Multimodality Multi-Lead ECG Arrhythmia Classification using Self-Supervised Learning

1st Thinh Phan  
Department of AI Convergence  
Chonnam National University  
Gwangju, South Korea 61186  
phantrandacthin2382@gmail.com

2nd Duc Le  
Department of CSCE  
University of Arkansas, Fayetteville, Arkansas 72703  
minhducl@uark.edu

3rd Patel Brijesh  
Department of Cardiology  
West Virginia University  
Morgantown, WV 26506  
brijesh.patel@wvumedicine.org

4th Donald Adjeroh  
Department of CSEE  
Morgantown, WV 26506  
don@csee.wvu.edu

5th Jingxian Wu  
Department of ELEG  
University of Arkansas, Fayetteville, Arkansas 72703  
wuj@uark.edu

6th Morten Olgaard Jensen  
Department of Biomedical Engineering  
University of Arkansas, Fayetteville, Arkansas 72703  
mojensen@uark.edu

7th Ngan Le  
Department of CSCE  
University of Arkansas, Fayetteville, Arkansas 72703  
thile@uark.edu

Abstract—Electrocardiogram (ECG) signal is one of the most effective sources of information mainly employed for the diagnosis and prediction of cardiovascular diseases (CVDs) connected with the abnormalities in heart rhythm. Clearly, single modality ECG (i.e. time series) cannot convey its complete characteristics, thus, exploiting both time and time-frequency modalities in the form of time-series data and spectrogram is needed. Leveraging the cutting-edge self-supervised learning (SSL) technique on unlabeled data, we propose SSL-based multimodality ECG classification. Our proposed network follows SSL learning paradigm and consists of two modules corresponding to pre-stream task, and down-stream task, respectively. In the SSL-pre-stream task, we utilize self-knowledge distillation (KD) techniques with no labeled data, on various transformations and in both time and frequency domains. In the down-stream task, which is trained on labeled data, we propose a gate fusion mechanism to fuse information from multimodality. To evaluate the effectiveness of our approach, ten-fold cross validation on the 12-lead PhysioNet 2020 dataset has been conducted. https://github.com/UARK-AICV/ECG-SSL.

Index Terms—ECG classification, self-supervised learning, contrastive learning, multimodalities, multi-lead

I. INTRODUCTION

CVDs are leading causes of deaths globally. The mortality rate can be considerably reduced by early treatment if occult signals linked with CVD are detected by ECG. The ECG signals, which record cardiac electrical activities, are widely adopted to diagnose abnormal heart rhythms and intra-cardiac conduction abnormalities.

Traditionally, cardiac feature extraction [1] and pattern classifiers [2] are separated. Notwithstanding the decent performance, they are not applicable to real-life cases because of high time consumption and complexity. Recent DNNs-based approaches have obtained amazing research progress in various domains [3]–[5]. DNN-based ECG classification in general can be categorized into single lead classification [6], [7] or multiple lead classification [8], [9]. Our proposed method belongs to the second category. In this group, [8] proposed a simple residual neural network used for classifying 6 types of abnormalities on the in-house 12-lead database, where some of its testing metrics are better than those of expert cardiologists. [9] trained a CNNs-based structure on multi-lead ECG data to diagnose myocardial infraction. Due to high complexity in higher-dimensional data of 12-lead ECG, different architectures have been adopted to model time correlation among ECG sample points [10], [11]. Besides time-series data, time-frequency has played an important role in ECG analysis. [12] used STFT-based spectrogram and 2D CNNs for ECG arrhythmia classification. [7] proposed using STFT and stationary wavelet transform (SWT) transformations to obtain two-dimensional (2-D) matrix input suitable for deep CNNs. [13] proposed a novel wavelet sequence based on deep bidirectional LSTM network model.

Furthermore, with the growing demand for medical examination and treatment, the healthcare industry steadily accumulates innumerable amounts of data but these unlabelled data might not be serviceable for most tasks. To address this limitation, we leverage the recent advanced SSL techniques, i.e., contrastive learning [14]. A common workflow to apply SSL is to train the network in an unsupervised manner by learning with a pre-stream task, and then finetuning the pre-trained network on a target downstream task. The suitable pre-stream tasks can be considered in four categories: context-based [15], generation-based [16], free semantic label-based [17], [18], and cross-modal-based [19], [20]. Recently, SSL-based DNNs have also been applied in ECG classification. [21] proposed a self-designed loss to bring the representations from the same patient closer and study the shared context of individual recordings through time and scenarios. [22] customized an SSL model that understands the differences among segments from individual patients and dissimilarities between patients’ recordings from same category. [23] conducted extensive experiments on four SSL methods with multiple combinations of transformations and demonstrated the improvement in macro
A. Self-Knowledge Distillation (KD): a revisit

