Masked Transformer for Electrocardiogram Classification

Ya Zhou[1,#,*], Xiaolin Diao[1,#], Yanni Huo[1], Yang Liu[1], Xiaohan Fan[2,*] and Wei Zhao[3,*,†]

[1]*Department of Information Center, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100037, China*
[2]*Cardiac Arrhythmia Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, 100037, China*
[3]*Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100037, China*

Abstract

Electrocardiogram (ECG) is one of the most important diagnostic tools in clinical applications. With the advent of advanced algorithms, various deep learning models have been adopted for ECG tasks. However, the potential of Transformer for ECG data has not been fully realized, despite their widespread success in computer vision and natural language processing. In this work, we present Masked Transformer for ECG classification (MTECG), a simple yet effective method which significantly outperforms recent state-of-the-art algorithms in ECG classification. Our approach adapts the image-based masked autoencoders to self-supervised representation learning from ECG time series. We utilize a lightweight Transformer for the encoder and a 1-layer Transformer for the decoder. The ECG signal is split into a sequence of non-overlapping segments along the time dimension, and learnable positional embeddings are added to preserve the sequential information. We construct the Fuwai dataset comprising 220,251 ECG recordings with a broad range of diagnoses, annotated by medical experts, to explore the potential of Transformer. A strong pre-training and fine-tuning recipe is proposed from the empirical study. The experiments demonstrate that the proposed method increases the macro F1 scores by 3.4%-27.5% on the Fuwai dataset, 9.9%-32.0% on the PTB-XL dataset, and 9.4%-39.1% on a multicenter dataset, compared to the alternative methods. We hope that this study could direct future research on the application of Transformer to more ECG tasks.

Keywords: Electrocardiogram, Transformer, Masked autoencoders, Self-supervised learning, Classification

1 Introduction

The electrocardiogram (ECG) is one of the most commonly used non-invasive tools in clinical applications, recording the cardiac electrical activity and providing valuable diagnostic clues

[#]The authors contributed equally.
[*]Corresponding authors: Ya Zhou (zhouya@fuwai.com), Xiaohan Fan (fanxiaohan@fuwaihospital.org), Wei Zhao (zw@fuwai.com)
[†]Supervision: Wei Zhao

for systemic conditions (Siontis et al., 2021). With the algorithmic advances in computer vision (CV) several years ago, the convolutional neural network (CNN) has been adapted to solve various ECG classification tasks, such as detecting arrhythmia (Hannun et al., 2019; Ribeiro et al., 2020; Hughes et al., 2021), left ventricular systolic dysfunction (Attia et al., 2019, 2022; Yao et al., 2021) and hypertrophic cardiomyopathy (Ko et al., 2020).

The field of CV has been transformed recently by the excellent performance of Transformer models (Vaswani et al., 2017; Dosovitskiy et al., 2020). However, the progress of Transformer methods for ECG tasks lags behind those in CV. CNN remain the most common architecture for ECG data analysis (Siontis et al., 2021; Somani et al., 2021; Wu et al., 2021; Pourbabaee et al., 2017). A natural question is that can Transformer also be leveraged for ECG tasks. We attempt to answer the question in this paper.

One challenge is to adapt Transformer to effectively model ECG data. Commonly used ECG data are time-series signals with a unique temporal and spatial structure. For example, the standard 10 seconds 12-lead ECG signals with 500Hz sampling rate have 5000 time points. These signals typically consist of several beats with multiple waves, such as the P wave, QRS complex, T wave, and sometimes U wave (Sharma and Sunkaria, 2021). Directly modeling each time step individually in Transformer might ruin the structural information and lead to heavy computation. One potential solution is to split the signal into a sequence of non-overlapping segments. By treating these segments as tokens or image patches, one can readily utilize the design of Vision Transformer (ViT, Dosovitskiy et al., 2020). However, previous studies have demonstrated that the straightforward approach yields suboptimal classification accuracy (Li et al., 2021). We believe that this is partly due to inappropriate training strategies. In this work, we will show this straightforward approach is capable of achieving excellent performance if appropriate training strategies are adopted.

Another challenge is the lack of available labeled datasets. It is known that the successful training of Transformer requires a large-scale dataset with labeled data, e.g., ImageNet-21k (Deng et al., 2009), due to the lack of inductive bias (Dosovitskiy et al., 2020). However, creating labeled ECG datasets at scale is challenging since accurate annotation requires medical experts and considerable time. Although recent advancements have introduced larger ECG datasets like PTB-XL (Wagner et al., 2020) and Chapman-Shaoxing (Zheng et al., 2020b), they still contain only a few tens of thousands of data points, making the Transformer model prone to overfitting (Li et al., 2021). To address this limitation, we build the Fuwai dataset comprising 220,251 ECG recordings with a broad range of ECG diagnoses annotated by medical experts.

Although the Fuwai dataset is much larger than publicly available ECG datasets, our experiments show that naively training a Transformer on it yields unsatisfactory results. We still observe overfitting, even when using a lightweight Transformer with 5.7 million parameters. This problem might be caused by the information redundancy (Wei et al., 2019), which typically exists in ECG time series. For example, each time point can be inferred from neighboring points, and some heartbeat cycles could be inferred from neighboring heartbeat cycles.

In the field of natural images, the information redundancy has been effectively addressed by masked modeling methods such as masked autoencoders (MAE, He et al., 2022) and SimMIM (Xie et al., 2022).

Inspired by their success, we propose a masked Transformer method for ECG classification, referred as MTECG. This method is a simple extension of the image-based MAE to

self-supervised representation learning from ECG time series. We split the ECG signal into a sequence of non-overlapping segments along the time dimension and adopt learnable positional embeddings to preserve the sequential order information and distinguish the segments. To reduce the information redundancy, we utilize the lightweight Transformer with masked pre-training on unlabeled ECG data. To encourage the encoder learning useful wave shape features, we adopt a 1-layer Transformer as the decoder and a fluctuated reconstruction target in the pre-training stage. To further reduce overfitting, we explore the regularization strategies such as layer-wise LR decay (Bao et al., 2021; Clark et al., 2020) and DropPath (Huang et al., 2016) rate in the fine-tuning stage.

The main contributions of this work are summarized as follows.

1. **Novel Application**: We propose MTECG, a useful masked Transformer method for ECG time series, as the extension of MAE originally proposed for the natural image analysis (He et al., 2022). The derived lightweight model demonstrates the ability to classify a wide range of ECG diagnoses effectively, while remaining convenient for deployment in the clinical environment.

2. **New Dataset**: We build a comprehensive ECG dataset to evaluate various deep learning algorithms. The dataset consists of 220,251 recordings with 28 common ECG diagnoses annotated by medical experts and significantly surpasses the sample size of publicly available ECG datasets.

