

Article

A Multi-Level Multiple Contrastive Learning Method for Single-Lead Electrocardiogram Atrial Fibrillation Detection

Yonggang Zou^{1,2}, Peng Wang¹, Lidong Du¹, Xianxiang Chen¹, Zhenfeng Li¹, Junxian Song^{3,*}
and Zhen Fang^{1,2,4,*}

¹ Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; zouyonggang20@mails.ucas.ac.cn (Y.Z.); wangpeng01@aircas.ac.cn (P.W.); lddu@mail.ie.ac.cn (L.D.); chenxx@aircas.ac.cn (X.C.); lizhenfeng@aircas.ac.cn (Z.L.)

² School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China

³ Department of Cardiology, Peking University People's Hospital, Beijing 100044, China

⁴ Research Unit of Personalized Management of Chronic Respiratory Disease, Chinese Academy of Medical Sciences, Beijing 100700, China

* Correspondence: sjx221@163.com (J.S.); zfang@mail.ie.ac.cn (Z.F.)

Abstract: Atrial fibrillation (AF) is the most common persistent arrhythmia, and it is crucial to develop generalizable automatic AF detection methods. However, supervised AF detection is often limited in performance due to the difficulty in obtaining labeled data. To address the gap between limited labeled data and the requirements for model robustness and generalization in single-lead ECG AF detection, we proposed a semi-supervised contrastive learning method named MLMCL for AF detection. The MLMCL method utilizes the multi-level feature representations of the encoder to perform multiple contrastive learning to fully exploit temporal consistency, channel consistency, and label consistency. Meanwhile, it combines labeled and unlabeled data for pre-training to obtain robust features for downstream tasks. In addition, it uses the domain knowledge in the field of AF diagnosis for domain knowledge augmentation to generate hard samples and improve the distinguishability of ECG representations. In the cross-dataset testing mode, MLMCL had better performance and good stability on different test sets, demonstrating its effectiveness and robustness in the AF detection task. The comparison results with existing studies show that MLMCL outperformed existing methods in external tests. The MLMCL method can be extended and applied to multi-lead scenarios and has reference significance for the development of contrastive learning methods for other arrhythmia.

Keywords: atrial fibrillation; deep learning; electrocardiogram (ECG); contrastive learning (CL)



Academic Editors: Mario Petretta and Giuseppe Baselli

Received: 8 December 2024

Revised: 31 December 2024

Accepted: 6 January 2025

Published: 8 January 2025

Citation: Zou, Y.; Wang, P.; Du, L.; Chen, X.; Li, Z.; Song, J.; Fang, Z. A Multi-Level Multiple Contrastive Learning Method for Single-Lead Electrocardiogram Atrial Fibrillation Detection. *Bioengineering* **2025**, *12*, 44. <https://doi.org/10.3390/bioengineering12010044>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Atrial fibrillation (AF) is the most common type of sustained arrhythmia [1]. It is defined and characterized by extremely rapid and uncoordinated atrial activities [2]. Atrial fibrillation can significantly increase the risk of developing other high-risk cardiovascular diseases, including stroke [3], systemic embolism [4], vascular dementia [4], heart failure [5], myocardial infarction [6], and sudden cardiac death [7]. Atrial fibrillation usually affects the elderly and initially presents as paroxysmal. Without intervention, paroxysmal AF tends to progress to persistent and permanent AF [2]. Currently, the electrocardiogram (ECG) is the most commonly used monitoring technique for detecting and quantifying the electrical activities of the heart [8]. The occurrence of arrhythmia is usually manifested as the change in ECG morphology and rhythm. For example, the common characteristics of AF on the

ECG include irregular RR intervals, the disappearance of P waves, and the appearance of f waves [9]. According to diagnostic conventions, a fibrillation episode lasting at least 30 s can be diagnosed as clinical AF [10]. However, the recognized complexity of ECG is much higher than that of ordinary images, which makes the diagnosis process time-consuming and error-prone. In addition, the quality of ECG diagnosis depends on the professional level of cardiologists, and it may even require multiple experts to resolve differences in decision making. As a result, most of the ECG data collected in the medical field have not been systematically organized and diagnosed. Therefore, it is of great significance to develop efficient and reliable automated detection methods to analyze and interpret ECG recordings.

So far, deep learning (DL) technology has made remarkable progress in the field of automated AF detection. Among them, supervised learning methods dominate current research. Representative deep learning network models include CNN [11–13], LSTM [14–17], Transformer [18,19], the fusion of multiple networks [20,21], etc. These methods have demonstrated excellent performance in multi-lead and single-lead AF detection. Some studies [22] have shown that DL has the potential to reach or even exceed the proficiency level of cardiologists in arrhythmia detection. However, for AF detection, the good generalization performance of supervised learning models depends on a large amount of high-quality annotated ECG data, which is always difficult to obtain. Therefore, it poses a serious problem for AF detection: the AF detection methods based on supervised learning are prone to overfitting in the case of limited annotated data, thus limiting their generalizability.

Considering that multiple instances sharing the same label should share some consistent representations of data that can be learned by the network, AF detection methods based on contrastive learning (CL) [23] attempt to learn this data consistency from unlabeled data to alleviate overfitting, thereby mitigating the impact of limited labeled data and reducing the dependence on labeled data. As a mainstream technique of self-supervised learning (SSL), CL includes two processes: the pre-training task and the downstream task. Among them, the pre-training task usually takes unlabeled data as input, while the labeled data are used for the downstream task. The core idea of this method is to encourage the model to distinguish similar and different data without labels so that the model maps similar samples from the same class closer in the high-dimensional feature space while separating samples from different classes, thus providing meaningful representations for fine-tuning with labeled data in the downstream task and generating more accurate predictions. Specifically, mainstream CL methods bring the representations of positive sample pairs composed of target samples and similar (positive) samples closer in the embedding space while pushing the representations of negative sample pairs composed of target samples and dissimilar (negative) samples farther apart. Therefore, in CL, the selection of positive sample pairs and the design of the corresponding contrastive loss function have received great attention from researchers.

In the field of computer vision, numerous effective CL frameworks such as SimCLR [24], MoCo [25], BYOL [26] widely apply data augmentation to generate positive samples. Among them, SimCLR and its improved versions [27] are the most commonly used frameworks. They take the target sample and the augmented sample as a positive pair, and they regard the target sample and other samples as negative pairs. In the field of time series, some studies attempt to implement augmentation techniques applicable to time series data, such as DTW data augmentation proposed by TimeCLR [28] and random cropping and timestamp masking proposed by TS2Vec [29].

Unlike image data and traditional time series data, there are inherent connections in multiple attribute dimensions among different ECG samples. The data consistency

resulting from these internal connections may be beneficial to downstream tasks and can serve as a basis for the selection of positive samples other than data augmentation. For example, TNC [30] conducts subject consistency by taking the time-neighboring samples of the same subject as positive pairs. The CLOCS [31] method further takes into account the samples of different leads and the samples of the same subject to achieve contrastive learning at the channel consistency and subject consistency.

However, despite the certain degree of progress made in the CL methods for ECG, the following problems still exist, which may lead to suboptimal performance in the AF detection task. Firstly, existing CL-based studies usually randomly select negative samples from training data, which means that negative samples may be selected from the samples of the same category as the target sample, resulting in a decline in the quality of representation. Secondly, the downstream task is not used to guide the entire process of positive sample selection in the pre-training task, including data augmentation and the inherent connections of ECG data. Some augmentation methods are unable to generate hard samples with sufficient differences for the downstream task, which is not conducive to learning distinguishable features. Some inherent connections of ECG data are unstable. For example, the paroxysmal nature of AF may cause the ECG data of the same subject to come from multiple distributions, making it impossible for CL to capture the subject consistency. In addition, existing deep learning usually uses data with similar distributions during the training and testing processes, which may exhibit instability in real-world scenarios. The cross-dataset testing method, which evaluates the performance on external datasets that are significantly different from the training data, is becoming an important solution to effectively address this problem.

To address the above issues, we propose a multi-level and multiple contrast learning (MLMCL) solution for AF detection using single-lead ECGs. This method introduces semi-supervised pre-training, which uses both labeled and unlabeled data during the pre-training stage. It constructs robust representations by selecting negative samples based on labels to prevent the decline in representation quality caused by randomly selecting negative samples. To fully tap into the full potential of ECG data, MLMCL performs multiple contrastive learning on the multi-level feature representations extracted by the encoder, which systematically mine the temporal consistency, channel consistency, and label consistency. Among them, the temporal contrastive learning focuses on the representative morphology at a single timestamp, the channel contrastive learning focuses on the invariant information across leads, and the label contrastive learning focuses on the information retained across subjects. In addition, by leveraging the ECG knowledge in the AF-related medical domain, we propose using vertical flipping and T-wave masking to achieve diagnostic region augmentation and non-diagnostic region augmentation, respectively, which together form the domain knowledge augmentation to generate hard samples. We apply MLMCL on the basis of domain knowledge augmentation to learn the generalizable feature representations in ECG, maximizing the potential of both labeled and unlabeled data and reducing the dependence on labeled data. Finally, we conduct pre-training under the cross-dataset testing setting. After obtaining the ECG representations, the model is fine-tuned for AF detection on the labeled dataset. The cross-dataset testing evaluation is carried through linear probing and full fine-tuning to ensure its robustness. Our contributions are as follows:

- This paper proposes a semi-supervised contrastive learning framework that uses both labeled and unlabeled data during the pre-training stage to construct robust representations by selecting negative samples based on labels.
- For AF detection, an MLMCL contrastive learning method is proposed. It performs multiple contrastive learning to extract the temporal consistency, channel consistency,

and label consistency on multi-level feature representations, thereby learning generalizable representations.

- By utilizing the knowledge in the AF diagnosis domain, a domain knowledge augmentation combining diagnostic region augmentation and non-diagnostic region augmentation is proposed for generating hard samples to learn distinguishable features.
- The proposed method outperforms existing methods under the cross-dataset testing mode, and the external tests on multiple datasets demonstrate the generalizability of the proposed method.

The remaining parts of this study are arranged as follows. Section 2 introduces the related work. Section 3 outlines the methods used in this study. Section 4 presents the datasets, experiments, and results. Section 5 summarizes this study.

