ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

Classification of Electrocardiogram (ECG) Signals Based on Self-Supervised Learning

Chunyan Liu[1,a], Xuande Zhang[1,b,*], Xin Huang[2,c,*], Long Xu[2,d]

[1]*School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, China*
[2]*School of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China*
[a]*15297722274@163.com,* [b]*love_truth@126.com,* [c]*huangxin@nbu.edu.cn,* [d]*xulong1@nbu.edu.cn*
[*]*Corresponding author*

Abstract: *Electrocardiogram (ECG) analysis based on deep learning models has garnered significant research interest in recent years. Nevertheless, the performance of such models is often constrained by the limited availability of annotated ECG data.This paper proposes a novel self-supervised learning framework for ECG signal classification. Our method combines augmented contrastive learning with ECG-specific temporal augmentations (time truncation and random resized cropping). Experiments conducted on datasets (Cinc2020, Chapman, and Ribeiro) demonstrate that our approach achieves an average accuracy of 89.0%, an average AUC of 77.3%, and an F1-score of 63.3%. This represents improvements of 4.4%, 7.2%, and 3.8% in accuracy, AUC, and F1-score, respectively. When using only 60% of the labeled data, our method outperforms the fully supervised baseline by 7.7%. Ablation studies validate the effectiveness of our data augmentation strategies and contrastive learning design. Our approach offers a promising solution for label-efficient ECG analysis, with potential applications in clinical screening and remote monitoring systems.*

Keywords: *ECG, Self-Supervised Learning, Electrocardiogram Signal Classification, Contrastive Learning, Residual Network*

1. Introduction

Electrocardiography (ECG) is a safe, non-invasive, inexpensive, and simple-to-operate preliminary clinical diagnostic tool used for monitoring cardiac electrical activity and related information. It depicts the changes in body surface potential caused by the human heartbeat, which contain a wealth of physiological information and can serve as quantitative indicators reflecting the physiological and pathological states of the heart. Clinicians can assist in diagnosing cardiomyopathies by examining characteristic changes in ECG features such as P-waves, QRS complexes, and ST segments. However, manually identifying these subtle ECG variations typically requires sufficient experience and is time-consuming, with interpretation efficiency and consistency prone to subjective influences. To address this, deep learning algorithms have emerged, capable of accurately processing vast amounts of data in a short time, thereby helping to improve diagnostic accuracy and efficiency for physicians.

The performance of early machine learning methods depended on feature extraction and selection. Yet, extracting effective features is highly challenging. In contrast, deep learning can automatically extract salient features from high-dimensional raw data. Following the successful application of deep learning in medical imaging, deep neural networks such as Convolutional Neural Networks (CNNs)[1], Recurrent Neural Networks (RNNs)[2], and Long Short-Term Memory networks (LSTMs)[3]have also been employed for ECG signal analysis.

In recent years, deep learning models have seen widespread application in ECG diagnosis due to their strong feature extraction and classification capabilities. However, in deep learning, obtaining high-quality labeled datasets often presents a significant challenge. Considerable time and effort are required to annotate data, and the annotation process itself is not trivial. Particularly in the medical field, defining clinical ground truth is often difficult even for physicians. In the real world, data imbalance is another major issue. The number of patients with specific diseases or clinical symptoms is highly imbalanced; the incidence of some abnormal ECG patterns is very low, while normal ECGs constitute the majority of ECG results. Additionally, the volume of unlabeled data typically exceeds that of labeled data by several orders of magnitude. Due to the aforementioned points, the development of supervised deep learning

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

methods in the ECG field has been constrained.Over the past few years, self-supervised learning has made significant strides in fields such as Natural Language Processing (NLP)[4], Speech Recognition[1][5], and Computer Vision (CV)[2][3]. Among these, contrastive self-supervised learning is a widely adopted strategy that has demonstrated powerful learning capabilities in the ECG domain. Introducing self-supervised methods from these fields for ECG diagnosis may be a viable path. Examples include contrastive learning methods from Computer Vision (CV)[2], masked representation learning methods from Natural Language Processing (NLP), and contrastive predictive coding methods from speech processing[6]. Models are first pre-trained on large unlabeled datasets and then fine-tuned on smaller, task-specific labeled datasets. Self-supervised learning methods have been proven capable of learning effective representations for various downstream tasks from non-human-annotated datasets. In the ECG field, self-supervised contrastive methods can be broadly categorized into three types: The first type is Siamese networks based on negative sample pairs, including SimCLR[2] and MoCo v2[7]. The second type is Siamese networks that do not rely on negative sample pairs, where BYOL[8] primarily uses momentum updates, and SimSiam uses stop-gradient. SwAV[9] uses clustering to prevent model collapse. The last type is the auto-regressive-based CPC[6].