KD [25] is a learning paradigm where a student network \( S_0 \) is trained to match the output of a given teacher network \( T_0 \), parameterized by \( \theta \) and \( \phi \), respectively. Given an input signal \( x \), KD is explained in the following steps: (i) Apply different transformations \( T_1, T_2 \) on \( x \) to generate global view \( x_1 = T_1(x) \) and local view \( x_2 = T_2(x) \) at different distorted views, or crops. On the same transformation, the global transformation \( T_1 \) and local transformation \( T_2 \) are defined with different crop ratios i.e. greater than 50% is considered as global; otherwise, it is local. ; (ii) Pass \( x_1 \) and \( x_2 \) to networks of \( S_0 \) and \( T_0 \); (iii) Compute similarity between probability distributions of output from \( S_0 \) and \( T_0 \) using a cross-entropy loss. Notably, the output of \( T_0 \) is centered with a mean computed over the batch.

Unlike typical KD, Self-KD [14] build \( T_0 \) using the past iterations of \( S_0 \). Thus, both \( T_0 \) and \( S_0 \) are sharing the same network architecture with different sets of parameters \( \theta \) and \( \phi \). Notably, both \( T_1 \) and \( T_2 \) are sharing the same transformations with different crop ratios. The overall flowchart of self-KD is illustrated in Fig. 2.

![Fig. 2. Illustration of SSL with self-KD mechanism with no labeled data.](Image)

B. SSL: Pre-stream Task

The pre-stream task makes use of self-KD technique to pre-train models of time-series data and time-frequency signal using two components as shown in the yellow block in Fig. 1. To apply self-KD technique to two different modalities, we need to define the network architectures of \( T_0, S_0 \) and transformations of \( T_1, T_2 \) as follows:

1. Pre-stream task with time-series data This component takes time-series \( \mathcal{X} \) as an input. The SSL KD in this task is performed by Time Cutout (TC) and Gaussian Noise (GN) transformation \( T_1 \) as follows:

   \[ X_{i,TC} : \{X_i\}_{i=1}^{n} = X_{i,TC} + U(0,0.5) \cdot X \]

   where \( t_1 = \text{rand}(0,n - \alpha), t_2 = t_1 + \alpha, \alpha = U(0,0.5).n \), with \( n \) denoting the length of \( \mathcal{X} \), and \( U \) denoting the uniform distribution.

   Let \( N(0,\sigma) \) denote the Gaussian Noise (GN). The GN transformation is defined as: \( X_{i,GN} : X_i + N(0,\sigma) \) The \( T_1 \) transformation is \( T_1 : X_{i,TC} + X_{i,GN} \).

   We utilize xresnet1d50 [26] as our backbone network for both \( T_0 \) and \( S_0 \), which are trained on different parameters of \( \phi \) and \( \theta \).

2. Pre-stream task with time-frequency data This component takes time-frequency \( \mathcal{F} \) as an input. \( \mathcal{F} \) is generated by applying Short Time Fourier Transform (STFT) to the time-series \( \mathcal{X} \). The equation for STFT is shown in Eq:2 where \( S \) is the STFT function and \( g(n - m) \) is the window function. Usually a Hann or a Gaussian window is used and the width of window is specified by \( m \).

   \[ \{s_i\}_{i=1}^{n} = S(\{X_i\}_{i=1}^{n}) \]

   \[ S(\{X_i\})(k,m) = \sum_{n=0}^{N-1} X(n) g(n - m) e^{-\frac{12 \pi k n}{N}} \]  

Fig. 1. Overall flowchart of our proposed network consisting two modules i.e., SSL pre-stream task and SSL down-stream task. In the SSL pre-stream task, the pre-trained models are student networks and they are trained on unlabeled data with self-KD mechanism. In the SSL pre-stream task, the pre-trained models are fine-tuned on labeled data. The features \( f_1 \) from time-series and \( f_2 \) from spectrogram are fused by our proposed gate fusion.

AUC when using SSL. However, the existing SSL-based ECG classification approaches only focus on time domain with time-series data and they have not exploited the ECG characteristics in frequency domain.

In this work, we adopt the advanced SSL technique, i.e., self-distillation SSL with no labels [14], [24]. Our proposed network contains two modules i.e., SSL pre-stream task, SSL down-stream task. The first module is trained on unlabeled data and contains two components corresponding to 1D CNNs for time-series data and 2D CNNs for time-frequency spectral (spectrogram) signal. The second module fine-tunes the pre-trained models i.e. 1D CNNs and 2D CNNs from the first module to perform mutli-lead ECG classification on labeled data. In the down-stream task, the features from the two networks are fused by our proposed gate fusion mechanism. The overall flowchart of our proposed network is shown in Fig. 1.