3. **Strong Pre-training and Fine-tuning Recipe**: We conduct comprehensive experiments to explore the training strategies on the proposed ECG dataset. The key components contributing to the proposed method are presented, including the masking ratio, training schedule length, fluctuated reconstruction target, layer-wise LR decay and DropPath rate.

4. **Excellent Performance**: We compare the proposed method with recent state-of-the-art algorithms on a wide variety of tasks in both private and public ECG datasets. The experiments demonstrate that the proposed algorithm outperforms others significantly. Notably, it increases the macro F1 scores by 3.4%-27.5% on the Fuwai dataset, 9.9%-32.0% on the PTB-XL dataset, and 9.4%-39.1% on a multicenter dataset, compared to other methods.

The rest of this article is organized as follows. Section 2 briefly outlines the related works. Section 3 introduces the proposed method. The experiment setting, the properties and the practical performance of the proposed method can be founded in Section 4. We provide the additional discussion in Section 5. The conclusion of this work is summarized in Section 6.

2 Related work

2.1 ECG classification

ECG classification is one of the most important ECG data analysis tasks (Pyakillya et al., 2017; Somani et al., 2021; Siontis et al., 2021). Traditionally, the classification for ECG signals is mainly based on expert features and conventional machine learning methods (Hong et al., 2020). However, the algorithm performance is limited by data quality and expert knowledge (Hong et al., 2020), precluding the usage in clinical applications (Ribeiro et al., 2020). To overcome these limitations, numerous works have adapted end-to-end deep learning approaches to ECG tasks (Hannun et al., 2019). In comparison to the traditional methods, the deep learning approach can automatically learn useful features from raw ECG signals and achieve better performance in many classification tasks (Strodthoff et al., 2023).

2.2 Transformer

Transformer is an attention-based architecture which has rapidly become the dominant choice in natural language processing (NLP) since it was proposed (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2018). Due to its excellent performance, the field of computer vision (CV) has also been recently transformed by Transformer (Han et al., 2022; Khan et al., 2022). Compared to CNN, Transformer can better utilize the global structure of the data and share shape recognition capabilities similar to those of the human visual system (Naseer et al., 2021). However, CNN still remains the most common architecture for ECG data analysis (Somani et al., 2021). Although some works have explored Transformer on ECG data analysis, such as BaT (Li et al., 2021), MaeFE (Zhang et al., 2022) and CRT (Zhang et al., 2023), a comprehensive evaluation remains lacking and the potential of Transformer in this area is not fully realized. We will compare the performance of the proposed method with BaT, MaeFE and CRT.

2.3 Self-supervised learning

Self-supervised learning is a popular method to utilize the unlabeled data and enhance the algorithm performance (Liu et al., 2021). Contrastive and generative learning are two main self-supervised learning approaches (Krishnan et al., 2022).

Contrastive learning aims to model sample similarity and dissimilarity through a contrastive estimation objective (Gutmann and Hyvärinen, 2010; Liu et al., 2021), and its performance heavily relies on the data augmentation strategy (He et al., 2022; Wang et al., 2023b). Various contrastive learning methods have been proposed for ECG data analysis, notable examples including CLECG (Chen et al., 2021) and CLOCS (Kiyasseh et al., 2021). A previous study suggests that CLECG generally outperforms CLOCS across multiple datasets (Chen et al., 2021). In this paper, we compare CLECG with our method.

Generative learning, such as GPT (Radford et al., 2018), BERT(Devlin et al., 2019), BEiT (Bao et al., 2021) and MAE (He et al., 2022), is another main self-supervised learning method. Although steadily gaining attention in NLP and CV, its effectiveness for ECG analysis tasks has yet to show significant performance improvements. Recent studies have explored the use of masked modeling techniques on ECG data, introducing models such as MaeFE (Zhang et al., 2022) and CRT (Zhang et al., 2023), which have achieved superior results in ECG classification compared to other approaches. Nonetheless, the full potential of masked modeling for ECG classification remains untapped. In this paper, we present a novel approach that significantly outperforms the state-of-the-art in this domain.

3 Methodology

Our approach expands image MAE (He et al., 2022) to ECG time series and follows the pretrain-and-finetune paradigm. The framework consists of 5 major components, i.e., segment operation, mask operation, encoder, decoder, and reconstruction target. As shown in Figure 1, the pre-training stage involves all components. As for fine-tuning, the decoder, reconstruction target and mask operation will be discarded. In the subsequent subsections, we detail each component and the training strategy.

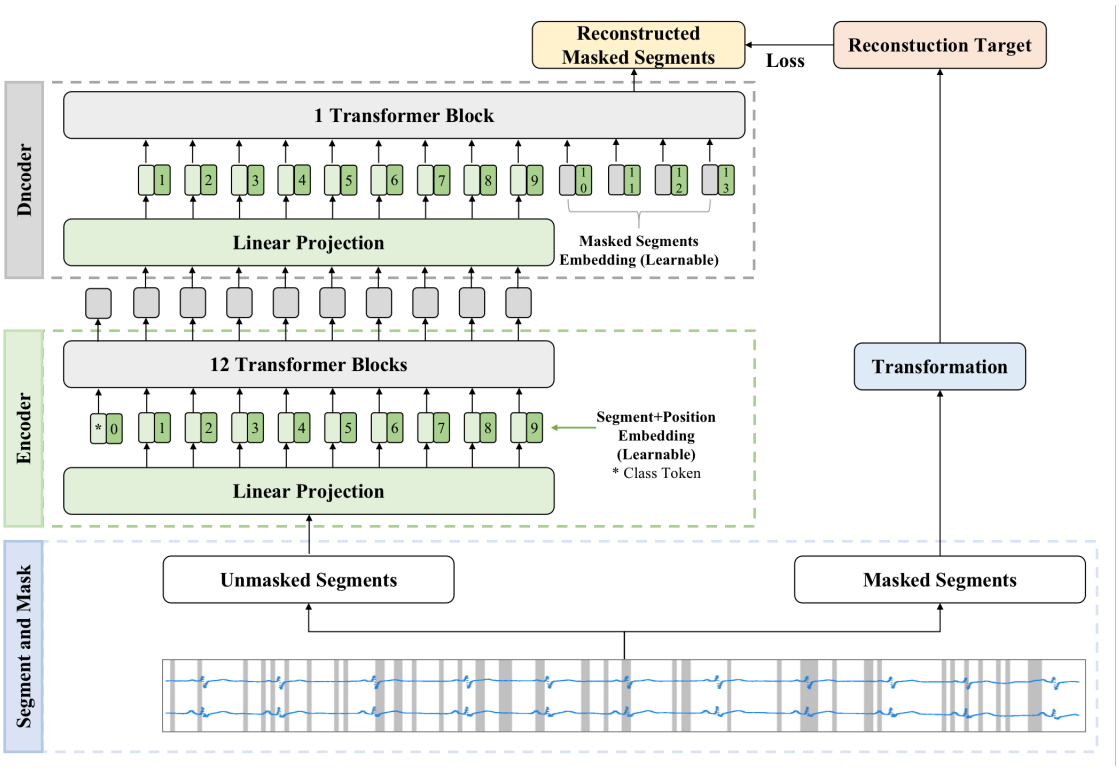

Figure 1: An example of the masked pre-training method. The original ECG signals are split to non-overlapping segments and a subset of these segments is masked out. The unmasked segments were used to reconstruct the fluctuated transformation of the masked segments through Transformer blocks. The lead and sequence information of the ECG signals are preserved by learnable positional embeddings. The masked segments are represented by learnable embeddings in the reconstruction task.