SimCLR holds distinct advantages in the field of self-supervised learning due to its simple yet effective contrastive learning framework and strong data augmentation strategies. This paper proposes a multi-classification model for ECG signals based on the Simple framework for Contrastive Learning of Visual Representations (SimCLR). Leveraging the unique characteristics of ECG signals, such as their one-dimensionality and periodicity, we specifically design data augmentation methods tailored for ECG signals. We employ two data augmentation methods: time-out and Randomly Resized Cropping (RRC). The encoder uses a 1D-ResNet50 for experiments, and comparisons are made with fully supervised methods. The experimental results indicate that the proposed method outperforms fully supervised methods while reducing the reliance on labeled data.

2. Related Work

2.1. The Application of Deep Learning in Electrocardiogram (Ecg) Signal Classification

Cardiovascular disease (CVD) is one of the leading causes of human mortality, with approximately 85% of deaths attributable to heart attacks[10]. Most patients develop arrhythmias before being diagnosed with CVD. Therefore, early screening and diagnosis using electrocardiograms are of great significance for identifying potential cardiac pathologies, improving patient prognosis, and reducing mortality rates. According to the standards set by the American Association for Medical Instrumentation (AAMI), heartbeats are classified into five major categories: N, S, V, F, and Q, with each major category further divided into several subcategories. In the commonly used MIT-BIH database, there are a total of 18 label values, comprising one normal heart rate and 17 abnormal heart rates. The AHA database includes eight major categories of arrhythmias. The input to neural networks typically consists of electrocardiogram signals in various forms, most commonly as one-dimensional raw ECG signals[11][12][13][14][15], but also as two-dimensional images[16][17] (grayscale images of the signals) or spectrograms[18][19].

Georgios et al.[15] proposed a hybrid CNN-LSTM network for detecting atrial fibrillation on an imbalanced dataset, categorizing electrocardiogram rhythms into four classes: normal (N), atrial fibrillation (AFIB), atrial flutter (AFL), and AV junctional rhythm (J). The model achieved a sensitivity of 97.87% and a specificity of 99.29%. Compared to all previous related studies, this model provided an optimal combination of performance and demonstrated robust effectiveness even in highly imbalanced datasets.

Ullah et al.[17] proposed a robust algorithm by converting one-dimensional data into two-dimensional (2D) images, enabling accurate classification of ECG signals even in the presence of environmental noise. The entire process consists of four steps: signal processing, 2D image generation, data augmentation, feature extraction from the data, and classification. The proposed 2D model achieved a classification accuracy of 99.02%.

Chen et al.[20] proposed a bidirectional RNN model incorporating an attention mechanism, composed of five convolutional blocks. This model achieved state-of-the-art results in the classification and detection of heart rhythm disorders, securing first place in the CPSC2018 competition. Addressing the slow progress in the field of automated ECG analysis—attributed to the lack of adequate training datasets and standardized evaluation procedures to ensure comparability among different algorithms—Wagner et al.[21] employed several classic time-series classification models on the PTB-XL dataset, such as full

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

CNNs, one-dimensional deep residual networks, InceptionTime, and bidirectional LSTM networks, to perform ECG classification tasks and achieved favorable classification accuracy.

Huang et al.[22] transformed time-domain ECG data into time-frequency spectrograms and subsequently applied 2D CNNs for classification, attaining an average accuracy of 99.00% and demonstrating high classification precision. In comparison, traditional 1D CNN models achieved an average accuracy of 90.93%, validating that the proposed CNN classifier using ECG spectrograms as input can achieve higher classification accuracy without requiring additional manual operations.