2. Related Works

The essence of CL is to mine the common patterns of similar data. According to the requirements of downstream tasks, CL-based methods usually select positive and negative sample pairs in a customized manner. They learn data consistency by maximizing the similarity of the representations of positive sample pairs while minimizing the similarity of the representations of negative sample pairs. Selecting appropriate positive samples is crucial for ensuring the performance of downstream tasks [32]. For example, the SimCLR proposed by Chen et al. [24] defines positive samples as the augmented views of the target sample and directly regards the views from other samples in the current batch as negative samples. Tian et al. [33] use different modal views of the same sample as positive sample pairs. The way of selecting positive samples determines the semantic information of the learned representations. Therefore, it is important to develop a selection strategy of positive sample applicable to ECG.

The CL methods in the ECG field have been studied for extracting effective ECG representations. Some studies directly transfer CL methods from other fields to the ECG field. For example, Mehari et al. [34] directly compared instance-based CL methods (SimCLR, BYOL, and SwAV) and latent forecasting methods (CPC) to demonstrate the feasibility of learning useful representations from 12-lead ECG data through self-supervised learning. Soltanieh et al. [35] used multiple different augmentations and parameters to evaluate the effectiveness of the ECG representations of three contrastive learning methods (SimCLR, BYOL, and SwAV) on out-of-distribution datasets. Zhang et al. [36] adopted a contrastive learning method that manipulates temporal–spatial reverse augmentation to learn ECG representations and explored the impact of different combinations of horizontal flipping (temporal reverse) and vertical flipping (space reverse) in the pre-training stage on the downstream AF detection task. Although these methods have shown improvements over the fully supervised baseline, their selection of positive samples largely depends on data augmentation methods. More ECG contrastive learning methods construct positive samples by combining the data consistencies of the ECG inherent attributes. Such consistencies take various attribute dimensions, and each will be considered in turn, including the consistency of the time, subject, channel, rhythm, morphology, and label.

Temporal Consistency: Temporal consistency, also known as contextual consistency, encourages the feature representations at the same timestamp in different augmented views to be similar. TS2vec [29] proposes contextual consistency in the time dimension of sample representations. It regards the representations of the same timestamp in two augmented views as positive pairs and the representations of different timestamps as negative pairs, focusing on the representative morphology at a single timestamp in the sample.

Subject Consistency: The ECG data of the same subject usually maintains a highly identical pattern within a short period of time, which is the prerequisite for subject con-

sistency. Cheng et al. [37] incorporated a subject-aware condition into the SSL framework to promote the extraction of subject invariance through contrastive loss and adversarial training. Lan et al. [38] proposed an intra-inter subject self-supervised learning (ISL) model for arrhythmia diagnosis. The inter-subject SSL maximizes the subject consistency between different augmented views of the same subject and minimizes the similarity between different subjects to learn the unique representations of differences between different subjects. However, when the ECG data of the same patient comes from multiple distributions, the assumed prerequisite of subject consistency is not satisfied.

Channel Consistency: Different leads data of ECG share the same rhythmic characteristics and represent the same cardiac activity with different waveform morphologies. Channel consistency encourages the learning of invariance across leads, especially rhythmic invariance. Liu et al. [39] proposed a dense lead contrast (DLC) method, which explores the intra-lead and inter-lead invariance through contrastive learning between different leads. In the follow-up work, Liu et al. further [40] proposed a direct lead assignment (DLA) contrastive learning method. In pre-training, DLA simultaneously focuses on the global ECG representation and lead-specific features by performing contrastive learning between multi-lead and single-lead representations, thus improving the quality of single-lead representation. Some studies focused on both the subject consistency and channel consistency of ECG data, such as the 3KG method proposed by Gopal et al. [41] and the CLOCS method proposed by Kiyasseh et al. [31].

Rhythmic Consistency and Morphological Consistency: Liu et al. [42] proposed a morphology–rhythm contrast (MRC) learning framework. MRC performs dual contrastive learning through random beat selection (morphological view) and 0–1 pulse generation (rhythm view), thereby unifying the morphological and rhythmic features. Zhu et al. [43] designed pre-training tasks for intra-period and inter-period representation learning to capture the stable morphological features of a single period and the rhythmic features of multiple periods, respectively. The morphological consistency adopted by these methods [42,43] is a kind of subsequence consistency, which encourages the representation of the time series to be closer to its sampled subsequence. However, the representation of single-period subsequence features usually cannot reflect the overall morphological features. Due to the loss of information, subsequence consistency may not be applicable to ECG data.

Label Consistency: The label consistency representation is the goal of downstream tasks. Incorporating label information into contrastive learning can provide good initialization parameters for downstream tasks. The supervised CL proposed by Khosla et al. [44] extended CL from the self-supervised domain to the full-supervised domain. The proposed supervised contrastive loss function (SupCon), by introducing label information, makes samples belonging to the same class cluster more closely in the embedding space, while samples of different classes are pushed away from each other. On this basis, Le et al. [45] applied supervised contrastive learning to the classification of multi-lead ECG arrhythmias.

The above studies have demonstrated that using CL methods for ECG representation learning can achieve performance comparable to or even surpassing that of fully supervised methods. Summarizing the current research, existing methods mainly consider the consistency of attributes such as the time, channel, subject, rhythm, morphology (subsequence), and label of the ECG. However, existing studies usually only consider one or two attribute consistencies of ECG data and overlook the degree of matching of these strategies with the downstream task data, such as subsequence consistency. Few studies have proposed customized data augmentation methods for downstream tasks. Additionally, existing methods usually only use unlabeled data or only use labeled data for pre-training. These limitations result in the inability of existing methods to fully tap into the full potential

of ECG data. To overcome these limitations, we develop a CL method for AF detection by leveraging the ECG data consistencies of multiple attribute dimensions such as time, channel, and label, as well as AF domain knowledge augmentation methods. We do not specifically implement rhythm consistency, since it has already been achieved in channel consistency and diagnostic region augmentation. Due to the presence of a large number of paroxysmal samples in AF records, which may cause the data of the same subject to have different distributions, we abandon the subject consistency strategy in AF detection.

3. Method

The overall process of the proposed MLMCL algorithm is shown in Figure 1, which includes two parts: the semi-supervised pre-training task and the downstream task. In the pre-training part, multiple contrast learning is performed on multi-level representations of labeled and unlabeled data, aiming to learn the general representations of single-lead ECGs. The downstream task is the AF detection in single-lead ECG. The MLMCL method contains three key components: domain knowledge augmentation, a multi-level encoder, and multiple contrastive losses. The following provides a more detailed description from three aspects: data augmentation, contrastive learning pre-training, and the downstream task.

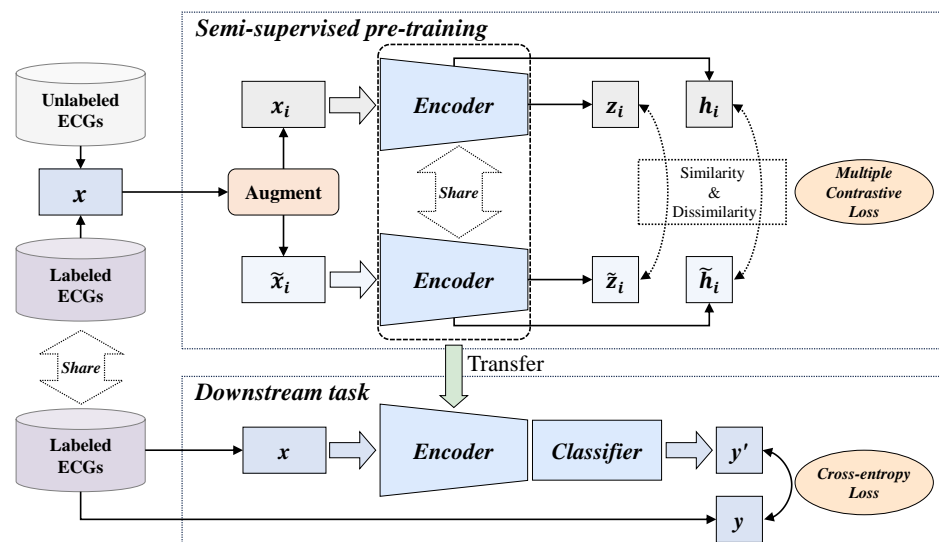


Figure 1. The overall flow of the proposed MLMCL algorithm.

3.1. Domain Knowledge Augmentation

Data augmentation is a crucial part for the success of CL methods. It does not need to utilize the inherent connections (such as label categories) of different sample data to obtain similar positive samples. Instead, it only needs to directly transform a single sample to generate similar views. When the data augmentation transformation operators are applied to each ECG instance, the guarantee for extracting efficient representations is to maintain the invariance of important information in the ECG records. However, sometimes series data augmentation methods produce less difference between the augmented sample and the original ECG sample, making it impossible to generate reliable “hard” samples to help the model locate the key invariant features related to downstream tasks.

In ECG, arrhythmia may change the ECG rhythm and the morphology of specific regions, and each type of arrhythmia has distinguishable characteristic patterns. We attempted to utilize the medical domain knowledge of ECG diagnosis to address the issue of hard sample generation [46]. The proposed domain knowledge augmentation is a method that combines diagnostic region augmentation and non-diagnostic region augmentation. Both the diagnostic and non-diagnostic region augmentation modify the

signal values in the corresponding regions to generate “hard” samples for achieving robust learning. The aim is to transform the waveform of the diagnostic/non-diagnostic regions to a large extent so that it is distinguishable from the original sample and other ECG categories, thereby explicitly guiding the model to learn distinguishable features.

Diagnostic Region Augmentation: As shown in Figure 2, the key diagnostic regions of AF lie in the P-wave and QRS-wave regions. Considering that the positive and negative morphologies of P wave/QRS wave are symmetrical in the vertical direction, we thus adopt vertical flipping for diagnostic region augmentation. Vertical flipping will create a sufficiently large difference between the original ECG sample and the augmented sample, without causing confusion between the P wave and f wave. The vertical flipping augmentation of the raw ECG signal x can be represented as $\tilde{x} = -x$.

