arXiv:2203.06604*.

[27] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," 2022, *arXiv:2203.12602*.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

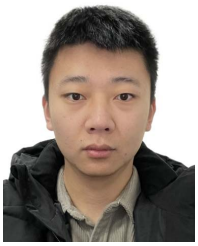
[29] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[31] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.
[32] T. Mehari and N. Strodthoff, "Self-supervised representation learning from 12-lead ECG data," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105114.
[33] H. Liu, Z. Zhao, and Q. She, "Self-supervised ECG pre-training," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103010.
[34] D. Kiyasseh, T. Zhu, and D. A. Clifton, "Clocs: Contrastive learning of cardiac signals across space, time, and patients," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5606–5615.
[35] D. Gedon, A. H. Ribeiro, N. Wahlstrom, and T. B. Schon, "First steps towards self-supervised pretraining of the 12-lead ECG," in *Proc. Comput. Cardiol. (CinC)*, Sep. 2021, pp. 1–4.
[36] F. Liu et al., "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *J. Med. Imag. Health Informat.*, vol. 8, no. 7, pp. 1368–1373, Jul. 2018.
[37] H. Martin, U. Morar, W. Izquierdo, M. Cabrerizo, A. Cabrera, and M. Adjouadi, "Real-time frequency-independent single-lead and single-beat myocardial infarction detection," *Artif. Intell. Med.*, vol. 121, Nov. 2021, Art. no. 102179.
[38] Z. He, Z. Yuan, P. An, J. Zhao, and B. Du, "MFB-LANN: A lightweight and updatable myocardial infarction diagnosis system based on convolutional neural networks and active learning," *Comput. Methods Programs Biomed.*, vol. 210, Oct. 2021, Art. no. 106379.
[39] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1519–1528, May 2021.
[40] J. Niu, Y. Tang, Z. Sun, and W. Zhang, "Inter-patient ECG classification with symbolic representations and multi-perspective convolutional neural networks," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1321–1332, May 2020.
[41] J. Zheng et al., "Optimal multi-stage arrhythmia classification approach," *Sci. Rep.*, vol. 10, no. 1, pp. 1–17, Feb. 2020.
[42] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform," *Biomed. Signal Process. Control*, vol. 8, no. 5, pp. 437–448, Sep. 2013.
[43] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[44] J. Hampton and J. Hampton, *The ECG Made Easy E-Book*. Amsterdam, The Netherlands: Elsevier, 2019.
[45] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3320–3328.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[47] M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," presented at the 29th Int. Conf. Mach. Learn., Edinburgh, Scotland, 2012.
[48] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16519–16529.
[49] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
[50] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are convolutional neural networks or transformers more like human vision," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 43, 2021, pp. 1–7.

**Huaicheng Zhang** was born in Fujian, China. He received the bachelor's degrees in microelectronics and solid-state electronics from Wuhan University, Wuhan, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Physics and Technology.

His current research interests mainly include deep learning, machine learning, and artificial intelligence for medical diagnosis.

**Wenhan Liu** received the bachelor's and master's degrees from Wuhan University, Wuhan, China, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree with the School of Physics and Technology.

From 2019 to 2021, he was an Algorithm Engineer with Alibaba Inc., Hangzhou, China. He has authored or coauthored ten papers on intelligent electrocardiogram analysis. His current research interests include biosignal processing, deep learning, and machine learning.



**Jiguang Shi** received the bachelor's degree in microelectronics and solid-state electronics from Wuhan University, Wuhan, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Physics and Technology.

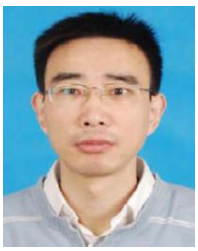
His current research interests mainly include machine learning, deep learning, and AI for medical diagnosis.

**Sheng Chang** (Senior Member, IEEE) received the B.S. degree in applied physics and the M.S. and Ph.D. degrees in microelectronics and solid-state electronics from Wuhan University, Wuhan, China, in 2002, 2004, and 2009, respectively.

He is currently a Professor with the School of Physics and Technology, Wuhan University. His research interests mainly include machine learning and digital circuit.

**Hao Wang** (Member, IEEE) was born in Henan, China, in 1983. He received the B.S. degree in electronic engineering and the M.S. and Ph.D. degrees in microelectronics and solid-state electronics from Wuhan University, Wuhan, China, in 2003, 2006, and 2009, respectively.

He is currently an Associate Professor with Wuhan University. His research interests mainly include the modeling and simulation of novel semiconductor devices.



**Jin He** (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University, Singapore, in 2011.

He has been working on analog/RF/mm-Wave in both academia and industry from 2003 to 2013 in China and Singapore, respectively, holding positions of an Engineer, a Research Associate, a Senior Research Engineer, a Scientist, and a Project Leader. He joined Wuhan University, Wuhan, China in 2013, where he is currently an Associate Professor with the School of Physics and Technology. His current research interests include analog/RF/mm-Wave/THz integrated circuit design in CMOS/BiCMOS/SiGe for optical and wireless applications.

Dr. He has served as a Technical Reviewer for the IEEE TRANSACTION ON MICROWAVE AND THEORY TECHNIQUES, the IEEE TRANSACTION ON ELECTRON DEVICES, the IEEE MICROWAVE AND WIRELESS COMPONENTS LETTERS, the *Electronics and Telecommunications Research Institute* (ETRL) *Journal*, and the *Journal of Electronic Science and Technology*.

**Qijun Huang** was born in Guangxi, China, in 1965. He received the B.S. degree in semiconductor physics and the Ph.D. degree in microelectronics and solid-state electronics from Wuhan University, Wuhan, China, in 1986 and 2010, respectively.

He is currently a Professor with the School of Physics and Technology, Wuhan University. He has published more than 240 papers, including more than 70 SCI-indexed papers. His current research focuses on biosignal processing, machine learning and its hardware implementation, simulation, and design of microelectronic devices and circuits.