3.1 Segment operation

For simplicity, this paper focuses on fixed-length ECG data. This type of ECG signal can be denoted as $\mathbf{X}^\sharp = (X^\sharp_{k,q}) \in \mathbb{R}^{K\times Q}$, where K is the number of leads and Q is the length of ECG. In many applications, Q is much larger than K. For example, in the case of standard 12-lead ECGs, with a duration of 10 seconds and a sampling frequency of 500 Hz, we have $Q = 5000$ and $K = 12$. When dealing with ECG signals of this length, directly modeling each time step individually using a Transformer would lead to heavy computation due to the quadratic cost of the self-attention operation. Moreover, ECG signals comprise multiple beats, each of which typically consists of various components such as the P wave, QRS complex, T wave, and U wave. If we were to model the signals at the individual time point level, we might ignore the wave shape information.

Alternatively, we split the ECG signal $\mathbf{X}^\sharp$ into a sequence of non-overlapping segments $\mathcal{S} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$, where T is the sequence length. The structural information of waves could be preserved in the segments and the sequence length could be remarkably reduced. Each segment $\mathbf{X}_t = (X^t_{k,q}) \in \mathbb{R}^{K\times(Q/T)}$ represents a subset of the original signal, where $\mathbf{X}_t = (X^t_{k,q}) \in \mathbb{R}^{K\times(Q/T)}$ and $X^t_{k,q} = X^\sharp_{k,\{(t-1)*Q/T+q\}}$ for $t = 1, \ldots, T$, $k = 1, \ldots, K, q =$

$1, \ldots, Q/T$. These ECG segments can be regarded as image patches in computer vision, or tokens (words) in natural language processing. Similar to image MAE (He et al., 2022), we utilize a linear projection to resize the vectorization of the ECG segments. However, unlike their work, we incorporate learnable positional embeddings to preserve the sequential order information and distinguish the segments, as shown in (1) and (2).

3.2 Mask operation

As mentioned in Section 1, we employ a masked pre-training method to address information redundancy and overfitting. Specifically, we employ a uniform random sampling technique to select segments from $\mathcal{S}$ without replacement, while the remaining segments are masked. For notational simplicity, we denote the unmasked segments as

$$\mathcal{S}_{unmask} = \{\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_S}\},$$

and the masked segments as

$$\mathcal{S}_{mask} = \{\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_{S'}}\},$$

Here, i_s and $j_{s'}$ are chosen from the set $\{1, \ldots, T\}$, with $s = 1, \ldots, S$ and $s' = 1, \ldots, S'$. The segments in $\mathcal{S}_{unmask}$ are sent to the encoder, while those in $\mathcal{S}_{mask}$ will be transformed into the reconstruction targets. Although the case $T = S + S'$ is the main focus of this paper, the entire framework is also applicable to the case where $T < S + S'$.

3.3 Encoder

In our settings, we can easily utilize the design of ViT (Dosovitskiy et al., 2020). We use a stack of standard Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2020) as the encoder. For the natural image, MAE (He et al., 2022) incorporates vanilla positional embeddings (the sine-cosine version) to extract location information. However, previous studies demonstrate that this vanilla positional approach might be unable to fully exploit the important features of times series data (Wen et al., 2022). Therefore, we utilize learnable positional embeddings to effectively model the lead and sequence information in the ECG time series.

We briefly introduce the encoder below. Denote the layer normalization (Ba et al., 2016), the multi-headed self-attention and multi-linear perception blocks used in ViT (Dosovitskiy et al., 2020) as $LN(\cdot)$, $MSA(\cdot)$ and $MLP(\cdot)$, respectively. We use $\mathbf{x}_{i_s}^\intercal \in \mathbb{R}^{KQ/T}$ to denote the vectorization of $\mathbf{X}_{i_s}$ for $s = 1, \ldots, S$. Suppose D is the latent vector size. Let the linear projection matrix $\mathbf{E} \in \mathbb{R}^{(KQ/T)\times D}$, the auxiliary token $\mathbf{x}_{class}^\intercal \in \mathbb{R}^D$, and the positional embedding $\mathbf{e}_{pos} = (\mathbf{e}_0, \mathbf{e}_1, \cdots, \mathbf{e}_T)^\intercal \in \mathbb{R}^{D(T+1)}$ be learnable. Here $\mathbf{e}_t^\intercal \in \mathbb{R}^D, t = 0, 1, \ldots, T$ are used to preserve the sequential order information and distinguish the segments. During pre-training, only the unmasked segments are fed to the encoder, which can be written as

$$\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{class}; \mathbf{x}_{i_1}\mathbf{E}, \cdots, \mathbf{x}_{i_S}\mathbf{E}] + [\mathbf{e}_0; \mathbf{e}_{i_1}; \ldots; \mathbf{e}_{i_S}], \\
\mathbf{z}'_l &= MSA(LN(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, l = 1, \ldots, L, \\
\mathbf{z}_l &= MLP(LN(\mathbf{z}'_l)) + \mathbf{z}'_l, l = 1, \ldots, L,
\end{aligned} \tag{1}$$

where L is the number of blocks. For simplicity, we denote the output $\mathbf{z}_L$ as

$$\mathbf{z}_L = [\mathbf{z}_L^0; \mathbf{z}_L^{i_1}, \cdots; \mathbf{z}_L^{i_S}],$$

where the encoded auxiliary token $(\mathbf{z}_L^0)^\intercal \in \mathbb{R}^D$ is discarded and the encoded unmasked segments $(\mathbf{z}_L^{i_s})^\intercal \in \mathbb{R}^D, s = 1, \ldots, S$ are used for the reconstruction targets in pre-training. In contrast, during fine-tuning, the encoder is applied to all segments and both the encoded auxiliary token and segments could be used for the downstream task.