2.2. The Application of Self-Supervised Learning in Electrocardiogram (ECG) Signal Classification

In recent years, self-supervised learning models have been widely applied to electrocardiogram (ECG) representation learning, with contrastive learning-based methods dominating the field. Their performance often relies on carefully designed data augmentation strategies. Such methods typically require the model to distinguish similarities between different segments of the same ECG signal or between different signals, thereby learning discriminative ECG features.Beyond contrastive learning, other self-supervised learning approaches have also demonstrated value in the ECG domain. Temporal reconstruction tasks require models to restore the temporal order of ECG signal segments. By training models to identify or reorganize scrambled signal segments, these tasks enhance the ability to model temporal dependencies, improving the quality of feature representations and the accuracy of signal reconstruction. Adversarial training leverages the generative adversarial network framework, where the dynamic interplay between the generator and discriminator deepens the model's understanding of ECG data distribution during synthesis and discrimination, thereby strengthening the robustness of the learned representations. Additionally, autoencoders utilize an encoder-decoder architecture to learn low-dimensional compressed representations of ECG signals. By reconstructing the input signals, they automatically extract key features, achieving effective dimensionality reduction and feature learning.The common advantage of these methods lies in their ability to automatically learn meaningful representations by uncovering the intrinsic structure of data, even in scenarios with limited labeled data. This provides more reliable initialization features for training models in ECG analysis tasks, ultimately enhancing the performance and generalization capability of downstream diagnostic tasks.

Han Liu et al.[23] designed two ECG-specific data augmentation methods—random baseline drift and random high-frequency interference—and integrated them with six mainstream contrastive self-supervised learning approaches for ECG morphology recognition. The downstream task involved an eight-class multi-label ECG classification task. Experiments demonstrated that self-supervised pre-training relying on negative sample pairs can achieve ECG representations significantly superior to those of baseline networks. With only a limited number of labeled samples, these methods were able to approach the performance level of human experts, substantially reducing the demand for labeled data. However, due to hardware limitations in the experiments, the mini-batch size could only reach 2049. The lack of negative sample pairs lowered the difficulty of contrastive learning, preventing the SimCLR model from generalizing sufficient features to distinguish between positive and negative samples, thereby constraining the model's performance.

Shunxiang Yang et al.[24] proposed a masked self-supervised learning model based on the multi-view information bottleneck principle. This method applies high-ratio masking to ECG signal instances in both the time and frequency domains, followed by reconstructing the original input using an autoencoder. The model was pre-trained simultaneously on three datasets—PTB-XL, CPSC, and Chapman—and subsequently fine-tuned for each ECG classification task. Control groups included supervised learning methods and previous self-supervised learning approaches. Experiments demonstrated that the proposed model not only outperforms state-of-the-art self-supervised learning models but also surpasses supervised learning models. However, a limitation of this method lies in its architecture: the model incorporates both a time-domain view autoencoder and a frequency-domain view autoencoder, and additional dimensionality expansion of some output features is introduced to align the two views, resulting in a large number of parameters.

Temesgen et al.[25] applied the Contrastive Predictive Coding (CPC) method to ECG diagnosis. In the first step, contrastive representations were learned, and their quality was evaluated based on the linear evaluation performance for comprehensive clinical ECG classification tasks. In the second step, the impact of self-supervised pre-training on fine-tuning ECG classifiers was analyzed in comparison to purely supervised performance. The results showed that the self-supervised representations achieved scores only 0.5% lower than supervised performance during linear evaluation, while improving supervised performance by 1.0% during fine-tuning. This fully demonstrates the effectiveness of self-

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

supervised learning in the field of ECG analysis.

Jiewei et al.[26] proposed a self-supervised learning classification framework capable of identifying 60 ECG diagnostic terms based on a large-scale dataset. On the offline test set, the model achieved an average AUROC of 0.975, an average AUPRC of 0.646, and an average F1-score of 0.575. On the online test set, it achieved an average sensitivity of 0.736, an average specificity of 0.954, and an average F1-score of 0.468. The distribution of ECG diagnostic terms in the online test set better reflects real-world scenarios and includes various artifacts, indicating that although the model's performance declined in practical applications, it still maintained acceptable sensitivity and specificity levels. During the pre-training phase, the Siamese convolutional network was pre-trained using Momentum Contrast (MoCo), and the learned weights were then used as the initialization weights for the downstream classification network.