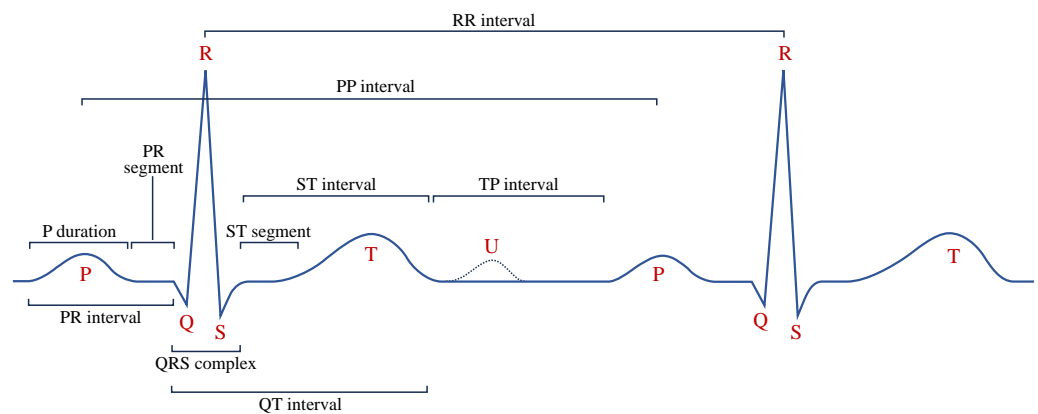


Figure 2. The classical ECG waveform and crucial segments with measurement points.

Non-Diagnostic Region Augmentation: For the non-diagnostic regions of AF, considering that the morphology and position of the T wave are not the key factors for diagnosing AF, we adopt the method of T-wave masking to reduce the interference of the T-wave part on the extraction of diagnostic features, making the model focus more on the diagnostic regions such as the P wave and QRS wave. T-wave masking sets the specific ST interval in the single-lead ECG signal to a fixed value. Specifically, its implementation includes two steps. First, the QRS-wave region of each heartbeat needs to be identified on the entire ECG record, since the QRS wave is the most distinctive wave region and is widely used as a reference for locating other characteristic wave regions. Second, the $c\%$ number of the ST intervals in the single-lead ECG sample are set to a fixed value. Usually, the T wave appears within the range of 300 ms after the R wave. Considering that the duration of the QRS wave is usually 80 ms to 100 ms, we achieved T-wave masking by setting the region from 50 ms to 300 ms after the R wave to a fixed value. In the experiments, a typical value for the masking parameter is $c = 50$.

Typically, contrastive learning methods use two augmented variants with different strengths to improve the robustness of the learned representations [47]. In this paper, the weak augmented variant directly uses the pre-processed data without transforming the original ECG sample, since the ECG samples are not augmented in the downstream task. The strong augmented variant adopts the domain knowledge augmentation strategy to generate “hard” samples. A typical implementation of domain knowledge augmentation is shown in Figure 3. In fact, we applied diagnostic region augmentation on top of non-diagnostic region augmentation with a 50% probability to achieve domain knowledge enhancement. Given a raw single-lead ECG sample x_i , augmentation produces two views that can be represented as x_i and \tilde{x}_i , respectively. These views are passed to the encoder to extract their high-dimensional latent representations.

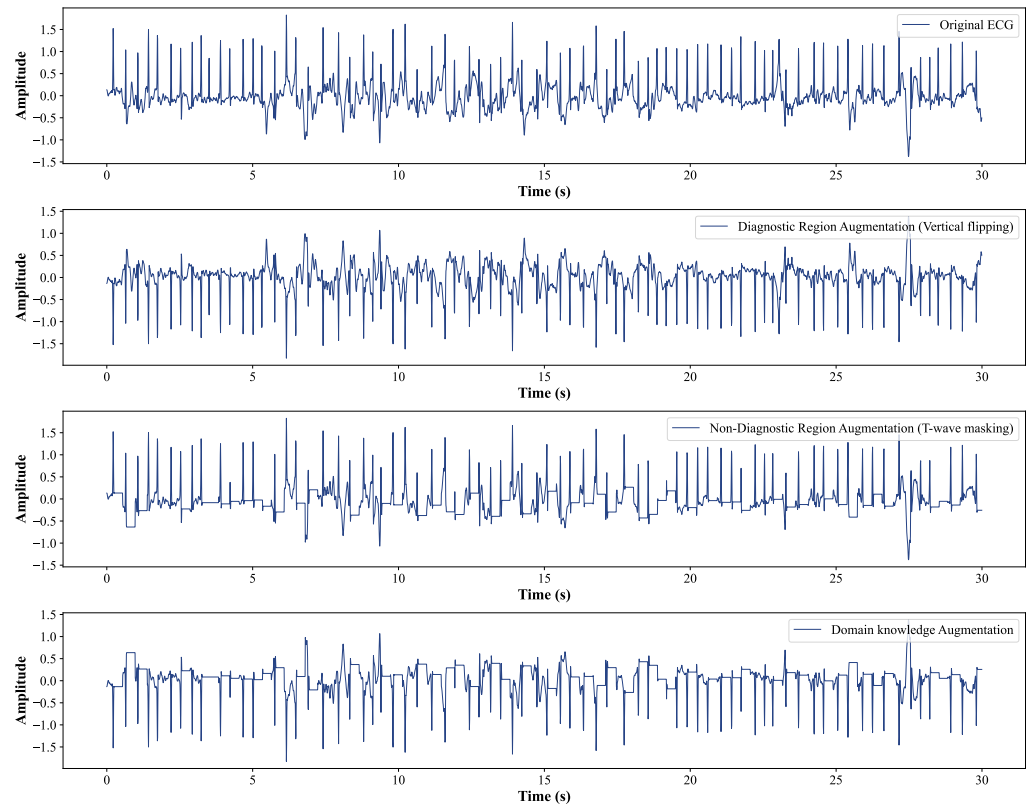


Figure 3. The diagram of domain knowledge augmentation.

3.2. Multi-Level Multiple Contrastive Learning Pre-Training

The proposed MLMCL is a semi-supervised contrastive learning method for single-lead ECGs, which can learn ECG representations from both labeled and unlabeled data simultaneously. Each single-lead ECG $x \in \mathbb{R}^{T_0 \times 1}$ is regarded as single-channel 1D data with a sequence length of T_0 . Given a labeled dataset $D_L = \{x_i, y_i\}_{i=1 \dots M}$ that contains M instance/label pairs and an unlabeled dataset $D_U = \{x_i\}_{i=1 \dots N}$ that contains N instances, the goal of the MLMCL pre-training is to learn an encoder to extract effective representations $z_i \in \mathbb{R}^{D_z}$ relevant to the downstream task from each x_i .

During the pre-training stage, two augmented copies of the same sample are fed into a multi-level encoder to obtain the intermediate hidden representations and the output encoder representations, and data consistency in multiple attribute dimensions on the multi-layer representations is encouraged. The temporal contrastive loss is calculated on the hidden representations, and the channel contrastive loss and label contrastive loss are calculated on the encoder representations. By utilizing the inherent data consistency in multiple dimensions through contrastive learning, the encoder is optimized with multiple contrastive losses to help learn a representation that is both representative and generalizable.

As shown in Figure 4, the multi-level encoder consists of two parts, including a convolutional feature extraction network $f(\cdot)$ and a temporal-spatial feature fusion network $g(\cdot)$. The convolutional feature extraction network adopts a residual block design. It extracts the context representation of each sample through one input convolutional layer and six residual blocks. It maps x_i to a high-dimensional latent space to obtain the hidden representation $h_i = f(x_i) \in \mathbb{R}^{T \times D_h}$. The hidden representation $h_i = \{h_{i,1}, h_{i,2}, \dots, h_{i,T}\}$ contains T representation vectors in the time dimension. The feature vector $h_{i,t} \in \mathbb{R}^{D_h}$ at the t -th timestamp is a representation vector with D_h dimensions and is used to calculate the temporal contrastive loss.

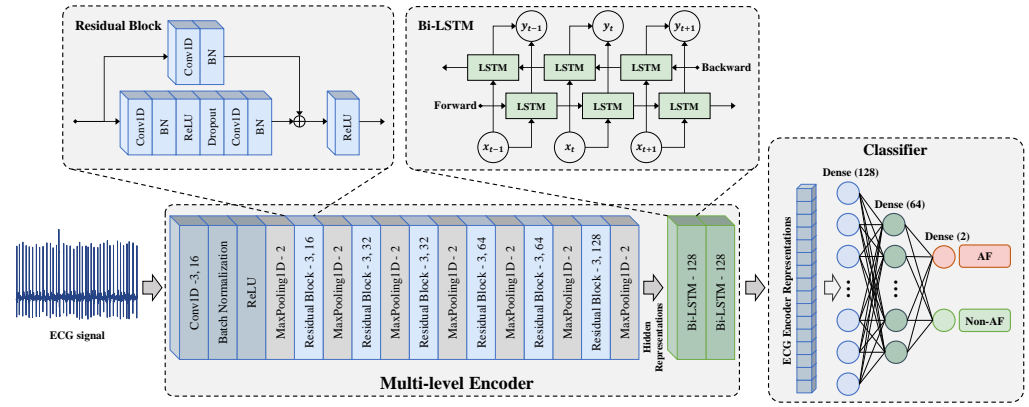


Figure 4. The encoder architecture used for pre-training and the classifier architecture used for fine-tuning.

The temporal–spatial feature fusion network $g(\cdot)$ is constructed using the Bidirectional Long Short-Term Memory Network (Bi-LSTM) to extract the temporal-spatial features of ECG. The $g(\cdot)$ learns from the hidden representation h_i to obtain the representation $r_i = g(h_i) \in \mathbb{R}^{T \times D_z}$, which also contains T representation vectors in the time dimension. Since the output of the last time step of the LSTM network integrates the temporal–spatial information of h_i in order to reduce the amount of computation, the output $z_i = r_{i,T}$ of the last time step is taken as the output encoder representation. The output z_i of the encoder is normalized onto the unit hypersphere in \mathbb{R}^{D_z} , which makes it possible to use the inner product to measure the distance in the latent space.