3.4 Decoder

To encourage the encoder learning useful wave shape features and reduce the computation cost, we use a 1-layer Transformer as the decoder. Suppose D' is the latent vector size. Let $\mathbf{E}' \in \mathbb{R}^{D \times D'}$, $\mathbf{E}_0 \in \mathbb{R}^{D' \times (KQ/T)}$, $\mathbf{e}_m^\intercal \in \mathbb{R}^{D'}$ and $\mathbf{e}'_{pos} = (\mathbf{e}'_1, \cdots, \mathbf{e}'_T)^\intercal \in \mathbb{R}^{TD'}$ be learnable components. Here $(\mathbf{e}'_t)^\intercal \in \mathbb{R}^{D'}$ for $t = 1, \ldots, T$ serve as the positional embeddings and $\mathbf{e}_m$ is used to denote the masked segments. The decoder can be written as

$$\begin{aligned} \tilde{\mathbf{z}}_0 &= [\mathbf{z}_L^{i_1}\mathbf{E}'; \cdots; \mathbf{z}_L^{i_S}\mathbf{E}'; \mathbf{e}_m; \cdots; \mathbf{e}_m] \\ &\quad + [\mathbf{e}'_{i_1}; \ldots; \mathbf{e}'_{i_S}; \mathbf{e}'_{j_1}; \cdots; \mathbf{e}'_{j_{S'}}] \\ \tilde{\mathbf{z}}'_1 &= MSA(LN(\tilde{\mathbf{z}}_0)) + \tilde{\mathbf{z}}_0, \\ \tilde{\mathbf{z}}_1 &= MLP(LN(\mathbf{z}'_1)) + \mathbf{z}'_1, \end{aligned} \tag{2}$$

where $\tilde{\mathbf{z}}'_1$ can be written as $\tilde{\mathbf{z}}'_1 = [\tilde{\mathbf{z}}_1^{i_1}; \cdots; \tilde{\mathbf{z}}_1^{i_S}; \tilde{\mathbf{z}}_1^{j_1}; \cdots; \tilde{\mathbf{z}}_1^{j_{S'}}]$. Here $(\tilde{\mathbf{z}}_1^{i_s})^\intercal, (\tilde{\mathbf{z}}_1^{j_{s'}})^\intercal \in \mathbb{R}^{D'}$ for $s = 1, \ldots, S$, $s' = 1, \ldots, S'$ and $\tilde{\mathbf{z}}_1^{j_{s'}}$ will be used to generate the output $\tilde{\mathbf{x}}_{j_{s'}}$ though

$$\tilde{\mathbf{x}}_{j_{s'}} = \tilde{\mathbf{z}}_1^{j_{s'}} \mathbf{E}_0, s' = 1, \ldots, S'. \tag{3}$$

3.5 Reconstruction target

The art of masked modeling lies in the choice of pretext tasks (Liu et al., 2021). Due to the cyclic nature of the ECG signals, directly reconstructing the original signal might weaken the learning of wave shape features. To encourage the encoder learning useful features, we adopt a fluctuated reconstruction target in pre-training. To be more specific, we denote $f : \mathbb{R}^{QW/T} \to \mathbb{R}^{QW/T}$ as a pre-defined fluctuated mapping. We utilize the following mean square error (MSE) as the pre-training loss function

$$MSE = \sum_{s'=1}^{S'} \|\tilde{\mathbf{x}}_{j_{s'}} - f(\mathbf{x}_{j'_s})\|_2^2.$$

Here $\tilde{\mathbf{x}}_{j_{s'}}$ is defined in (3), $\mathbf{x}_{j'_s} = (x_{j'_s,j})_{j=1}^{KQ/T}$ is the vectorization of $\mathbf{X}_{j_{s'}}$, where $\mathbf{X}_{j_{s'}} \in \mathcal{S}_{mask}$ is a masked segment. In this paper, we explore two options for the fluctuated mapping f. The first one is per-segment normalization and the j-th element of $f(\mathbf{x}_{j'_s})$ is defined as

$$f(\mathbf{x}_{j'_s})_j = \frac{x_{j'_s,j} - \mu_{j'_s}}{\sqrt{\sigma_{j'_s}^2 + \epsilon}}, \tag{4}$$

where

$$\mu_{j'_s} = \frac{T}{KQ} \sum_{j=1}^{KQ/T} x_{j'_s,j}, \quad \sigma_{j'_s}^2 = \frac{T}{KQ} \sum_{j=1}^{KQ/T} (x_{j'_s,j} - \mu_{j'_s}),$$

and ϵ is a constant used for the numerical stability. The second choice is just a a simple squaring operation. We define the j-th element of $f(\mathbf{x}_{j'_s})$ as

$$f(\mathbf{x}_{j'_s})_j = sign(x_{j'_s,j})|x_{j'_s,j}|^{0.5}. \tag{5}$$

Both of these options have been demonstrated to enhance local contrast in time-series ECG data. The per-segment normalization (4) is the counterpart of the per-patch normalization employed in MAE (He et al., 2022), whereas the squaring operation (5) is a simple and effective approach proposed in this study.

3.6 Training strategy

Our approach follows the the pretrain-and-finetune paradigm. Firstly, we pre-train the model to learn a useful ECG representation through masked modeling, as illustrated in Figure 1. Secondly, we follow MAE (He et al., 2022) and fine-tune the pre-trained encoder with an added classification head for the downstream tasks. Since ECG signals have a distinct nature compared to images and language, directly applying the training recipe of images or language might lead to poor results. We investigate several components such as masking ratio and reconstruction targets to effectively train the model. Moreover, considering the heavier redundancy in ECG signals, we also dive into the regularization hyper-parameters, such as layer-wise LR decay (Bao et al., 2021; Clark et al., 2020) and DropPath (Huang et al., 2016) rate, during fine-tuning to further reduce overfitting.

4 Experiments

In this section, we explore the property of the proposed method based on the Fuwai dataset consisting 220,251 ECG recordings with a broad range of diagnoses annotated by medical experts. Additionally, we evaluate the prediction performance of the proposed method on both private and public ECG datasets.

4.1 Datasets

The **Fuwai dataset** consists of 220,251 ECG recordings from 173,951 adult patients in Fuwai Hospital of Chinese Academy of Medical Sciences. This dataset encompasses 28 different diagnoses, and the diagnoses for each ECG recording were annotated by two certified ECG physicians. The dataset was split into three sets based on patients, with a ratio of 8:1:1 for the training, validation, and testing sets, respectively. This study was approved by the Ethics Committee at Fuwai Hospital, and no identifiable personal data of patients were used.