Chuankai Luo et al.[27] proposed a novel self-supervised pre-training method called Segment Origin Prediction (SOP) to enhance model performance in arrhythmia classification. To validate the general applicability of the SOP method, six widely used and advanced models were tested. In the downstream task phase, labeled data were denoised and randomly cropped to obtain data xi, which was then input into the DNN model. The pre-trained model weights were used as the initialization weights for the downstream model. Three sets of experiments were conducted. The first set of experiments showed that models using the SOP method achieved higher F1-scores. The second set further improved model performance on the basis of the first set. Specifically, after applying SOP, the F1-score of the original state-of-the-art model increased from 0.852 to 0.863, and with the introduction of external data, it further improved to 0.875. The third set of results indicated that adding a feedforward layer during pre-training enhanced the classification performance of the model.

3. Methodology and Model

3.1. Dataset

The model was pre-trained using a combined dataset comprising three sources: CinC2020[28], Chapman[29], and Ribeiro[13], totaling 54,566 records. Among these, CinC2020 includes the PTB-XL dataset, which is annotated with 71 labels (based on the SCP-ECG standard). These labels cover a wide range of diagnoses, morphological statements, and rhythm statements. The 44 diagnostic statements can be grouped into five superclasses (Normal/Conduction Disturbance/Myocardial Infarction/Hypertrophy/ST-T Changes). The 19 form statements mainly describe morphological variations in specific ECG segments, such as abnormal QRS complexes, while the 21 rhythm statements include characterizations of normal heart rhythms as well as arrhythmia statements. The overall dataset was split into ten stratified and label-balanced folds. During pre-training, all data were used, and model performance was validated using the PTB-XL subset. The dataset composition is shown in Table 1.

Table 1: ECG Datasets.

dataset	# samples	# patients
Pretraining	54,566	unknown
-Cinc2020	43,093	unknown
-Chapman	10,646	10,646
-Ribeiro	827	827
Evaluation	21,837	18,885
-PTB-XL	21,837	18,885

3.2. Data Augmentation Methods

3.2.1. Time out(TO)

TimeOut (also known as Time Masking or Temporal Masking) is a data augmentation technique for time series signals. Its core principle involves randomly masking or truncating local temporal segments of the signal, thereby enabling the model to learn robust temporal features. The key parameter in Time Out is t, which is uniformly sampled from the interval $(t_l\text{-}t_u)$. In our experiments, $(t_l\text{-}t_u)$ is set to (0.0, 0.5).

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

3.2.2. Randomly Resized Cropping (RRC)

Randomly Resized Cropping (RRC) involves continuously cropping an ECG waveform at a random length and then resizing it back to its original dimensions. The core function of RRC is to generate augmented versions of each original sample with varying cropping regions and length scaling. In our method, the cropping length p is uniformly sampled from the interval (l, m). In our experiments, we set (l, m) = (0.6, 1.0), meaning the signal is cropped to segments ranging from 60% to 100% of its original length.

3.2.3. Combination of Time Out and Randomly Resized Cropping (TO-RRC)

The two methods mentioned above are combined, with particular attention to the sequence of operations: for each ECG signal, the Randomly Resized Cropping (RRC) is applied first, followed by the Time Out operation. This order is crucial because if local intervals are masked first and cropping is performed afterward, the resulting cropped segment might only contain masked regions, leading to the loss of effective features.

3.3. Model Architecture

In self-supervised learning, contrastive learning frameworks are commonly employed. These frameworks leverage large amounts of unlabeled data for pre-training through contrastive learning, followed by fine-tuning on labeled data for downstream tasks. This approach enables the training of high-precision models with limited labeled data. In this experiment, SimCLR is used as the self-supervised contrastive learning framework for the pre-training phase, with a one-dimensional ResNet-50 model serving as the encoder. The contrastive loss of the framework is subsequently recalculated to optimize the model.

The typical training process for augmented contrastive learning is as follows: First, data augmentation is applied to the ECG signal x_i to obtain an augmented sample $x_i^{'}$. Then, the original samples x_i, x_j, and the augmented sample $x_i^{'}$ are input into the encoder $f(\cdot)$ to extract feature embeddings for each sample.

$$h_i=f(x_i,\theta) \tag{1}$$

$$h_j=f(x_j,\theta) \tag{2}$$

$$h_{i'}=f(x_i^{'},\theta) \tag{3}$$

Where h_i represents the feature embedding of the original ECG signal x_i, h_j represents the feature embedding of the original ECG signal x_j, and $h_{i'}$ represents the feature embedding of the augmented sample $x_i^{'}$. Subsequently, a contrastive loss is applied to measure the relative similarity between these feature embeddings.