Unlike the models with the SimCLR architecture, we adopted a projector-free design. The experimental results (Section 4.4.3) show that our method did not lead to a performance decline without using a projector. In fact, the absence of an additional projector could further reduce the pre-training parameters and time consumption.

3.2.1. Temporal Contrastive Learning

Contrastive learning generally assumes that the representation of an augmented sample will carry information similar to that of the corresponding original sample. Analogously, their representation vectors at the same timestamp will also carry similar context information, especially for high-dimensional latent representations extracted by convolutional networks because of the same receptive field and the same network parameters. In order to learn discriminative representations that change over time, we chose to learn temporal consistency. For two augmentations of the same sample, representation vectors with the same timestamp are regarded as positive pairs, and representation vectors from different timestamps are treated as negative pairs [29]. Temporal contrastive learning helps the encoder focus on the representative features at a single timestamp. Since temporal consistency does not rely on labels, temporal contrastive learning was performed on both labeled and unlabeled data.

For a given ECG sample x_i and its augmentation \tilde{x}_i , the convolutional feature extraction network in the encoder extracts their respective hidden representations as $h_i = f(x_i)$ and $\tilde{h}_i = f(\tilde{x}_i)$, respectively. To capture the temporal consistency, the positive pair is $(h_{i,t}, \tilde{h}_{i,t})$, and the negative pairs are $(h_{i,t}, h_{i,t'})$ and $(h_{i,t}, \tilde{h}_{i,t'})$. In practice, the contrastive loss is calculated within a mini-batch of data. In a batch with B raw ECG samples, the temporal contrastive loss [29] is defined as

$$\mathcal{L}_{tCL} = -\frac{1}{B} \sum_{i=1}^B \frac{1}{T} \sum_{t=1}^T \log \frac{\exp(h_{i,t} \odot \tilde{h}_{i,t} / \tau)}{\sum_{t'=1}^T (\exp(h_{i,t} \odot \tilde{h}_{i,t'} / \tau) + \mathbf{1}_{\{t' \neq t\}} \exp(h_{i,t} \odot h_{i,t'} / \tau))} \quad (1)$$

where $\mathbf{1}_{\{t' \neq t\}}$ is an indicator function, which equals 1 when the condition $t' \neq t$ is satisfied and equals 0 otherwise. The function $\exp(\cdot)$ is the exponential function and the function $\log(\cdot)$ is the natural logarithm function. The symbol $\tau \in \mathbb{R}^+$ is a scalar temperature parameter used to adjust the slope of the loss function. The symbol \odot represents the inner product operation and is used to calculate the cosine similarity between two vectors. The cosine similarity of vectors \mathbf{u} and \mathbf{v} is defined as $\mathbf{u} \odot \mathbf{v} = \mathbf{u}^T \mathbf{v} / (\|\mathbf{u}\| \|\mathbf{v}\|)$.

3.2.2. Channel Contrastive Learning

The 12 lead signals in the same ECG can be thought of as natural augmentation of each other [31], since multiple lead ECG signals collected simultaneously will reflect the same cardiac activity, and they are associated with the same class. Although some arrhythmias affect specific parts of the heart so that they can only be detected by a few leads, the irregular rhythm of AF is special and can be observed in all leads. Minimizing the inter-lead differences helps to discover the rhythm invariance among the leads [48]. Therefore, we utilize the invariance of different leads for channel contrastive learning.

For a given ECG sample x_i , its time-aligned sample of other leads can be represented as x'_i . The samples x_i and x'_i are input into the encoder to obtain their respective encoder representations \mathbf{z}_i and \mathbf{z}'_i . To capture the channel consistency, the encoder representation \mathbf{z}_i should be close to \mathbf{z}'_i , and conversely, away from the representation \mathbf{z}_j and \mathbf{z}'_j of any other different samples. Specifically, the positive pair is $(\mathbf{z}_i, \mathbf{z}'_i)$, and the negative pairs are $(\mathbf{z}_i, \mathbf{z}_j)$ and $(\mathbf{z}_i, \mathbf{z}'_j)$. In a batch with B initial ECG samples, the channel contrastive loss [39] is defined as

$$\mathcal{L}_{cCL} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\mathbf{z}_i \odot \mathbf{z}'_i / \tau)}{\sum_{j=1}^N (\exp(\mathbf{z}_i \odot \mathbf{z}'_j / \tau) + \mathbf{1}_{\{j \neq i\}} \exp(\mathbf{z}_i \odot \mathbf{z}_j / \tau))} \quad (2)$$

where the indicator function $\mathbf{1}_{\{j \neq i\}}$ equals 1 when the condition $j \neq i$ is satisfied and equals 0 otherwise.

3.2.3. Label Contrastive Learning

In self-supervised contrastive learning, the negative samples of the target sample are composed of samples randomly selected from the mini-batch data. When multiple negative samples and the target sample have the same label, the contrastive loss may push samples of the same class further apart. For downstream supervised tasks, this may lead to a decline in the quality of the representation. Therefore, label consistency was introduced to solve this problem. Label consistency is beneficial for learning domain-adaptive representations of certain diseases across patients or even datasets because it indicates that samples with the same label should exhibit shared patterns even if they are collected from different subjects in different ways. Here, we adopted semi-supervised learning in pre-training to introduce labeled data and label contrastive loss.

In a batch with B initial ECG samples, the positive samples of the target sample x_i will be generalized to any number of samples x_p that have the same label as the target sample. Different from the self-supervised contrastive loss, the label contrastive loss contrasts the set of all samples with the same class label with the remaining samples of other classes in the batch. The calculation formula of label contrastive loss [44] is as follows:

$$\mathcal{L}_{lCL} = -\frac{1}{B} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \frac{\exp(\mathbf{z}_i \odot \mathbf{z}_p / \tau)}{\sum_{j \in \mathcal{A}_i} \exp(\mathbf{z}_i \odot \mathbf{z}_j / \tau)} \quad (3)$$

In the formula, $\mathcal{A}_i = \{I \setminus \{i\}\}$ is the set of indices in the batch index $I = 1, 2, \dots, B$ that do not include i , $\mathcal{P}_i = \{p \in \mathcal{A}_i: y_p = y_i\}$ is the set of sample indices in the batch that have the same label as sample i , and $|\mathcal{P}_i|$ is the number of its elements.

3.2.4. Multiple Contrastive Loss

The calculation process of the multiple contrastive loss used in MLMCL is shown as Figure 5. The multiple contrastive loss \mathcal{L}_{CL} consists of three loss terms. The temporal contrastive loss and the label contrastive loss help to learn robust representations that are invariant to transformations, while the channel contrastive loss encourages the encoder to learn lead-invariant representations. To sum up, the multiple contrastive loss of the MLMCL method is as follows:

$$\mathcal{L}_{CL} = \lambda_1 \mathcal{L}_{tCL} + \lambda_2 \mathcal{L}_{cCL} + \lambda_3 \mathcal{L}_{lCL} \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$ are hyperparameters that adjust the scale of each loss and satisfy $\lambda_1 + \lambda_2 + \lambda_3 = 1$. For the labeled dataset, three contrastive losses were adopted. For the unlabeled data, only temporal contrastive loss and channel contrastive loss were adopted, that is, λ_3 was set to 0.

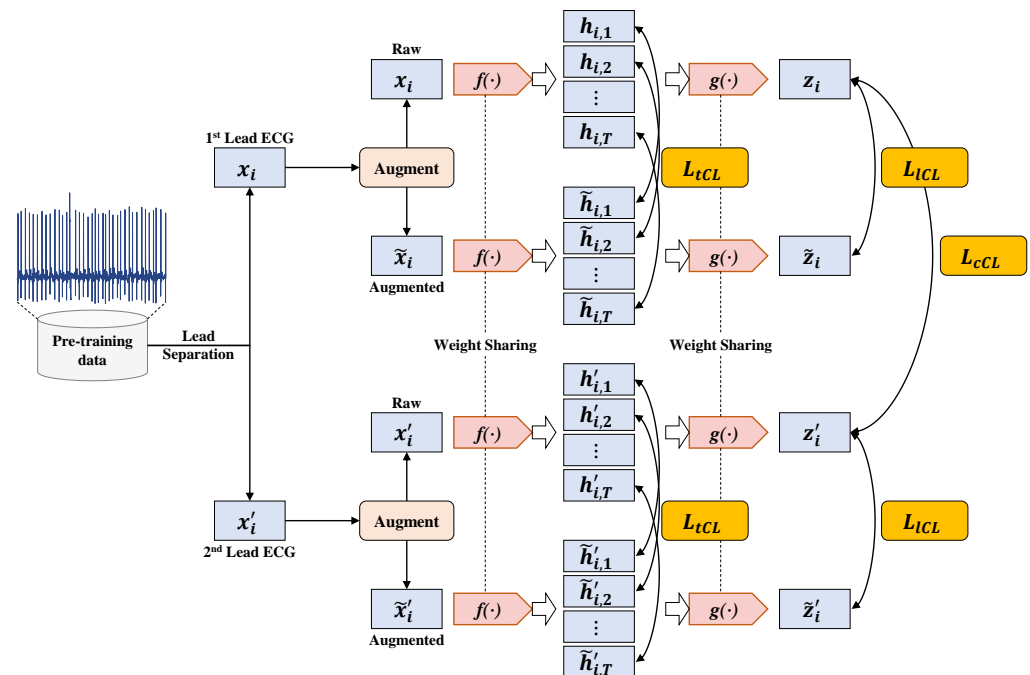


Figure 5. The calculation process of multiple contrastive loss.