The **PCinC dataset** was collected from the PhysioNet/Computing in Cardiology Challenge 2021 (Reyna et al., 2021). The public portion of the challenge datasets contains 88,253 12-lead ECG recordings gathered from 8 datasets, including PTB (Bousseljot et al., 1995), PTB-XL (Wagner et al., 2020), Chapman-Shaoxing (Zheng et al., 2020b), Ningbo (Zheng et al., 2020a), CPSC (Liu et al., 2018), CPSC-Extra (Liu et al., 2018), INCART (Tihonenko et al., 2007), and G12EC(Reyna et al., 2021). The diagnoses of these ECG datasets were encoded to 133 diagnoses using approximate Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) codes, with 30 diagnoses were selected as being of clinical interest

(Reyna et al., 2022). Following the competition, we merged 8 diagnoses into 4 labels. We then consider subjects with complete 12-lead signals, sampled at a rate of 500 Hz over a duration of 10 seconds, resulting in a total of 79,574 ECG recordings. Furthermore, we refine the dataset by selecting only labels with an incidence rate of at least 0.5%, further reducing the number of labels to 25. To evaluate the algorithms, we randomly divided the entire dataset into training, validation, and testing sets in a ratio of 7:1:2, respectively.

The **PTB-XL dataset** is a commonly used dataset to evaluation ECG classification algorithms (Strodthoff et al., 2020). It is a part of the aforementioned challenge datasets and includes 21,837 ECG recordings. Consistently, we utilize SNOMED-CT labels with an incidence of at least 0.5% and remove the sample without complete 12-lead signals, resulting in the consideration of 21,836 recordings with 22 labels. We follow the suggested 10-fold split by Strodthoff et al. (2020), where folds 1-8, 9 and 10 are training, validation, testing sets, respectively.

4.2 Metrics

Our clinical objective is to detect the majority of positive cases while minimizing the occurrence of false alarms. In binary classification, the F1 score is a commonly used performance metric that takes this trade-off into account. It is calculated as the harmonic mean of sensitivity (recall) and precision. When dealing with multi-label classification in imbalanced data, such as our ECG datasets, the macro metrics can provide a better assessment of the overall classification performance (Bagui and Li, 2021). Therefore, we use the macro F1 score as the major evaluation metric throughout this paper.

4.3 Default setting

We adopt MTECG-T as the default backbone, a lightweight Transformer architecture, which will be introduced in Section 4.5. Following Li et al. (2021), we use 25 as the default segment size, which yields the sequence length $T = 200$ in our experiments. In addition, we use a mask ratio of 25%, a per-segment normalization reconstruction target, a light weight decoder with a latent dimension $D' = 128$ alongside a self-attention head of 4, and a global pooling classifier as the default setting.

Initially, we pre-train the model on the training set. During pre-training, we utilize an AdamW optimizer with a cosine learning rate schedule. The default hyperparameters include a momentum of $(\beta_1, \beta_2) = (0.9, 0.95)$, a weight decay of 0.05, a learning rate of 0.001, a batch size of 256, and a total of 1600 training epochs with a warm-up epoch of 40.

After pre-training, we fine-tune the pre-trained encoder on the same dataset. For fine-tuning, we also use the AdamW optimizer and the cosine learning rate schedule. The default hyperparameters includes a momentum of $(\beta_1, \beta_2) = (0.9, 0.999)$, a weight decay of 0.05, a learning rate of 0.001, a batch size of 256, a DropPath rate of 0.4, a layer decay rate of 0.6, and a warm-up epoch of 5. We train the model for 50 epochs using the binary cross entropy loss and select the one corresponding to the highest macro F1 scores on the validation set as our final model.

For comparison, we also train the model from scratch for 200 epochs using the same recipe as the fine-tuning process. Figure 2 depicts the comparison result, from which we can see the masked pre-training significantly improve the classification performance.

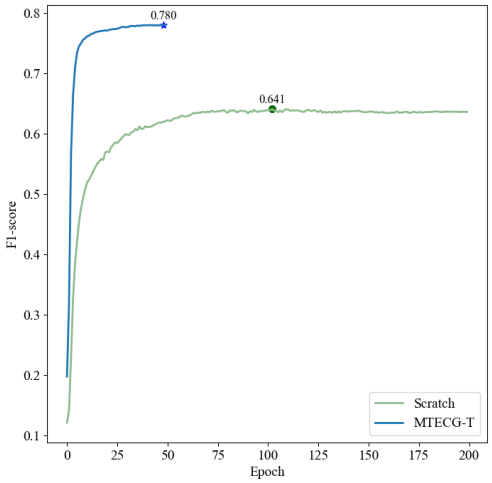

Figure 2: Performance comparison between masked pre-training and training from scratch on the Fuwai dataset. The optimal epoch for masked pre-training is found to be 48, whereas for training from scratch, it is 102.

4.4 Ablation study

In the ablation study, we explore the properties of important components for the proposed method on the Fuwai dataset, and report the marco F1 score on the validation set.

4.4.1 Masking ratio

The validation marco F1 scores of the proposed MTECG-T under multiple masking ratios are summarized in Figure 3 (A). The optimal masking ratio of the proposed method is 25% and it performs stably well across a broad range of ratios (5%-75%). Interestingly, we observed that even at the extreme ratios of 1% and 99%, our proposed method significantly outperforms the model trained from scratch. These findings highlight the effectiveness of the masked pre-training.

4.4.2 Training schedule

We investigate the influence of the pre-training schedule length on the fine-tuning performance. Figure 3 (B) presents our findings. The performance improves with longer training and begins to saturate at 1000 epochs. Interestingly, even with shorter training schedules that had not yet reached saturation, such as the 100-epoch pre-training case, the proposed method exhibits significantly superior performance compared to the model trained from scratch.

4.4.3 Reconstruction target

In this subsection, we compare different reconstruction targets. Figure 4 presents the reconstruction examples from one patient in the validation set and the corresponding fine-tuning