SimCLR, proposed by Google in 2020, is a simple yet efficient self-supervised contrastive learning framework. Its core idea is to maximize the correlation between representations of similar samples in the embedding space while minimizing the correlation between representations of dissimilar samples. The fundamental principle of the SimCLR framework is illustrated in Figure 1.

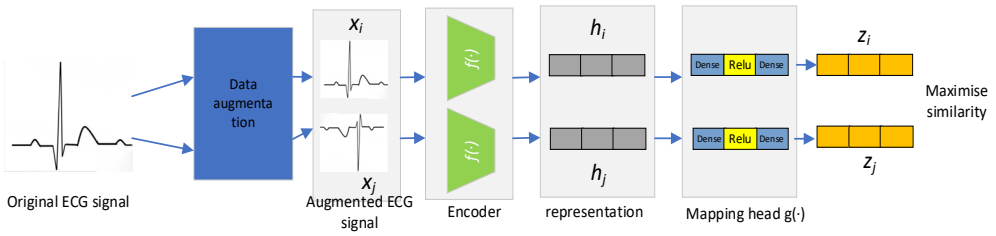

Figure 1: SimCLR Framework Schematic Diagram.

During the pre-training phase, for a given sample x, augmented samples x_i and x_j are generated through Time Out (TO) and Randomly Resized Cropping (RRC), respectively. These two augmented samples are treated as a positive pair in the contrastive learning framework. Subsequently, they are processed by an encoder (a 1D-ResNet50 is used in this experiment) to obtain the data representation vectors h_i and h_j. To prevent the learned features from being suboptimal for downstream tasks, SimCLR further applies a nonlinear projection to these vectors, resulting in the final representation

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

vectors z_i and z_j.

The optimization objective of SimCLR is to bring the vectors z_i and z_j as close as possible in the feature space, while pushing the output vectors of different samples as far apart as possible. The closeness or distance is determined by the similarity between vectors: more similar vectors are brought closer together, while less similar vectors are pushed further apart. In the SimCLR framework, similarity is measured by computing the cosine similarity between two vectors, as shown in Formula (4).

$$SIM(u,v)=\frac{u\cdot v}{||u||\times||v||}=\frac{\sum_{i=1}^{n}(u_i\times v_i)}{\sqrt{\sum_{i=1}^{n}u_i^2}\times\sqrt{\sum_{i=1}^{n}v_i^2}} \tag{4}$$

In the formula, u and v represent two vectors, $\cdot$ denotes the dot product operation, and $||u||$ and $||v||$ represent the norms of the two vectors, respectively. When the cosine similarity approaches 1, it indicates that the directions of the two vectors are similar, implying high similarity. When the cosine similarity approaches -1, it indicates that the directions of the two vectors are opposite, implying low similarity.

The SimCLR framework uses the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss as its objective function. For a pair of positive samples, the calculation formula is shown in Equation (5).

$$l_{Z_x,Z_{x'}}=-log\frac{exp(SIM(Z_x,Z_{x'})/\tau)}{\sum_{y=1,y\neq x}^{2N}exp(SIM(Z_x,Z_y)/\tau)} \tag{5}$$

Where N represents the number of randomly selected samples that form a mini-batch (referred to as a batch hereafter). All samples in this batch are augmented to generate a set of 2N samples. For a given anchor sample Z_x, there is only one positive sample $Z_{x'}$ within the batch, while the remaining 2(N-1) samples serve as negative samples. τ is a temperature parameter used for scaling. The final loss function is computed over all positive pairs in the batch, including both$(Z_x,Z_{x'})$and$(Z_{x'},Z_x)$.

During the fine-tuning phase for downstream tasks, the parameters of the encoder $f(\cdot)$ are fixed as a feature extractor, and the projection head is replaced with a structure tailored to the downstream task,such as a classification head for the ECG classification task in this study. Through these steps, a high-performing model can be trained using only a limited amount of labeled data.