II. PROPOSED METHODS

A. Self-Knowledge Distillation (KD): a revisit

KD [25] is a learning paradigm where a student network \( S_0 \) is trained to match the output of a given teacher network \( T_0 \), parameterized by \( \theta \) and \( \phi \), respectively. Given an input signal \( x \), self-KD is explained in the following steps: (i) Apply different transformations \( T_1, T_2 \) on \( x \) to generate global view \( x_1 = T_1(x) \) and local view \( x_2 = T_2(x) \) at different distorted views, or crops. On the same transformation, the global transformation \( T_1 \) and local transformation \( T_2 \) are defined with different crop ratios i.e. greater than 50% is considered as global; otherwise, it is local. ; (ii) Pass \( x_1 \) and \( x_2 \) to networks of \( S_0 \) and \( T_0 \); (iii) Compute similarity between probability distributions of output from \( S_0 \) and \( T_0 \) using a cross-entropy loss. Notably, the output of \( T_0 \) is centered with a mean computed over the batch.

Unlike typical KD, Self-KD [14] build \( T_0 \) using the past iterations of \( S_0 \). Thus, both \( T_0 \) and \( S_0 \) are sharing the same network architecture with different sets of parameters \( \theta \) and \( \phi \). Notably, both \( T_1 \) and \( T_2 \) are sharing the same transformations with different crop ratios. The overall flowchart of self-KD is illustrated in Fig. 2.
where \( k \) and \( m \) are the time index and frequency index, respectively. The time-frequency responses \( \{s_i\}_{i=1}^{n} \) is passed through transformations \( T^2 \), (i.e. \( T^2_1 \) and \( T^2_2 \) are global and local transformations). The SSL KD in this task is performed by Time Cutout (TC) and Frequency Cutout (FC) transformation \( T^2 \) as follows:
\[
T^2 : \{F_{x,y}\}_{x=t_1,y=f_1} = 0
\]
where \( \{F_{x,y}\}_{x=t_1,y=f_1} = 0 \) and \( \{F_{x,y}\}_{y=f_1} = 0 \) are TC and FC transformations, respectively.

We utilize SE-ResNet34, a modified version of ResNet [27] as our backbone network for both \( T^2_1 \) and \( S^2_1 \) which are trained on different parameters of \( \hat{\phi} \) and \( \hat{\theta} \).

C. SSL: Down-stream Task

Different from pre-stream task trained on unlabeled data, the down-stream task is trained on labeled data. Given a signal \( \mathcal{X} \), we transfer student models (i.e., \( S^1_1, S^2_1 \)), which were unsupervised trained by the previous module, to extract ECG characteristic from time-series (\( f_1 = S^1_1(\mathcal{X}) \)) and time-frequency signal (\( f_2 = S^2_1(\mathcal{X}) \)). Because features \( f_1 \) and \( f_2 \) carry different characteristic of \( \mathcal{X} \), we propose a gate fusion mechanism as follows: (i) We first process features \( f_1 \) and \( f_2 \) through a soft attention [28] to learn impact factors (\( w_1, w_2 \)) of each feature. The output are weighted and denoted as \( \hat{f}_1 \) and \( \hat{f}_2 \). (ii) We then stack features \( \hat{f}_1 \) and \( \hat{f}_2 \) and form \( f = [\hat{f}_1, \hat{f}_2] \). A multi-layer perceptron (MLP) and sigmoid are then applied into \( f \). (iii) We finally train the down-stream task with a binary cross entropy loss between the predicted label output from the sigmoid function and groundtruth label.

III. EXPERIMENTAL RESULTS

A. Dataset, Metrics

We conduct our experiments on PhysioNet 2020 dataset [29]. PhysioNet comprises six ECG datasets and the total number of records are 43,101 for 111 diagnoses. To fairly compare with other existing methods, we adopt data stratification from [23] and only 25 classes are selected according to the challenge. Our proposed network is trained using ten-fold cross validation. For a broader analysis of performance the results are described under all recommended metrics i.e., AUROC, AUPRC, Accuracy (Acc), F1, F2, G-score and the challenge metric (ChM) as in [29].