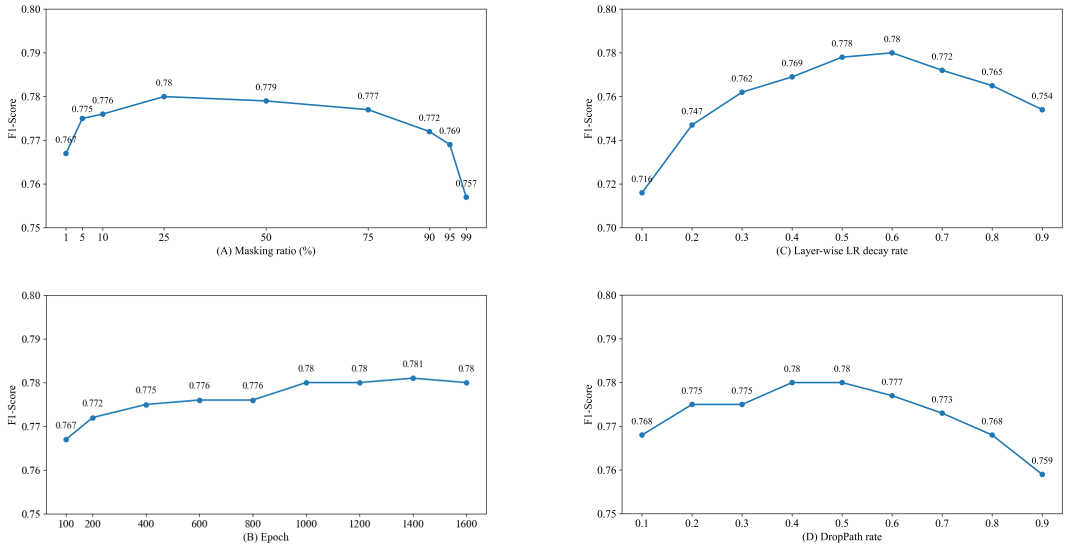

Figure 3: Classification performance in the ablation study. The first column, from top to bottom, corresponds to masking ratio and pre-training schedule lengths, respectively. Similarly, the second column, from top to bottom, corresponds to layer-wise LR decay and DropPath rate, respectively.

performance. Our experimental results clearly demonstrate the beneficial impact of two transformation operations, i.e., per-segment normalization and squaring operation. We observe that both operations fluctuate the original signals, which might enhance the local contrast and amplify subtle changes.

4.4.4 Fine-tuning strategy

We investigate the effects of regularization hyper-parameters, layer-wise LR decay (Bao et al., 2021; Clark et al., 2020) and DropPath (Huang et al., 2016) rate on the fine-tuning performance. The results are presented in Figure 3 (C and D), which show the performance across a range of hyper-parameter values from 0.1 to 0.9. The layer-wise LR decay rate of 0.6 is identified as the optimal setting, while for the DropPath rate, options of 0.4 and 0.5 are considered. However, for simplicity, a DropPath rate of 0.4 is utilized later on. Interestingly, we observe that the performance is slightly more sensitive to these hyper-parameters than to the masking ratio.

Furthermore, we conducted experiments on the encoded auxiliary token and segments for classification and found that their performances are comparable. Specifically, when using all encoded segments for classification through global pooling, the macro validation F1 score is 0.780, while using the auxiliary token directly yields a macro F1 score of 0.779.

4.5 Scaling experiments

Noting that the model capacity matters to the generalization ability, we adopt Transformer of different model size for experiments, including MTECG-A (Atomic), MTECG-M (Molecular), MTECG-T(Tiny), MTECG-S (Small) and MTECG-B (Base), for which we change the number of heads h for the encoder, keeping $D/h = 64$ and the layers $L = 12$. Table 1 summarizes

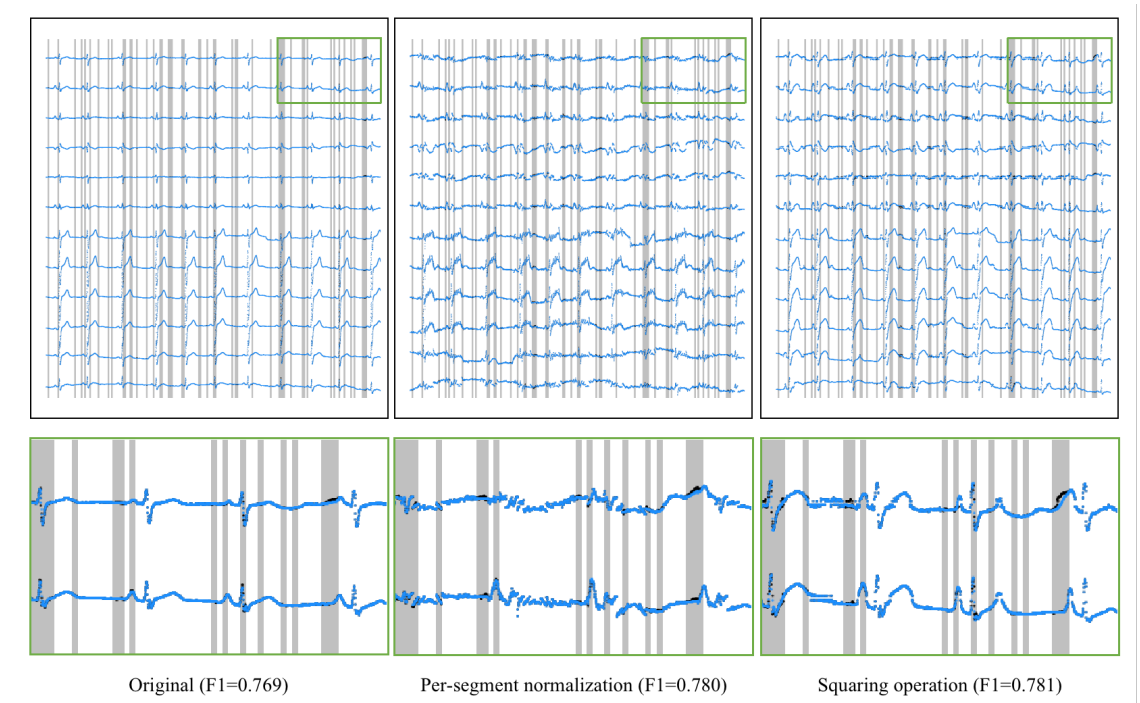

Figure 4: Reconstruction examples from one patient in the validation set under different reconstruction targets. The first row corresponds to the whole ECG signal, while the second row corresponds to a sequence of signal segments. The columns represent the original, per-segment normalization (4), and squaring operation (5) targets, respectively. The gray regions indicate the masked segments. The black point represents the original signal, while the blue point within the masked region represents the reconstructed signal.

the architectures and their fine-tuning performance on the validation set of the Fuwai dataset. The fine-tuning performance initially increase and then fall as the model size increases.

Table 1: Variants of our MTECG architecture and their fine-tuning performance. The #params column represents the number of trainable parameters.

Model	Segment size	D	h	L	#params	F1
MTECG-A	25	64	1	12	0.9M	0.752
MTECG-M	25	128	2	12	2.7M	0.775
MTECG-T	25	192	3	12	5.7M	**0.780**
MTECG-S	25	384	6	12	21. 8M	0.775
MTECG-B	25	768	12	12	85.8M	0.776

4.6 Comparison with the state-of-the-art algorithms

We compare our proposed MTECG-T with four state-of-the-art approaches: (i) CLECG (Chen et al., 2021), (ii) MaeFE (Zhang et al., 2022), (iii) CRT (Zhang et al., 2023), and (iv)

BaT (Li et al., 2021). These methods represent distinct techniques in the domain of ECG classification. Specifically, CLECG utilizes contrastive learning to enhance CNN performance. MaeFE and CRT employ masked pre-training for Transformer, which differs from our method in terms of model architecture, training strategies, and data preprocessing. BaT is specifically designed with a Transformer architecture to learn information from ECG heartbeats.