3.4. Overall Framework

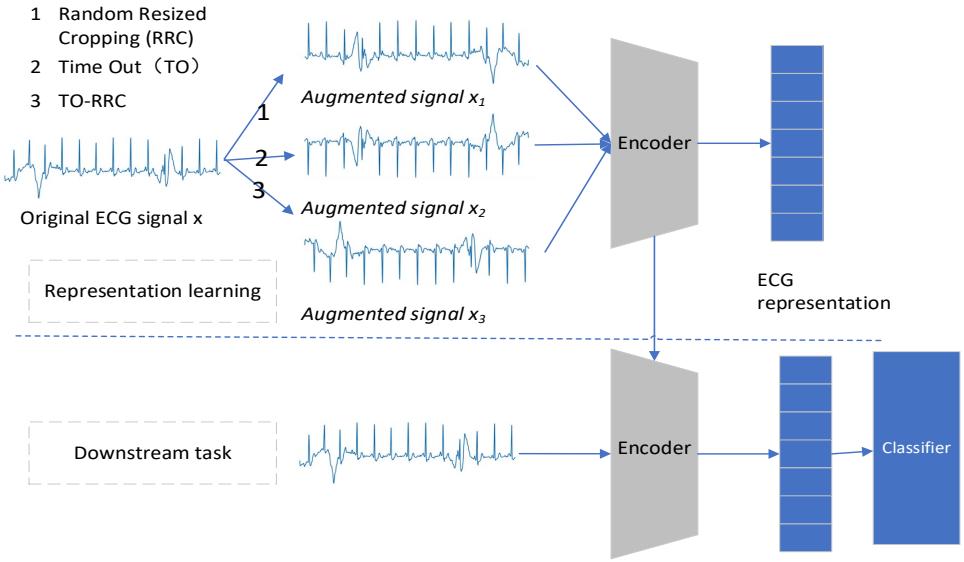

Figure 2: Overview of our two-stage framework.

Our two-stage framework is illustrated in Figure 2.The overall framework comprises two key phases. The first phase is self-supervised pre-training: during this stage, contrastive learning is employed to pre-train the encoder, enabling it to learn discriminative and generalizable representations from large-scale unlabeled electrocardiogram (ECG) data. The second phase is supervised fine-tuning: after pre-training, the trained encoder is transferred as a feature extractor to downstream tasks, typically followed by a

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

trainable classifier such as a fully connected layer. In this study, the framework is specifically applied to ECG signal classification—leveraging the representations learned in the first phase to fine-tune the model with a limited amount of labeled data, thereby achieving efficient and accurate automatic classification of ECG signals. The entire process demonstrates the advantages of the "pre-training and fine-tuning" paradigm, significantly reducing reliance on labeled data while enhancing the performance and robustness of the model on downstream tasks.

4. Experiments

4.1. Baseline Model

To evaluate the performance of the proposed method, four representative representation learning methods were selected as baselines for comparison, including Principal Component Analysis (PCA)[30], Random Projection (RP)[31], Autoencoder (AE)[32], and the method proposed by Oh et al.[33].

The proposed method was compared with the four baseline methods listed in Table 2 on the same downstream task (arrhythmia classification). Experimental results demonstrate that the proposed method consistently outperforms all baselines across all evaluation metrics. Specifically, compared to the best-performing baseline method, our approach achieves improvements of 4.4% in accuracy (ACC), 7.2% in the area under the receiver operating characteristic curve (AUC), and 3.8% in the F1-score. These results validate that the proposed pre-training strategy provides superior initialization weights for the model, enabling the encoder to learn representations embedded with prior knowledge of ECG signals, thereby significantly enhancing the performance of downstream classification tasks.

Table 2: Experimental results of baseline methods and our method.

Method	Classfication task		
	Ave ACC	Ave AUC	Ave F1
PCA[30]	0.781	0.579	0.467
RP[31]	0.610	0.478	0.388
AE[32]	0.829	0.668	0.544
Oh et al[33]	0.846	0.701	0.595
ours	0.890	0.773	0.633

4.2. Ablation Study

To separately evaluate the contributions of the two data augmentation strategies, Time Out (TO) and Randomly Resized Cropping (RRC), to representation learning, the following ablation experiments were designed.