| Model | AUCR | AUPRC | Acc | F1 | F2 | G2 | ChM |
|-------|------|-------|-----|----|----|----|-----|
| Existing Works | | | | | | | |
| [30] | 84.6 | 92.2 | 20.9 | 25.3 | 71.9 | 23.6 |
| [31] | 81.6 | 89.5 | 21.4 | 29.7 | 67.3 | 25.1 |
| [32] | 80.6 | 82.2 | 21.1 | 29.0 | 68.4 | 24.8 |
| [33] | 82.5 | 81.2 | 22.8 | 31.3 | 68.0 | 24.8 |
| [34] | 87.2 | 87.2 | 23.9 | 33.7 | 70.2 | 24.8 |
| [35] | 55.1 | 52.3 | 32.8 | 38.3 | 63.2 | 23.7 |
| [36] | 57.3 | 54.2 | 33.5 | 39.8 | 63.9 | 24.8 |
| [37] | 57.3 | 54.2 | 33.5 | 39.8 | 63.9 | 24.8 |
| [38] | 57.3 | 54.2 | 33.5 | 39.8 | 63.9 | 24.8 |

Our

| SSL-T | 95.7 | 64.8 | 48.8 | 61.4 | 64.9 | 40.2 | 64.7 |
| SSL-S | 93.8 | 56.3 | 40.7 | 53.2 | 57.4 | 33.9 | 58.0 |
| SSL-TSG | 96.0 | 65.8 | 48.9 | 62.1 | 65.9 | 41.0 | 65.4 |

Transformations are crucial to the success of the adopted SSL [14]; hence, we further conduct an ablation study on different transformations as shown in Fig. 3. On each domain, a baseline is defined when no transformation is used. In time series domain, we report the performance on transformation of TC, GN and \( T^1 \) as defined in Subsec.II-B.1. In time-frequency domain, we report the performance on transformation of TC, FC and \( T^2 \) as defined in Subsec.II-B.2. The baseline obtains good performance – thanks to the inherent temporal dependency modeling in our network design. Given a long ECG record, most of existing work pre-processed the ECG signal by partitioning it into a set of overlapping segments. In such approaches, the temporal relationship between segments mainly depends on the overlap threshold, which is typically set as a pre-defined parameter. Unlike the existing work, we randomly crop the ECG record into a set of segments at each
iteration, thus, the overlap between segments is more flexible and the inter-segment coherence is modeled better.

**Conclusion**

In this paper, we have proposed an SSL-based multimodal network for multi-lead ECG classification. Our approach includes two modules, one for the pre-stream task to pre-train networks on unlabeled data, and the other on the down-stream task to fine-tune the networks on labeled data. Our SSL is built based on a self-KD mechanism. We have investigated two modalities corresponding respectively to time-series data, and spectrogram in time-frequency domain. To combine features from multiple modalities, we have introduced a gate fusion network for multi-lead ECG classification. Our approach has shown that SSL on multi-modality data is an effective approach to classify multi-lead ECG.

**Acknowledgments:** This material is based upon work supported in part by the US National Science Foundation, under Award No. OIA-1946391, NSF 1920920.

**References**

[1] V Chouhan and S Mehta. Threshold-based detection of p and t-wave in ecg using new feature signal. *Com.Sci.& Net.Sec.*, 8(2):144–153, 2008.

[2] M Zahibi, A Rad, et al. Detection of atrial fibrillation in ecg hand-held devices using a random forest classifier. In *CinC*, pages 1–4, 2017.

[3] Kushu Yamazaki, Viet-Ho Ho, Darshan Bulsara, and Ngan Le. Spiking neural networks and their applications: A review. *Brain Sciences*, 12(7):363, 2022.

[4] Ngan Le, Vidyeshwar Singh Rathour, Kushu Yamazaki, Khoo Luu, and Marios Savvides. Deep reinforcement learning in computer vision: a comprehensive survey. *Artificial Intelligence Review*, pages 1–87, 2021.

[5] S Kevin Zhou, Hoang Ngan Le, Khoo Luu, and Nicholas Ayache. Deep reinforcement learning in medical imaging: A literature review. *Medical image analysis*, 73:102193, 2021.

[6] P Rajpurkar, A.Y Hannun, et al. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

[7] M.D Le, V.S Rathour, et al. Multi-module recurrent convolutional neural net with transformer encoder for ecg arrhythmia classification. In *BIH*, pages 1–5, IEEE, 2021.

[8] A.H Ribeiro, M.Horta Ribeiro, G.M.M Paixão, D.M. Oliveira, P.R Gomes, J.A Canazart, M.P.S Ferreira, C.R Andersson, P.W MacFarlane, W Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1–9, 2020.

[9] U.B Baloglu, M Talo, O Yildirim, R San Tan, and U.R Acharya. Classification of myocardial infarction with multi-lead ecg signals and deep cnn. *Pattern Recognition Letters*, 122:23–30, 2019.

[10] R Zhang, P Isola, and A.A Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.