3.3. Fine-Tuning for AF Detection Task

After the MLMCL pre-training is completed, the encoder composed of the convolutional feature extraction network $f(\cdot)$ and the temporal–spatial feature fusion network $g(\cdot)$ are passed to the downstream AF detection task as a feature extractor. During fine-tuning, in order to use the pre-trained encoder for classification, the encoder is fine-tuned and a classifier is trained. The classifier is usually designed as a multi-layer perceptron (MLP) with a hidden layer, and ReLU is the activation function used between each fully connected layer. As shown in Figure 4, the classifier used in the downstream task was designed with three fully connected layers, and a batch normalization layer was added before the activation function to stabilize the training. The cross-entropy loss was used for the supervised training of AF detection:

$$\mathcal{L}_{CE} = -\frac{1}{B} \sum_{j=1}^B \sum_{i=1}^C (y_i \log y'_i + (1 - y_i) \log(1 - y'_i)) \quad (5)$$

Among them, C is the number of categories, y_i is the i -th element of the true label \mathbf{y} , and y'_i is the i -th element of the prediction result \mathbf{y}' .

4. Experiments and Results

4.1. Databases and Data Preprocessing

In this work, four public databases from Physionet [49] were used to evaluate the proposed method. There is no subject overlap among these databases. The specific introductions are as follows:

- MIT-BIH Atrial Fibrillation Database (AFDB) [50]: It consists of long-term ECG recordings of 23 human subjects with AF (mostly paroxysmal). Each recording contains two different lead ECG signals with unspecified lead names and lasts for 10 h with a sampling rate of 250 Hz. Rhythm annotation files and heartbeat annotation files are provided separately. The rhythm annotation files are manually prepared, and most of the heartbeat annotation files are obtained by an automatic detector without manual correction.
- The 4th China Physiological Signal Challenge 2021 (CPSC2021) [51]: This database has publicly released two training sets, which altogether include 1406 dual-lead recordings from 105 patients. These recordings have been extracted from the ECG readings of 49 AF patients (23 paroxysmal AF patients) and 56 non-AF patients (usually including other abnormal and normal rhythms). The provided annotations include heartbeat annotations, rhythm annotations, and diagnoses of the global rhythm. Each recording consists of lead I and lead II signals, with a sampling frequency of 200 Hz, and the duration of each recording is not fixed.
- Long-Term AF Database (LTAfDB) [52]: It includes long-term ECG recordings of 84 subjects with paroxysmal or persistent AF. Each recording contains two simultaneously recorded ECG signals with a sampling frequency of 128 Hz, and the recording duration usually ranges from 24 to 25 h.
- MIT-BIH Arrhythmia Database (MITDB) [53]: It contains 48 half-hour dual-channel ECG recordings from 47 subjects, and the recordings are digitized at 360 Hz. Two or more cardiologists independently annotated each recording, and the reference annotations for each heartbeat were given after resolving differences. The lead names of the two ECG signals in each ECG recording are not fixed, mainly leads II and V1, and a few are leads V2, V4, and V5.

The detailed information of the four open-source datasets is summarized in Table 1. The channels of leads I and II were preferentially used when lead information was available. If it was unavailable, the channel similar to lead II among the two channels was selected for analysis.

Table 1. Details of the datasets used in this study.

Dataset	Freq	NR	NS	Lead	Record Length	Rhythms	TD	AFD	NEB
AFDB	250 Hz	25	25	ECG1, ECG2	10 h	4	234.28 h	93.40 h (39.87%)	/
CPSC2021	200 Hz	1436	105	I, II	0~6.8 h	3	480.19 h	164.44 h (34.24%)	93,545 (4.4%)
LTAfDB	128 Hz	84	84	ECG1, ECG2	6~26 h	9	1960.60 h	1030.89 h (52.58%)	285,100 (3.2%)
MITDB	360 Hz	48	47	II, V1, V2, V4, V5	0.5 h	15	24.07 h	2.21 h (9.18%)	34,442 (31.5%)

Freq: Sampling frequency, NR: Total number of records, NS: Number of subjects in the recording, TD: Total duration, AFD: AF duration, NEB: Number of ectopic beats.

Before the experiment, the ECG recordings were resampled to 128 Hz to ensure a uniform sampling rate. A band-pass filter with a frequency range of 0.5–40 Hz was used to remove the baseline drift and high-frequency noise in the ECG signals. Each recording was segmented into 30 s ECG segments for analysis. The parts shorter than 30 s were discarded. Each ECG segment was normalized using the z-score method. According to the annotations, each segment was labeled as AF or non-AF. Among them, all rhythms except AF and AFL were labeled as non-AF rhythms. To remove the segments containing severe noise, the beat signal quality index (bSQI) [53] of each ECG segment was used to evaluate the signal quality. The segments with $bSQI < 0.8$ were excluded from the analysis in this paper. In addition, we reserved the segments of another simultaneously collected lead for each ECG segment for subsequent experiments. The detailed segmentation information of the databases is shown in Table 2.

Table 2. Data description after segmentation.

Dataset	Segment Length	Overlap	Non-AF Segments	AF Segments
AFDB	30 s	15 s	31,547	21,383
CPSC2021	30 s	0 s	68,904	37,132
LTAADB	30 s	0 s	100,163	115,952
MITDB	30 s	15 s	4579	452

4.2. Experimental Settings

4.2.1. Evaluation Paradigms

It is worth noting that the quality of the representations from semi-supervised pre-training should be evaluated by the downstream AF detection task. Better performance of AF detection indicates higher quality of ECG representations. To evaluate the performance, the accuracy (Acc), macro sensitivity (Sen), macro precision (Pre), and macro F1-score (F1) were used as the metrics of classification performance. All these metrics range from 0 to 1. The larger the value, the better the performance.

4.2.2. Parameter Settings

For the input data, normalized single-lead ECGs with a sampling frequency of 128 Hz and a time length of 30 s were used for both pre-training and fine-tuning. In the pre-training stage, the length of the hidden representation vector was $D_h = 128$, and the length of the encoded representation vector was $D_z = 256$. In addition to $\tau = 1.0$ for temporal contrastive loss, $\tau = 0.1$ was used as the temperature hyperparameter for all other contrastive losses. For the labeled data, the loss coefficients adopted $\lambda_1 = \lambda_2 = \lambda_3$, while for the unlabeled data, the loss coefficients adopted $\lambda_1 = \lambda_2$. The entire pre-training process lasted for 40 epochs, and the batch size of the pre-training data was 256. The Adam optimizer was used with an initial learning rate of 0.001. The StepLR was adopted to adjust the learning rate, and the learning rate decayed to 0.8 of the original value every three epochs, ensuring the stability of the entire optimization process. In the fine-tuning stage, the Adam optimizer was also used for parameter optimization, and the initial learning rate η_{base} was set to the same as in the pre-training stage. The batch size for fine-tuning was 512, and the number of iterative epochs T_{max} was 30. The formula for the learning rate change with training epochs is

$$\eta = \eta_{base} / (1 + 10 \times T_{cur} / T_{max})^2 \quad (6)$$

where T_{cur} represents the current training epoch.

4.2.3. Implementation

All models were implemented based on PyTorch (v1.12.1). With the support of the Intel (R) Xeon (R) E5-2640 CPU and the NVIDIA GeForce RTX 3090 GPU in terms of hardware

configuration, Python (v3.9.15) was used to implement the pre-training, fine-tuning, and testing of the proposed MLMCL method.

In this work, we mainly adopted the cross-dataset testing mode, which tests with external datasets that have not been seen in the model training phase. The cross-dataset testing follows the inter-patient paradigm. The CPSC2021 dataset was used as the labeled training set for pre-training and downstream task training, because it has the largest number of individual subjects, a relatively balanced data distribution, and sufficient training samples. The AFDB and LTAfDB were respectively used as external independent test sets. Each test dataset was not used in the pre-training and fine-tuning processes but only for evaluation. During their testing, the remaining datasets were added to the pre-training as unlabeled data.

We tried linear probing and full fine-tuning in the training of downstream AF detection model. During training, linear probing freezes the encoder parameters and only updates the classifier parameters to perform AF detection. It is usually used to evaluate the quality of the learned representations for that downstream results completely depend on the pre-trained encoder. In contrast, full fine-tuning adjusts both the encoder and the classifier simultaneously to adapt to the AF detection task. In the full fine-tuning mode, pre-training is equivalent to providing more effective initial parameter values.

4.2.4. Baselines

To compare with the proposed MLMCL method, we conducted four baseline methods. (1) Fully supervised (FS): Supervised training was started on randomly initialized encoders and classifiers. (2) Fully supervised with data augmentation (FS + DA): Domain knowledge augmentation was introduced on the basis of full supervision. (3) SimCLR [24]: As a classical framework of CL, SimCLR forms positive samples of target sample only through data augmentation and takes other samples of the same batch as negative samples. (4) T-S [36]: Temporal reverse, spatial reverse, and temporal-spatial reverse were performed on the original signals, and then pre-training was completed by classifying four signals, including the original signal.

It is worth noting that these baseline methods used the same encoder architecture and datasets as the proposed MLMCL method for comparison.

4.3. Results

In this section, we evaluated the effectiveness of the proposed MLMCL method and compared it with various baseline methods on the AF detection task.

4.3.1. Linear Probing

Linear probing aims to assess the quality of representations. Table 3 shows the linear probing results of AF detection on the AFDB and LTAfDB databases. In the linear probing mode, the fully supervised method used a randomly initialized encoder and did not have any prior knowledge about ECG information. The results show that MLMCL outperformed the fully supervised method by 47.54% and 33.54% in terms of Acc and by 50.54% and 33.61% in terms of F1-score on AFDB and LTAfDB, which verifies the effectiveness of the pre-training. Similarly, MLMCL also obtained results that were significantly higher than those of the fully supervised method with data augmentation. Compared to SimCLR and T-S that used self-supervised training, MLMCL achieved better performance. Overall, the Acc and F1-scores of MLMCL on all external test sets were higher than those of all the baselines, which indicates that MLMCL is able to provide high-quality representations.

Table 3. Linear probing results of different methods on the external ECG datasets.