4.2.1. Ablation with TO Only

Since Time Out is inherently an augmentation method that generates positive sample pairs, and the core of the NT-Xent loss lies in learning features by distinguishing between "different augmented views of the same signal (positive samples)" and "augmented views of different signals (negative samples)," the NT-Xent loss function was selected for this experiment.

4.2.2. Ablation with RRC Only

When applying RRC to ECG signals, it was adapted as 1D-RRC. The cropping ratio was set to 0.6–1.0 times the original signal length to avoid excessive shortening that could lead to the loss of critical features such as the QRS complex. After cropping, the signal was interpolated and rescaled back to the original length to ensure consistent input dimensions for the encoder. Cropping positions were randomly sampled across the entire time series to ensure coverage of the full P-QRS-T waveform. The NT-Xent loss function was also used here.

As shown in the ablation results in Table 3, RRC alone performed poorly in helping the model learn ECG features. This may be because TO only discards partial temporal information without distorting signal morphology, whereas RRC risks losing global cross-temporal dependencies. Furthermore, TO exhibits stronger robustness to low-quality signals, while RRC relies heavily on high signal integrity. If the original signal contains locally concentrated noise and the random crop happens to select these regions, rescaling could amplify noise features and thereby interfere with model training. The proposed method effectively addresses these shortcomings, leading to a clear improvement in model performance.

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

Table 3: Ablation study.

Method	Classfication task		
	Ave ACC	Ave AUC	Ave F1
RRC	0.937	0.756	0.612
TO	0.952	0.819	0.611
ours	0.958	0.826	0.670

4.3. ECG Classification Based on TO-RRC Augmentation and the SimCLR Framework

To validate the effectiveness of the proposed self-supervised electrocardiogram (ECG) classification method combining TO-RRC augmentation and the SimCLR framework, a fully supervised learning approach was adopted as the baseline model for performance comparison. To ensure fairness and rigor in the comparison, the baseline model employs a 1D-ResNet50 as the encoder, with its data preprocessing pipeline and training hyperparameters—including the Adam optimizer, initial learning rate, weight decay coefficient, etc.—kept entirely consistent with the pre-training phase of the proposed self-supervised method. After pre-training, the self-supervised method transfers the encoder's weights to the downstream ECG classification task for fine-tuning, while the fully supervised method directly performs end-to-end training based on the same 1D-ResNet50 encoder. The performance comparison between the two methods on the ECG classification task is presented in Table 4.

Table 4: Comparison between fully-supervised methods and our self-supervised approach.

Method	Classfication task		
	Ave ACC	Ave AUC	Ave F1
Supervised	0.910	0.822	0.567
ours	0.987	0.850	0.716

5. Discussion and Conclusion

The proposed method demonstrates strong applicability in ECG analysis tasks where labeled data is scarce. Compared to fully supervised deep learning methods, it significantly reduces dependence on annotated data. With the same amount of labeled ECG signals, the classification accuracy is improved by 4.8%, while training efficiency is also further enhanced. Experiments indicate that self-supervised learning can enhance model generalization and training stability by effectively leveraging unlabeled data, offering a viable pathway to alleviate the widespread issue of label scarcity in the medical field. Moreover, the pre-trained encoder exhibits strong transferability: it can not only be quickly adapted to downstream tasks such as ECG classification but also be fine-tuned for other related medical time-series analysis tasks. This maintains performance while substantially reducing the time and labeling costs required for model training.

References

[1] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019) (pp. 3465–3469).

[2] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML) (pp. 1597–1607).

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

[4] Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171–4186).

*[5] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291*

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

[6] Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint. https://doi.org/10.48550/arXiv.1807.03748

[7] Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. Advances in Neural Information Processing Systems, 33, 20033–20044.

[8] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, R., Munos, R., & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33.

[9] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems, 33.

[10] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., Bittencourt, M. S., Boehme, A. K., Buxton, A. E., Carson, A. P., Commodore-Mensah, Y., et al. (2022). Heart disease and stroke statistics—2022 update: A report from the American Heart Association. Circulation, 145(8), e153–e639.