We conduct experiments across three different settings, indicated as Fuwai, PTB-XL, and PCinC in Table 2. For the two-stage methods, including CLECG, MaeFE, CRT and MTECG-T, we develop algorithms as follow. In the first setting, we pre-train and fine-tune the models on the training set of the Fuwai dataset. In the second setting, we pre-train the models on the PCinC dataset, excluding PTB-XL, and then fine-tune them on the training set of PTB-XL. In the third setting, both pre-training and fine-tuning are performed on the training set of the PCinC dataset. For the single-stage method, i.e., BaT, we train the model from scratch on the training set of Fuwai, PTB-XL and PCinC, respectively. For these methods involved in the comparison, we adopt the data preprocessing, training strategies and hyper-parameters recommended in the corresponding reference papers for each method (Chen et al., 2021; Zhang et al., 2022; Li et al., 2021; Strodthoff et al., 2020; Zhang et al., 2023). Particularly, for CLECG, we adopt xresnet1d101 (Strodthoff et al., 2020) as the backbone. In the case of MaeFE, we employ wavelet transform filtering for data preprocessing, as proposed by (Martis et al., 2013), and utilize the MTAE masking strategy with the ViT backbone, as suggested in (Zhang et al., 2022). For CRT, we adopt the training strategy and backbone as detailed in (Zhang et al., 2023). For BaT, we use the segmentation method (Makowski et al., 2021) to obtain ECG heartbeats. The algorithm comparison is demonstrated in Table 2, from which we can see MTECG-T significantly outperforms the alternatives across all datasets. Notably, MTECG-T increases the macro F1 scores by 3.4%-27.5% in Fuwai, 9.9%-32.0% in PTB-XL, and 9.4%-39.1% in PCinC, compared to the alternative methods.

Table 2: Prediction performance on 3 real datasets. Reported are the macro F1 scores on the testing set of each dataset.

Methods	Fuwai	PTB-XL	PCinC
CLECG (Chen et al., 2021)	0.740	0.518	0.590
MaeFE (Zhang et al., 2022)	0.705	0.533	0.605
CRT (Zhang et al., 2023)	0.600	0.444	0.476
BaT (Li et al., 2021)	0.702	0.462	0.561
MTECG-T	**0.765**	**0.586**	**0.662**

5 Discussion

The application of deep learning for ECG classification has gained significant popularity in various domains. In this study, we extensively evaluate and explore the application of the in-

novative Transformer architecture and masked modeling methods, leveraging both private and publicly available ECG datasets. Our experiments demonstrate promising performance of the proposed method, which also underscore their potential effectiveness in clinical applications. Moreover, our findings raise several questions as follows.

1. **A standard ECG is worth 200 words?** In this paper, the segments, architectures and training strategies are similar to techniques used in NLP. One may wonder whether the deep learning methods for ECG will also follow a similar trajectory as NLP. It is important to acknowledge that the simple self-supervised learning method in NLP enables benefits through model scaling. However, our scaling experiments demonstrate that the masked pre-training method fails to sustain scalable benefits, with performance exhibiting an initial rise followed by a subsequent decline as the model size increases. One potential explanation could be the insufficient size of the training data. Although larger than almost all publicly available ECG datasets, our dataset is still relatively small compared to the datasets commonly used in NLP and CV pre-training tasks.

2. **The Transformer is suitable for ECG time series in lightweight regimes?** On the other hand, our experiments show that the lightweight Transformer could achieve excellent performance in ECG classification if proper training strategies are adopted. Note that our networks stem from the vanilla ViT. The findings break the popular conception that the vanilla Transformer is not suitable for classification tasks in lightweight regimes, which is consistent with the observation of Wang et al. (2023a). These facts also suggest that employing appropriate pre-training techniques the naive network architectures could be better than specifically designed ones. It is worth to mention the naive lightweight model are friendly to deployment in the clinical environment.

3. **How important the training recipe is in the masked modeling?** Previous emphasis was primarily on the masking strategy, training schedule length, and reconstruction target. However, in this work, we demonstrate that the fine-tuning performance is not only sensitive to the those components, but also to other components such as the layer-wise LR decay and DropPath rate. These findings indicate that the successful application of masked modeling in domain-specific tasks might rely on the implementation of proper pre-training and fine-tuning recipes.

4. **Transformer or CNN for ECG classification?** Transformer has shown remarkable performance in various visual benchmarks, often matching or surpassing that of CNN. One might be curious about whether Transformer could potentially replace CNN for ECG classification tasks. We compare the Transformer-based model pre-trained by the proposed masked method with the benchmark CNN-based model (xresnet1d101 (Strodthoff et al., 2020)) pre-trained by the contrastive method (CLECG), and observe that the Transformer-based model achieves significantly better performance. However, we cannot conclude that Transformer is a better architecture than CNN for ECG classification due to the different pre-training strategy. It would be interesting to investigate whether CNN could achieve a comparable performance by developing an appropriate masked pre-training method.

5. **Can the proposed method be utilized in novel clinical scenarios?** In recent years, the field of ECG has experienced remarkable advancements, particularly in the area of deep learning. Numerous studies have revealed the ability of deep learning models to identify subtle changes in ECG signals that are unrecognizable by the human eye (Siontis et al., 2021). Remarkably, these subtle patterns may be indicative of asymptomatic diseases,

such as asymptomatic left ventricular systolic dysfunction (Attia et al., 2019, 2022; Yao et al., 2021). The detection of these patterns introduces novel clinical scenarios for ECG analysis (Somani et al., 2021), where the labels for ECG are obtained from alternative modalities such as echocardiogram and computed tomography. However, a significant challenge lies in the large number of ECG signals that lack corresponding labels from other modalities. Previous research has typically excluded these unlabeled ECG records (Siontis et al., 2021; Somani et al., 2021). Our proposed method, characterized by its self-supervised nature, could effectively utilize these unlabeled ECG records. Consequently, it becomes intriguing to explore the performance of the proposed method in such scenarios.

6 Conclusion

In conclusion, the paper presents a useful masked Transformer method which expands the application of MAE to ECG time series. We interpret an ECG time series as a sequence of segments and process it by the lightweight Transformer. Despite its simplicity, this method performs surprisingly well when adopting masked pre-training combined with proper training strategies. Therefore, the proposed algorithm outperforms recent state-of-the-art algorithms across multiple ECG classification datasets. In addition, the derived lightweight model offers deployment-friendly features, which is attractive in the clinical environment. We hope that this study could direct future research on the application of Transformer to more ECG tasks.

References

Attia, Z. I., Harmon, D. M., Dugan, J., Manka, L., Lopez-Jimenez, F., Lerman, A., Siontis, K. C., Noseworthy, P. A., Yao, X., Klavetter, E. W. et al. (2022) Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. *Nature Medicine*, 1–7.