Methods	AFDB				LTAADB			
	Acc (%)	Pre (%)	Sen (%)	F1 (%)	Acc (%)	Pre (%)	Sen (%)	F1 (%)
FS	50.64	47.63	47.75	47.56	63.07	63.99	63.71	62.99
FS + DA	47.39	44.03	44.34	44.06	59.15	61.15	60.37	58.53
T-S	67.12	69.10	61.20	59.84	62.71	69.14	64.59	61.00
SimCLR	80.63	81.77	81.31	80.61	81.70	83.58	78.72	79.81
MLMCL	98.18	98.33	97.90	98.10	96.61	96.56	96.64	96.60

4.3.2. Full Fine-Tuning

Full fine-tuning allows fine-tuning of the entire network parameters. Table 4 presents the full fine-tuning results of MLMCL and multiple baselines. Compared with the fully supervised training with a randomly initialized encoder, the performance of the model pre-trained by MLMCL was greatly improved, indicating that it provides a more effective network initialization. In addition, MLMCL consistently outperformed the fully supervised method with data augmentation on AFDB and LTAADB, which shows that MLMCL pre-training is more helpful for the encoder to extract key features from the augmented data. Overall, compared to fully supervised training, MLMCL can achieve better performance by simultaneously utilizing labeled and unlabeled data, alleviating the annotation burden in practical applications.

Table 4. Full fine-tuning results of different methods on the external ECG datasets.

Methods	AFDB				LTAADB			
	Acc (%)	Pre (%)	Sen (%)	F1 (%)	Acc (%)	Pre (%)	Sen (%)	F1 (%)
FS	89.98	90.85	88.47	89.32	95.88	95.82	95.92	95.86
FS + DA	95.03	95.70	94.11	94.76	95.82	95.75	95.90	95.81
T-S	93.65	94.65	92.40	93.25	96.24	96.18	96.28	96.22
SimCLR	96.55	96.41	96.43	96.42	95.68	95.61	95.78	95.67
MLMCL	99.14	99.19	99.03	99.10	97.34	97.34	97.31	97.33

In addition, the Acc and F1-score of MLMCL exceeded those of the best baseline SimCLR by 2.59% (96.55% vs. 99.14%) and 2.68% (96.42% vs. 99.10%), respectively, on AFDB. On the LTAADB dataset, the Acc and F1-score of MLMCL exceeded those of the best baseline T-S by 1.10% (96.24% vs. 97.34%) and 1.11% (96.22% vs. 97.33%), respectively. Compared to other contrastive learning baseline methods, MLMCL had a consistent advantage on different test sets, indicating that it has good stability. As shown in Figure 6, the fully fine-tuning method generally achieved better performance than linear probing.

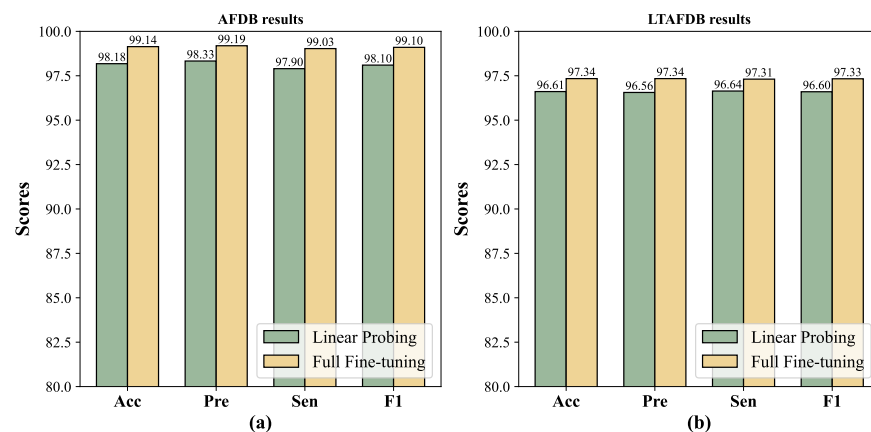


Figure 6. The comparison between linear probing and fully fine-tuning. (a) AFDB results. (b) LTAADB results.

4.4. Ablation Study

In this section, we investigate the necessity and effectiveness of the key components and settings in the proposed method for AF detection, including domain knowledge augmentation, multiple contrastive loss, and the design without a projector. The following ablation experiments all adopt the results of full fine-tuning.

4.4.1. Data Augmentation

The MLMCL method uses ECGs processed by domain knowledge augmentation, not just the raw ECG records. To evaluate the effectiveness of domain knowledge augmentation on AF detection, we separately evaluated the performance when the diagnostic region augmentation (vertical flipping) and non-diagnostic region augmentation (T-wave masking) were used alone. In addition, the following data augmentation methods commonly used for time series [35] were compared:

- Gaussian noise: A Gaussian noise signal $n(t)$ with a mean of 0 was added to the ECG sample $x(t)$ to obtain the augmented signal $\tilde{x}(t) = x(t) + n(t)$. A random standard deviation of noise in the range of $[0.01, 0.1]$ was adopted in the experiment.
- Baseline wander: A low-frequency sinusoidal wave with frequency f_w and maximum amplitude S_w were added to the ECG sample to achieve baseline wander. In the experiment, the range of f_w was $[1/30 \text{ Hz}, 1/10 \text{ Hz}]$, and the range of S_w was $[0.5, 0.6]$.
- Channel scaling: The ECG sample was multiplied by a scaling factor S to achieve scaling. A random scaling factor S in the range of $[0.1, 5]$ was tested in the experiment.
- Horizontal flipping: Horizontal flipping was considered as temporal reverse, which can be expressed as $\tilde{x}(t) = x(T_0 - t + 1)$, where $t = 1, 2, \dots, T_0$.
- Time warping: First, the ECG sample $x(t)$ was divided into w segments. Next, for each segment, half of the areas were randomly selected to stretch them by a factor of $r\%$ while squeezing the other half by the same amount. Finally, we connected the segments in the original order to generate the augmented sample, denoted as $\tilde{x}(t)$. In the experiment, $(w, r) = (1, 5)$ and $(w, r) = (3, 5)$ were used with the same probability.
- Random masking: It randomly sets $c\%$ of the ECG sample to a fixed value. In the experiment, the masking parameter was set to the typical value $c = 20$.

Each augmentation method was pre-trained under the same conditions as MLMCL. Table 5 shows the results of different data augmentation methods on external test sets. In all downstream tasks with a single data augmentation for pre-training, it can be seen that there are different suitable augmentation methods on different test sets. Generally, T-wave masking shows the best single augmentation result on AFDB, while vertical flipping obtains the second-best result. In the task of LTAFDB, vertical flipping showed the best result. As for domain knowledge augmentation, it achieved the best performance on AFDB. Although its performance on LTAFDB did not surpass vertical flipping, it also showed competitive performance. In other words, domain knowledge augmentation improved the quality of representations. This improvement may be due to the fact that T-wave masking and vertical flipping generate more diverse hard samples, which are sufficiently different from the original samples but retain the key diagnostic information. T-wave masking specifically modifies the non-diagnostic region, making the model pay more attention to the features of the P wave and QRS wave in the AF diagnostic region. Vertical flipping modifies the morphology of the P wave and QRS wave but does not affect the rhythm information. Existing studies [42] have shown that positive samples with large morphological differences but the same rhythm can better extract rhythm invariance. In addition, vertical flipping will not cause confusion between the P wave and f wave. These changes force the model to learn more discriminative and robust representations.

Table 5. Ablation study of data augmentation.

Data Augmentation	AFDB				LTAfDB			
	Acc (%)	Pre (%)	Sen (%)	F1 (%)	Acc (%)	Pre (%)	Sen (%)	F1 (%)
Gaussian Noise	96.56	95.90	95.57	95.73	96.89	96.85	96.91	96.88
Channel Scaling	96.04	95.56	94.60	95.08	97.40	97.38	97.39	97.38
Baseline Wander	97.55	97.72	96.19	96.95	96.32	96.26	96.36	96.31
Horizontal flipping	97.73	97.84	96.52	97.17	96.23	96.16	96.30	96.21
Time Warping	97.88	98.17	96.56	97.35	97.05	97.02	97.04	97.03
Random Masking	97.79	97.65	97.76	97.71	96.70	96.64	96.74	96.68
Vertical Flipping	98.82	98.54	98.54	98.54	97.44	97.43	97.43	97.43
T-wave Masking	99.10	99.14	98.99	99.10	96.85	96.80	96.88	96.84
MLMCL	99.14	99.19	99.03	99.10	97.34	97.34	97.31	97.33

4.4.2. Contrastive Loss

The ablation studies on different contrastive losses were evaluated on the AFDB and LTAfDB datasets. We examined the effectiveness of each contrastive loss and gradually combined each contrastive loss pairwise and finally combined all types of contrastive losses. Table 6 shows the ablation experiment results of six combination variants. Here, **T**, **C**, and **L** represent temporal contrastive loss, channel contrastive loss, and label contrastive loss, respectively. Among them, variant **T** only performs temporal contrastive learning on the intermediate hidden representations, thus updating only the convolutional feature extraction network in the encoder. The variant **L** indicates that only label CL is performed, so only labeled data are used and unlabeled data are not used. It should be noted that the equal λ value is used for different contrast losses across variants, whether labeled data or unlabeled data.

Table 6. Ablation study of multiple contrastive loss.

Methods	AFDB				LTAfDB			
	Acc (%)	Pre (%)	Sen (%)	F1 (%)	Acc (%)	Pre (%)	Sen (%)	F1 (%)
T	97.90	98.02	97.63	97.81	96.44	96.40	96.46	96.43
C	98.35	98.56	98.04	98.28	96.21	96.14	96.30	96.20
L	96.81	96.60	96.78	96.70	96.65	96.61	96.56	96.87
T + C	98.47	98.66	98.18	98.41	95.79	95.71	95.90	95.77
T + L	98.23	98.31	98.01	98.15	95.96	95.90	95.99	95.94
C + L	99.18	99.23	99.07	99.15	97.00	96.97	97.02	96.99
MLMCL	99.14	99.19	99.03	99.10	97.34	97.34	97.31	97.33

All variant experiments used domain knowledge augmentation and maintained the same training conditions as the MLMCL method. For the case of using only a single contrastive loss (**T**, **C**, or **L**), different optimal methods were obtained for different test sets, because each test task has a specific training dataset and data distribution. For the combination of two contrastive losses, it can be seen that on the AFDB database, the combinations of two contrastive losses achieved better Acc and F1 than using a single loss. On the LTAfDB database, only the **C + L** combination obtained better performance than using **C** or **L** alone. In addition, it is worth noting that the MLMCL method represented by the combination **T + C + L** achieved competitive or the best performance on all external test datasets, which means that the proposed MLMCL method is quite robust.