[11] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

[12] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint. https://doi.org/10.48550/arXiv.1707.01836

[13] Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira Jr., W., Schön, T. B., & Ribeiro, A. L. P. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications, 11(1), 1760. https://doi.org/10.1038/s41467-020-15432-4

[14] Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., Erickson, B. J., Kapa, S., & Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. The Lancet, 394(10201), 861–867. https://doi.org/10.1016/S0140-6736(19)31721-0

[15] Petmezas, G., Haris, K., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J. A., Katsaggelos, A. K., & Maglaveras, N. (2021). Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets. Biomedical Signal Processing and Control, 63, 102194. https://doi.org/10.1016/j.bspc.2020.102194

[16] Jun, T. J., Nguyen, H. M., Kang, D., Kim, D., Kim, D., & Kim, Y.-H. (2018). ECG arrhythmia classification using a 2-D convolutional neural network. arXiv preprint. https://doi.org/10.48550/arXiv.1804.06812

[17] Ullah, A., Rehman, S. U., Tu, S., Mehmood, R. M., & Fawad. (2021). A hybrid deep CNN model for abnormal arrhythmia detection based on cardiac ECG signal. Sensors, 21(3), 951. https://doi.org/10.3390/s21030951

[18] Mousavi, S., Afghah, F., Razi, A., & Acharya, U. R. (2019). ECGNET: Learning where to attend for detection of atrial fibrillation with deep visual attention. In 2019 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1-4). IEEE. https://doi.org/10.1109/BHI.2019.8834637

[19] Zihlmann, M., Perekrestenko, D., & Tschannen, M. (2017). Convolutional recurrent neural networks for electrocardiogram classification. In 2017 Computing in Cardiology (CinC) (pp. 1-4). IEEE.

[20] Chen, T., Huang, C., Shih, E. S., Hu, Y., & Hwang, M. (2020). Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. iScience, 23(3), 100886. https://doi.org/10.1016/j.isci.2020.100886

[21] Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2020). Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. IEEE Journal of Biomedical and Health Informatics, 25(5), 1519–1528. https://doi.org/10.1109/JBHI.2020.3022989

[22] Huang, J., Chen, B., Yao, B., & He, W. (2019). ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. IEEE Access, 7, 92871–92880. https://doi.org/10.1109/ACCESS.2019.2928017

[23] Liu, H., Zhao, Z., & She, Q. (2021). Self-supervised ECG pre-training. Biomedical Signal Processing and Control, 70, 103010. https://doi.org/10.1016/j.bspc.2021.103010

[24] Yang, S., Lian, C., Zeng, Z., Pan, J., Wang, Y., Luo, L., Du, X., & Lin, B. (2024). Masked self-supervised ECG representation learning via multiview information bottleneck. Neural Computing and Applications, 36, 7625–7637. https://doi.org/10.1007/s00521-024-09486-4

[25] Mehari, T., & Strodthoff, N. (2022). Self-supervised representation learning from 12-lead ECG data.

ISSN 2616-5775 Vol. 9, Issue 2: 1-10, DOI: 10.25236/AJCIS.2026.090201

Computers in Biology and Medicine, 141, 105114. https://doi.org/10.1016/j.compbiomed.2021.105114
[26] Lai, J., Tan, H., Wang, J., Wang, M., Li, J., Xiao, J., Liu, B., Yang, X., & Li, G. (2023). Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. Nature Communications, 14, 3741. https://doi.org/10.1038/s41467-023-39472-8
[27] Luo, C., Wang, G., Ding, Z., Zhang, Y., & Li, Y. (2021). Segment origin prediction: A self-supervised learning method for electrocardiogram arrhythmia classification. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1132–1135). IEEE.
[28] Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Wong, A.-K. I., Liu, C., Liu, F., Rad, A. B., Elola, A., Seyedi, S., Li, Q., Sharma, A., Clifford, G. D., & Reyna, M. A. (2020). Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. Physiological Measurement, 41(12), 124003. https://doi.org/10.1088/1361-6579/abc960
[29] Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., & Rakovski, C. (2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10, 000 patients. Scientific Data, 7(1), 54. https://doi.org/10.1038/s41597-020-0386-x
[30] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1–3), 37–52.
[31] Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 245–250).
[32] Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics, 59(4), 291–294.
[33] Oh, J., Chung, H., Kwon, J., Hong, D., & Choi, E. (2022). Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In Proceedings of the 2022 ACM Conference on Health, Inference, and Learning (CHIL '22). https://doi.org/10.48550/arXiv.2203.06889