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M. et al. (2019) Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine*, **25**, 70–74.

Ba, J. L., Kiros, J. R. and Hinton, G. E. (2016) Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bagui, S. and Li, K. (2021) Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, **8**, 1–41.

Bao, H., Dong, L., Piao, S. and Wei, F. (2021) Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bousseljot, R., Kreiseler, D. and Schnabel, A. (1995) Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet.

Chen, H., Wang, G., Zhang, G., Zhang, P. and Yang, H. (2021) Clecg: A novel contrastive learning framework for electrocardiogram arrhythmia classification. *IEEE Signal Processing Letters*, **28**, 1993–1997.

Clark, K., Luong, M.-T., Le, Q. V. and Manning, C. D. (2020) Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009) Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gutmann, M. and Hyvärinen, A. (2010) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. et al. (2022) A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 87–110.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P. and Ng, A. Y. (2019) Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, **25**, 65–69.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R. (2022) Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

Hong, S., Zhou, Y., Shang, J., Xiao, C. and Sun, J. (2020) Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, **122**, 103801.

Huang, G., Sun, Y., Liu, Z., Sedra, D. and Weinberger, K. Q. (2016) Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 646–661. Springer.

Hughes, J. W., Olgin, J. E., Avram, R., Abreau, S. A., Sittler, T., Radia, K., Hsia, H., Walters, T., Lee, B., Gonzalez, J. E. et al. (2021) Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA Cardiology*, **6**, 1285–1295.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S. and Shah, M. (2022) Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, **54**, 1–41.

Kiyasseh, D., Zhu, T. and Clifton, D. A. (2021) Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, 5606–5615. PMLR.

Ko, W.-Y., Siontis, K. C., Attia, Z. I., Carter, R. E., Kapa, S., Ommen, S. R., Demuth, S. J., Ackerman, M. J., Gersh, B. J., Arruda-Olson, A. M. et al. (2020) Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *Journal of the American College of Cardiology*, **75**, 722–733.

Krishnan, R., Rajpurkar, P. and Topol, E. J. (2022) Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 1–7.

Li, X., Li, C., Wei, Y., Sun, Y., Wei, J., Li, X. and Qian, B. (2021) Bat: Beat-aligned transformer for electrocardiogram classification. In *2021 IEEE International Conference on Data Mining (ICDM)*, 320–329. IEEE.

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z. et al. (2018) An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, **8**, 1368–1373.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J. and Tang, J. (2021) Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 857–876.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C. and Chen, S. A. (2021) Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, 1–8.

Martis, R. J., Acharya, U. R. and Min, L. C. (2013) Ecg beat classification using pca, lda, ica and discrete wavelet transform. *Biomedical Signal Processing and Control*, **8**, 437–448.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F. and Yang, M.-H. (2021) Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, **34**, 23296–23308.

Pourbabaee, B., Roshtkhari, M. J. and Khorasani, K. (2017) Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, **48**, 2095–2104.

Pyakillya, B., Kazachenko, N. and Mikhailovsky, N. (2017) Deep learning for ecg classification. In *Journal of physics: conference series*, vol. 913, 012004. IOP Publishing.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018) Improving language understanding by generative pre-training.

Reyna, M. A., Sadr, N., Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Rad, A. B., Elola, A., Seyedi, S., Ansari, S. et al. (2021) Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *2021 Computing in Cardiology (CinC)*, vol. 48, 1–4. IEEE.

— (2022) Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. *Physiological Measurement*, **43**, 084001.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P., Andersson, C. R., Macfarlane, P. W., Meira Jr, W. et al. (2020) Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature Communications*, **11**, 1760.

Sharma, L. D. and Sunkaria, R. K. (2021) Detection and delineation of the enigmatic u-wave in an electrocardiogram. *International Journal of Information Technology*, **13**, 2525–2532.

Siontis, K. C., Noseworthy, P. A., Attia, Z. I. and Friedman, P. A. (2021) Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, **18**, 465–478.

Somani, S., Russak, A. J., Richter, F., Zhao, S., Vaid, A., Chaudhry, F., De Freitas, J. K., Naik, N., Miotto, R., Nadkarni, G. N. et al. (2021) Deep learning and the electrocardiogram: review of the current state-of-the-art. *EP Europace*, **23**, 1179–1191.

Strodthoff, N., Mehari, T., Nagel, C., Aston, P. J., Sundar, A., Graff, C., Kanters, J. K., Haverkamp, W., Dössel, O., Loewe, A. et al. (2023) Ptb-xl+, a comprehensive electrocardiographic feature dataset. *Scientific Data*, **10**, 279.

Strodthoff, N., Wagner, P., Schaeffter, T. and Samek, W. (2020) Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, **25**, 1519–1528.

Tihonenko, V., Khaustov, A., Ivanov, S., Rivin, A. et al. (2007) St.-petersburg institute of cardiological technics 12-lead arrhythmia database. *Dataset on Physionet. org.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, **30**.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W. and Schaeffter, T. (2020) Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, **7**, 1–15.

Wang, S., Gao, J., Li, Z., Zhang, X. and Hu, W. (2023a) A closer look at self-supervised lightweight vision transformers. In *International Conference on Machine Learning*, 35624–35641. PMLR.

Wang, Z., Wang, Y., Hu, H. and Li, P. (2023b) Contrastive learning with consistent representations. *arXiv preprint arXiv:2302.01541*.

Wei, Y., Wang, X., Guan, W., Nie, L., Lin, Z. and Chen, B. (2019) Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing*, **29**, 1–14.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J. and Sun, L. (2022) Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

Wu, S.-C., Wei, S.-Y., Chang, C.-S., Swindlehurst, A. L. and Chiu, J.-K. (2021) A scalable open-set ecg identification system based on compressed cnns. *IEEE Transactions on Neural Networks and Learning Systems*.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q. and Hu, H. (2022) Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9653–9663.

Yao, X., Rushlow, D. R., Inselman, J. W., McCoy, R. G., Thacher, T. D., Behnken, E. M., Bernard, M. E., Rosas, S. L., Akfaly, A., Misra, A. et al. (2021) Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nature Medicine*, **27**, 815–819.

Zhang, H., Liu, W., Shi, J., Chang, S., Wang, H., He, J. and Huang, Q. (2022) Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, **72**, 1–15.

Zhang, W., Yang, L., Geng, S. and Hong, S. (2023) Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*.

Zheng, J., Chu, H., Struppa, D., Zhang, J., Yacoub, S. M., El-Askary, H., Chang, A., Ehwerhemuepha, L., Abudayyeh, I., Barrett, A. et al. (2020a) Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, **10**, 1–17.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H. and Rakovski, C. (2020b) A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, **7**, 1–8.