4.4.3. No Projecter

As shown in Figure 7, most CL-based methods employed a projector setup similar to SimCLR to achieve better performance for downstream tasks. They adopted a projector in pre-training to project the encoder representations into the latent space to calculate CL loss and discarded the projector in the downstream tasks. Differently, the proposed MLMCL

did not set an additional projector and simultaneously conducts contrastive learning on the encoder representations and the intermediate hidden representations. To explore the necessity of the projector in MLMCL, we conducted the ablation experiment of the projector. The projector was set as a two-layer MLP, and its output size was kept the same as the input size. Table 7 shows the AF detection results of the proposed method with and without a projector. It can be seen that having a projector did not achieve better performance on either the AFDB or LTAfDB database. Considering that the projector may reduce the performance, MLMCL did not use a projector for pre-training, which can also reduce the computational burden.

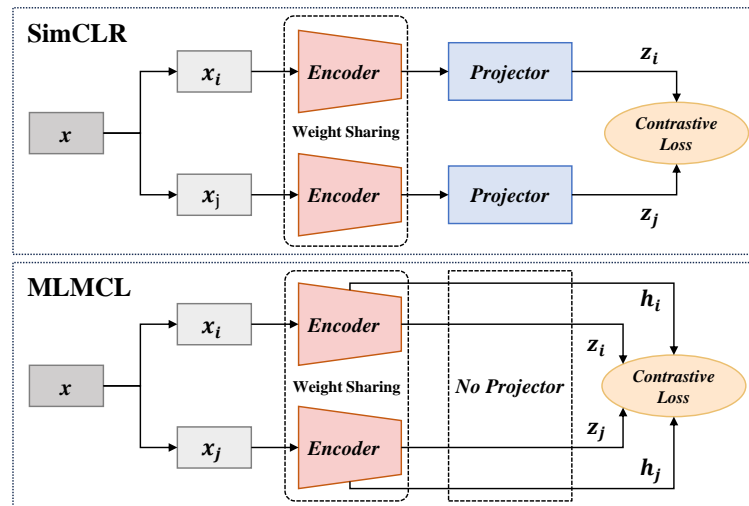


Figure 7. The comparison between MLMCL and SimCLR.

Table 7. Ablation study of the proposed method with/without a projector.

Test Set	Methods	Acc (%)	Pre (%)	Sen (%)	F1 (%)
AFDB	With a projector	99.04	99.09	98.92	99.00
	Without a projector	99.14	99.19	99.03	99.10
LTAfDB	With a projector	96.69	96.63	96.72	96.67
	Without a projector	97.34	97.34	97.31	97.33

4.5. Comparison with Existing Methods

In addition to the contrastive learning methods [36] for AF detection, there are also many studies on AF detection that adopted other strategies. However, few studies have reported cross-dataset test results, and some of them have used different training and testing datasets from those in this paper, which limits the number of studies available for comparison. Therefore, in order to achieve a good overall comparison with the cross-dataset evaluation results of other studies, we adjusted to use the same labeled training set and testing set for pre-training, fine-tuning, and testing. The comparison results with existing studies are shown in Table 8. Regarding the AF detection task, the method proposed in this paper outperformed most existing methods in different metrics, including Acc and F1. Compared to existing studies that only use labels for supervised learning, MLMCL simultaneously utilizes labels and the ECG inherent consistency to obtain higher-quality representations. In addition, we guided and optimized the pre-training according to the downstream AF detection task. We introduced domain knowledge augmentation for AF and used downstream labels for semi-supervised learning, which reduced the gap between the pre-trained representations and the requirements of downstream tasks, thereby achieving better performance.

Table 8. Comparison of the classification performance between the proposed method and previous works.

Study	Year	Training Set	Test Set	Acc (%)	Pre (%)	Sen (%)	F1 (%)
Andersen et al. [54]	2019	AFDB	MITDB	87.40	45.45	98.96	/
Shi et al. [55]	2020	AFDB	MITDB	87.4	81.11	97.46	/
Seo et al. [56]	2021	AFDB	MITDB	86.68	/	/	/
Liu et al. [57]	2022	AFDB	MITDB	92.23	53.92	95.17	68.84
MLMCL	2024	AFDB	MITDB	94.31	80.62	96.87	86.37
Wen et al. [58]	2022	CPSC2021	LTAfDB	/	94.60	89.50	91.98
Yun et al. [59]	2024	CPSC2021	LTAfDB	/	96.45	94.84	95.64
MLMCL	2024	CPSC2021	LTAfDB	97.34	97.34	97.31	97.33

5. Conclusions

In this work, we proposed a semi-supervised contrastive learning method, MLMCL, for AF detection, aiming to bridge the gap between the limited availability of labeled data in AF detection of single-lead ECG and the requirements for model robustness and generalization. MLMCL utilizes multi-level feature representations of the encoder to perform multiple contrastive learning, fully exploiting the temporal consistency, channel consistency, and label consistency. This method simultaneously uses both labeled and unlabeled data for pre-training to obtain the robust features required for downstream tasks, achieving better performance with limited labels and reducing the dependence on ECG labels. In addition, by utilizing the ECG diagnostic knowledge related to AF, a domain knowledge augmentation method was proposed to generate hard samples that would be sufficiently different from the original samples, enabling a full learning of distinguishable representations. We verified the stability and generalization of the MLMCL method on multiple different external data sets in a challenging cross-dataset testing environment. The abundant baseline settings and extensive ablation studies demonstrate that the proposed MLMCL effectively and robustly outperforms other existing AF detection methods. Our future work aims to test our method on more data sets to more fully examine its generalization ability. This method can be easily extended to the scenario of AF detection in multi-lead ECG signals and has reference significance for the development of other arrhythmia contrastive learning methods.

Author Contributions: Conceptualization, Y.Z.; methodology, Y.Z. and P.W.; software, Y.Z.; validation, Y.Z., P.W. and L.D.; formal analysis, Y.Z., L.D., X.C. and Z.L.; investigation, Y.Z. and P.W.; resources, J.S.; data curation, Y.Z. and P.W.; writing—original draft preparation, Y.Z.; writing—review and editing, L.D., X.C. and Z.L.; visualization, Y.Z.; supervision, J.S.; project administration, Z.F.; funding acquisition, Z.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (U21A20447 62371441 62331025 62401547) and CAMS Innovation Fund for Medical Sciences (2019-I2M-5-019).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Patient consent was waived because this work used public datasets.

Data Availability Statement: The CPSC2021, AFDB, LTAfDB, and MITDB datasets are publicly available at <https://physionet.org/content/cpsc2021/1.0.0/>, <https://physionet.org/content/afdb/1.0.0/>, <https://physionet.org/content/ltafdb/1.0.0/> and <https://physionet.org/content/mitdb/1.0.0/>, respectively

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Król-Józaga, B. Atrial fibrillation detection using convolutional neural networks on 2-dimensional representation of ECG signal. *Biomed. Signal Process. Control* **2022**, *74*, 103470. [[CrossRef](#)]
2. Wijesurendra, R.S.; Casadei, B. Mechanisms of atrial fibrillation. *Heart* **2019**, *105*, 1860–1867. [[CrossRef](#)] [[PubMed](#)]
3. Lin, H.J.; Wolf, P.A.; Kelly-Hayes, M.; Beiser, A.S.; Kase, C.S.; Benjamin, E.J.; D’Agostino, R.B. Stroke severity in atrial fibrillation: The Framingham Study. *Stroke* **1996**, *27*, 1760–1764. [[CrossRef](#)] [[PubMed](#)]
4. Papanastasiou, C.A.; Theochari, C.A.; Zareifopoulos, N.; Arfaras-Melainis, A.; Giannakoulas, G.; Karamitsos, T.D.; Palaiodimos, L.; Ntaios, G.; Avgerinos, K.I.; Kapogiannis, D. Atrial fibrillation is associated with cognitive impairment, all-cause dementia, vascular dementia, and Alzheimer’s disease: A systematic review and meta-analysis. *J. Gen. Intern. Med.* **2021**, *36*, 3122–3135. [[CrossRef](#)]
5. Vermond, R.A.; Geelhoed, B.; Verweij, N.; Tieleman, R.G.; Van der Harst, P.; Hillege, H.L.; Van Gilst, W.H.; Van Gelder, I.C.; Rienstra, M. Incidence of atrial fibrillation and relationship with cardiovascular events, heart failure, and mortality: A community-based study from the Netherlands. *J. Am. Coll. Cardiol.* **2015**, *66*, 1000–1007. [[CrossRef](#)] [[PubMed](#)]
6. Soliman, E.Z.; Safford, M.M.; Muntner, P.; Khodneva, Y.; Dawood, F.Z.; Zakai, N.A.; Thacker, E.L.; Judd, S.; Howard, V.J.; Howard, G. Atrial fibrillation and the risk of myocardial infarction. *JAMA Intern. Med.* **2014**, *174*, 107–114. [[CrossRef](#)] [[PubMed](#)]
7. Rattanawong, P.; Upala, S.; Riangwiwat, T.; Jaruvongvanich, V.; Sanguankeo, A.; Vutthikraivit, W.; Chung, E.H. Atrial fibrillation is associated with sudden cardiac death: A systematic review and meta-analysis. *J. Interv. Card. Electrophysiol.* **2018**, *51*, 91–104. [[CrossRef](#)]
8. Jones, S.A. *ECG Notes: Interpretation and Management Guide*; FA Davis: Philadelphia, PA, USA, 2021.
9. Ryden, L.; Asinger, R.; Cannom, D.; Crijns, H.; Frye, R.; Halperin, J.; Kay, G.; Klein, W.; Levy, S.; McNamara, R. ACC/AHA/ESC guidelines for the management of patients with atrial fibrillation: Executive summary—A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines and Policy Conferences (Committee to Develop Guidelines for the Management of Patients with Atrial Fibrillation). *J. Am. Coll. Cardiol.* **2001**, *38*, 1231–1265.
10. Čihák, R.; Haman, L.; Táborský, M. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Cor Et Vasa* **2016**, *6*, e636–e683. [[CrossRef](#)]
11. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **2019**, *25*, 65–69. [[CrossRef](#)] [[PubMed](#)]
12. Fan, X.; Yao, Q.; Cai, Y.; Miao, F.; Sun, F.; Li, Y. Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1744–1753. [[CrossRef](#)] [[PubMed](#)]
13. Zhang, P.; Ma, C.; Sun, Y.; Fan, G.; Song, F.; Feng, Y.; Zhang, G. Global hybrid multi-scale convolutional network for accurate and robust detection of atrial fibrillation using single-lead ECG recordings. *Comput. Biol. Med.* **2021**, *139*, 104880. [[CrossRef](#)] [[PubMed](#)]
14. Jin, Y.; Qin, C.; Huang, Y.; Zhao, W.; Liu, C. Multi-domain modeling of atrial fibrillation detection with twin attentional convolutional long short-term memory neural networks. *Knowl.-Based Syst.* **2020**, *193*, 105460. [[CrossRef](#)]
15. Petmezas, G.; Haris, K.; Stefanopoulos, L.; Kilintzis, V.; Tzavelis, A.; Rogers, J.A.; Katsaggelos, A.K.; Maglaveras, N. Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets. *Biomed. Signal Process. Control* **2021**, *63*, 102194. [[CrossRef](#)]
16. Zhang, P.; Ma, C.; Song, F.; Sun, Y.; Feng, Y.; He, Y.; Zhang, T.; Zhang, G. D2AFNet: A dual-domain attention cascade network for accurate and interpretable atrial fibrillation detection. *Biomed. Signal Process. Control* **2023**, *82*, 104615. [[CrossRef](#)]
17. Prabhakararao, E.; Dandapat, S. Atrial Fibrillation Burden Estimation Using Multi-Task Deep Convolutional Neural Network. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 5992–6002. [[CrossRef](#)] [[PubMed](#)]
18. Yan, G.; Liang, S.; Zhang, Y.; Liu, F. Fusing transformer model with temporal features for ECG heartbeat classification. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 898–905. [[CrossRef](#)]
19. Yang, M.U.; Lee, D.I.; Park, S. Automated diagnosis of atrial fibrillation using ECG component-aware transformer. *Comput. Biol. Med.* **2022**, *150*, 106115. [[CrossRef](#)]
20. Wang, J. A deep learning approach for atrial fibrillation signals classification based on convolutional and modified Elman neural network. *Future Gener. Comput. Syst.* **2020**, *102*, 670–679. [[CrossRef](#)]
21. Limam, M.; Precioso, F. Atrial fibrillation detection and ECG classification based on convolutional recurrent neural network. In Proceedings of the 2017 Computing in Cardiology (CinC), Rennes, France, 24–27 September 2017; pp. 1–4. [[CrossRef](#)]
22. 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, W., Jr. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat. Commun.* **2020**, *11*, 1760. [[CrossRef](#)]
23. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.

24. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Rennes, France, 24–27 September 2017; pp. 1597–1607.
25. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9729–9738.
26. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M. Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
27. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
28. Yang, X.; Zhang, Z.; Cui, R. Timeclr: A self-supervised contrastive learning framework for univariate time series representation. *Knowl.-Based Syst.* **2022**, *245*, 108606. [[CrossRef](#)]
29. Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; Xu, B. Ts2vec: Towards universal representation of time series. *Proc. Aaai Conf. Artif. Intell.* **2022**, *36*, 8980–8987. [[CrossRef](#)]
30. Tonekaboni, S.; Eytan, D.; Goldenberg, A. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv* **2021**, arXiv:2106.00750.
31. Kiyasseh, D.; Zhu, T.; Clifton, D.A. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In Proceedings of the International Conference on Machine Learning; PMLR: New York, NY, USA, 2021; pp. 5606–5615.
32. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6827–6839.
33. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XI 16; Springer: Berlin, Germany, 2020; pp. 776–794. [[CrossRef](#)]
34. Mehari, T.; Strodtzoff, N. Self-supervised representation learning from 12-lead ECG data. *Comput. Biol. Med.* **2022**, *141*, 105114. [[CrossRef](#)] [[PubMed](#)]
35. Soltanieh, S.; Hashemi, J.; Etemad, A. In-distribution and out-of-distribution self-supervised ecg representation learning for arrhythmia detection. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 789–800. [[CrossRef](#)] [[PubMed](#)]
36. Zhang, W.; Geng, S.; Hong, S. A simple self-supervised ECG representation learning method via manipulated temporal–spatial reverse detection. *Biomed. Signal Process. Control* **2023**, *79*, 104194. [[CrossRef](#)]
37. Cheng, J.Y.; Goh, H.; Dogrusoz, K.; Tuzel, O.; Azemi, E. Subject-aware contrastive learning for biosignals. *arXiv* **2020**, arXiv:2007.04871.
38. Lan, X.; Ng, D.; Hong, S.; Feng, M. Intra-inter subject self-supervised learning for multivariate cardiac signals. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 4532–4540. [[CrossRef](#)]
39. Liu, W.; Li, Z.; Zhang, H.; Chang, S.; Wang, H.; He, J.; Huang, Q. Dense lead contrast for self-supervised representation learning of multilead electrocardiograms. *Inf. Sci.* **2023**, *634*, 189–205. [[CrossRef](#)]
40. Liu, W.; Pan, S.; Chang, S.; Huang, Q.; Jiang, N. Direct Lead Assignment: A Simple and Scalable Contrastive Learning Method for ECG and Its IoMT Applications. *IEEE Internet Things J.* **2024**. [[CrossRef](#)]
41. Gopal, B.; Han, R.; Raghupathi, G.; Ng, A.; Tison, G.; Rajpurkar, P. 3KG: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In Proceedings of the Machine Learning for Health; PMLR: New York, NY, USA, 2021; pp. 156–167.
42. Liu, W.; Zhang, H.; Chang, S.; Wang, H.; He, J.; Huang, Q. Learning Representations for Multi-Lead Electrocardiograms from Morphology-Rhythm Contrast. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2509615. [[CrossRef](#)]
43. Zhu, X.; Shi, M.; Yu, X.; Liu, C.; Lian, X.; Fei, J.; Luo, J.; Jin, X.; Zhang, P.; Ji, X. Self-supervised inter–intra period-aware ECG representation learning for detecting atrial fibrillation. *Biomed. Signal Process. Control* **2025**, *100*, 106939. [[CrossRef](#)]
44. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
45. Le, D.; Truong, S.; Brijesh, P.; Adjeroh, D.A.; Le, N. scl-st: Supervised contrastive learning with semantic transformations for multiple lead ecg arrhythmia classification. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2818–2828. [[CrossRef](#)]
46. Zhou, S.; Huang, X.; Liu, N.; Zhang, W.; Zhang, Y.T.; Chung, F.L. Open-world electrocardiogram classification via domain knowledge-driven contrastive learning. *Neural Netw.* **2024**, *179*, 106551. [[CrossRef](#)]
47. Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwok, C.K.; Li, X.; Guan, C. Time-series representation learning via temporal and contextual contrasting. *arXiv* **2021**, arXiv:2106.14112.
48. Sangha, V.; Mortazavi, B.J.; Haimovich, A.D.; Ribeiro, A.H.; Brandt, C.A.; Jacoby, D.L.; Schulz, W.L.; Krumholz, H.M.; Ribeiro, A.L.P.; Khera, R. Automated multilabel diagnosis on electrocardiographic images and signals. *Nat. Commun.* **2022**, *13*, 1583. [[CrossRef](#)] [[PubMed](#)]
49. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [[CrossRef](#)]

50. Moody, G. A new method for detecting atrial fibrillation using RR intervals. *Proc. Comput. Cardiol.* **1983**, *10*, 227–230.
51. Wang, X.; Ma, C.; Zhang, X.; Gao, H.; Clifford, G.; Liu, C. Paroxysmal atrial fibrillation events detection from dynamic ECG recordings: The 4th China physiological signal challenge 2021. *PhysioNet* **2021**, 1–83. [[CrossRef](#)]
52. Petrutiu, S.; Sahakian, A.V.; Swiryn, S. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *Europace* **2007**, *9*, 466–470. [[CrossRef](#)] [[PubMed](#)]
53. Zou, Y.; Yu, X.; Li, S.; Mou, X.; Du, L.; Chen, X.; Li, Z.; Wang, P.; Li, X.; Du, M. A generalizable and robust deep learning method for atrial fibrillation detection from long-term electrocardiogram. *Biomed. Signal Process. Control* **2024**, *90*, 105797. [[CrossRef](#)]
54. Andersen, R.S.; Peimankar, A.; Puthusserypady, S. A deep learning approach for real-time detection of atrial fibrillation. *Expert Syst. Appl.* **2019**, *115*, 465–473. [[CrossRef](#)]
55. Shi, H.; Wang, H.; Qin, C.; Zhao, L.; Liu, C. An incremental learning system for atrial fibrillation detection based on transfer learning and active learning. *Comput. Methods Programs Biomed.* **2020**, *187*, 105219. [[CrossRef](#)] [[PubMed](#)]
56. Seo, H.C.; Oh, S.; Kim, H.; Joo, S. ECG data dependency for atrial fibrillation detection based on residual networks. *Sci. Rep.* **2021**, *11*, 18256. [[CrossRef](#)]
57. Liu, S.; Wang, A.; Deng, X.; Yang, C. MGNN: A multiscale grouped convolutional neural network for efficient atrial fibrillation detection. *Comput. Biol. Med.* **2022**, *148*, 105863. [[CrossRef](#)]
58. Wen, H.; Kang, J. A comparative study on neural networks for paroxysmal atrial fibrillation events detection from electrocardiography. *J. Electrocardiol.* **2022**, *75*, 19–27. [[CrossRef](#)] [[PubMed](#)]
59. Yun, D.; Yang, H.L.; Kwon, S.; Lee, S.R.; Kim, K.; Kim, K.; Lee, H.C.; Jung, C.W.; Kim, Y.S.; Han, S.S. Automatic segmentation of atrial fibrillation and flutter in single-lead electrocardiograms by self-supervised learning and Transformer architecture. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 79–88. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.